SIAM J. CONTROL OPTIM.
Vol. 41, No. 1, pp. 1–22

STOCHASTIC APPROXIMATION FOR NONEXPANSIVE MAPS: APPLICATION TO Q-LEARNING ALGORITHMS*

JINANE ABOUNADI†, DIMITRI P. BERTSEKAS†, AND VIVEK BORKAR‡

Abstract. We discuss synchronous and asynchronous iterations of the form

$$x^{k+1} = x^k + \gamma(k)(h(x^k) + w^k),$$

where h is a suitable map and $\{w^k\}$ is a deterministic or stochastic sequence satisfying suitable conditions. In particular, in the stochastic case, these are stochastic approximation iterations that can be analyzed using the ODE approach based either on Kushner and Clark's lemma for the synchronous case or on Borkar's theorem for the asynchronous case. However, the analysis requires that the iterates $\{x^k\}$ be bounded, a fact which is usually hard to prove. We develop a novel framework for proving boundedness in the deterministic framework, which is also applicable to the stochastic case when the deterministic hypotheses can be verified in the almost sure sense. This is based on scaling ideas and on the properties of Lyapunov functions. We then combine the boundedness property with Borkar's stability analysis of ODEs involving nonexpansive mappings to prove convergence (with probability 1 in the stochastic case). We also apply our convergence analysis to Q-learning algorithms for stochastic shortest path problems and are able to relax some of the assumptions of the currently available results.

Key words. stochastic approximation, Q-learning, neuro-dynamic programming

AMS subject classifications. 62L20

PII. S0363012998346621

1. Introduction.

The motivation for this paper has been the analysis of Q-learning algorithms, which have emerged as a powerful simulation tool for solving dynamic programming problems when a model is not known and/or the problem must be solved on-line as the data become available. Q-learning algorithms were first formulated by Watkins (1989), who gave a partial convergence analysis that was later amplified by Watkins and Dayan (1992). A more comprehensive analysis was given by Tsitsiklis (1994) (also reproduced in Bertsekas and Tsitsiklis (1996)), which made the connection between Q-learning and stochastic approximation. (A related treatment of a class of algorithms that include Q-learning and TD(λ) also appeared around the same time in Jaakola, Jordan, and Singh (1994). It may be recalled here that TD(λ) is a learning scheme for estimating the value function of a policy based on an exponentially weighted average (with weights λ^n for some $\lambda \in (0,1)$) of the so-called n-step truncated returns—see Bertsekas and Tsitsiklis (1996) for a detailed description.) In particular, Q-learning algorithms for discounted cost problems or stochastic shortest path (SSP) problems were viewed as asynchronous stochastic approximation versions of well-known value iteration algorithms in dynamic programming. This connection paved the way for a general analysis based on classic stochastic approximation techniques and dynamic programming-related contraction and monotonicity properties.

*Received by the editors October 29, 1998; accepted for publication August 31, 2001; published electronically March 27, 2002. This research was supported by the NSF under grant 9600494-DMI.
http://www.siam.org/journals/sicon/41-1/34662.html

†Department of Electrical Engineering and Computer Science, M.I.T., Cambridge, MA 02139 (jinane@mit.edu, dimitri@mit.edu).

‡School of Technology and Computer Science, Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400005, India (borkar@titr.res.in). The research of this author was supported in part by the Homi Bhabha Fellowship and by the government of India, Department of Science and Technology grant III 5(12)/96-ET.

A weakness of the methodology developed so far is that it deals in an ad hoc way with the question of boundedness of the Q-learning iterates. In particular, the analysis of Tsitsiklis required a special argument for proving boundedness with probability 1 (w.p.1), and for the case of SSP problems it also required that the cost per stage be nonnegative, unless boundedness is imposed as an assumption (see Bertsekas and Tsitsiklis (1996), Prop. 5.6).

Our purpose in this paper is to provide a new and powerful general framework for establishing boundedness and proving convergence in synchronous and asynchronous stochastic approximation methods involving nonexpansive maps, including as a special case Q-learning algorithms. Our framework relies strongly on nonexpansiveness and combines ideas from several fields, including asynchronous stochastic approximation analysis via the limiting ODE technique and nonlinear analysis of ODEs. Our method for dealing with boundedness bears a similarity to an idea from the paper by Jaakola, Jordan, and Singh (1994), which addressed the convergence of TD(λ) using stochastic approximation methods (see section 2). Also see Csibi (1975) and Gerencser (1992) for work in a similar spirit. As a special case of our analysis, we improve on Tsitsiklis' convergence result by dispensing with the boundedness assumption for the iterates of SSP Q-learning, in the case where the cost per stage may be negative. The methodology developed in this paper also provides an essential foundation for a convergence analysis of Q-learning algorithms for average cost dynamic programming problems given in a companion paper (Abounadi, Bertsekas, and Borkar (2001)).

Our results, in fact, can be cast as a powerful *deterministic* principle, because the conditions on the noise required to ensure its applicability can be cast in simple deterministic terms. These can, in turn, be verified in the almost sure sense for the stochastic approximation algorithms of interest here. The deterministic formulation also requires weaker conditions on the stepsizes. Thus we shall initially state our results in a deterministic framework, enlarging their scope beyond the applications to stochastic approximation.

The general framework that we propose applies to synchronous and asynchronous variants of algorithms of the form

$$x^{k+1} = x^k + \gamma(k)\big(h(x^k) + w^k\big). \tag{1}$$

Here x^k is a sequence in $\Re^n$, w^k is a deterministic noise sequence, h is Lipschitz, $\gamma(k)$ is a positive stepsize sequence, and the aim is to find a solution of the equation $h(x) = 0$. This is the synchronous implementation in which all components are updated together at each time with full information about past iterates. The asynchronous model that we use is based on the formulation of Borkar (1998) and is of the form

$$x_i^{k+1} = x_i^k + \gamma\big(\nu(k,i)\big)\big(h_i(x^k) + w_i^k\big)I(i \in Y^k) \tag{2}$$

for $i = 1, \ldots, n$, where Y^k is the subset of $\{1, 2, \ldots, n\}$ denoting components being updated at time k, $I(\cdot)$ is the indicator function, and $\nu(k,i)$ is the number of times the component x_i of the vector x has been updated by time k.

For the synchronous algorithm (1), a powerful analysis technique is the ODE method introduced by Ljung (1977), formally treated by Kushner and Clark (1978), and Benveniste, Metivier, and Priouret (1990). For the asynchronous algorithm (2), a similar technique has been developed by Borkar (1998). (See also Kushner and Yin (1997) and references therein for related work.) The major idea behind these two techniques is to find a limiting deterministic continuous-time ODE for the stochastic discrete-time processes, using interpolation with the appropriate time scaling. The

main result is that if the ODE has an asymptotically stable equilibrium point, then under appropriate assumptions, which include boundedness of the generated iterates, the discrete-time iteration converges to this point w.p.1. Thus, in ODE techniques, boundedness must be independently verified.

This paper's methodology for dealing with the boundedness issue involves three steps:

1. obtaining a related scaled iteration and establishing its convergence,
2. showing that the sequence $\{x^k\}$ generated by the original iteration is bounded as a consequence of the convergence of the scaled iteration,
3. showing that the boundedness of $\{x^k\}$ implies convergence by invoking a standard ODE limiting argument.

For each of the steps above, we will impose appropriate sufficient conditions on the mapping h, the stepsize, and the noise. A central assumption in our later applications is that the mapping h is of the form $h(x) = T(x) - x$, where the map T is nonexpansive with respect to some norm $\|\cdot\|_p$ with $p \in (1, \infty]$ for the synchronous case, and with respect to the sup-norm $\|\cdot\|_\infty$ for the asynchronous case. To our knowledge, ours is the first general method for dealing with the boundedness issues in the ODE approach where the underlying mapping T is not a contraction. (See, however, the recent work by Borkar and Meyn (2000), which is discussed later in this section.) Note that the class of fixed-point problems which involve nonexpansive mappings arises in a number of different applications (see the book by Bertsekas and Tsitsiklis (1989) and the papers by Tseng, Bertsekas, and Tsitsiklis (1990), Borkar and Soumyanath (1997), and Soumyanath and Borkar (1999)). In particular, it includes value iteration algorithms for various dynamic programming formulations, including Q-learning algorithms.

Step 1 of the scheme described above is carried out by choosing the scaling based on a Lyapunov function of an appropriate ODE. The scaling works like a projection on an appropriate bounded set when the iterates lie outside a certain level set of the Lyapunov function. Note that we do not need to know the Lyapunov function; all we need to know is that such a function exists. For this we will use a general converse Lyapunov theorem that guarantees the existence of a smooth Lyapunov function if the ODE has a globally asymptotically stable equilibrium point (Wilson (1969)). Given this scaling scheme, we will be able to show that the scaled iteration has the same deterministic limiting ODE and hence converges. The argument is similar to the standard limiting ODE argument of Kushner and Clark (1978). We need to consider the Skorohod topology instead of the "uniform convergence on compacts" topology on $C([0, \infty); \Re^n)$. Step 2 involves the idea of comparing the original iteration and its scaled counterpart and showing that the difference between the two is bounded due to the nonexpansiveness of the mapping F. The idea of comparing the two iterations appeared first in Jaakola, Jordan, and Singh (1994) in a more limited setting. Step 3 is an application of standard ODE limiting arguments since boundedness is already established.

It is instructive to compare this approach with that of Borkar and Meyn (2000). While both are motivated by the same class of algorithms, viz., Q-learning, they exploit different features of the latter. While our approach is solely based on the nonexpansivity of an associated map, Borkar and Meyn use a scaling limit of this map, in the spirit of fluid models in queueing theory. To underscore the difference, note that the stochastic gradient scheme can be viewed as a fixed-point seeking iteration of an L_2-nonexpansive map when the associated Hessian is uniformly bounded—see section III.B of Soumyanath and Borkar (1999). Thus it comes under the purview of

the present scheme, but not under that of Borkar and Meyn (2000) in the absence of any specification of how the gradient in question behaves near infinity. On the other hand, the requirement that a convenient scaling limit hold in their sense can be met without the map being nonexpansive: the former concerns only the behavior near infinity, but the latter is a global requirement. Thus the approach of Borkar and Meyn and that of the present paper are quite distinct, and given the paucity of general purpose criteria for the stability of stochastic recursions of this type, both are of interest, despite the fact that currently they are aimed at broadly the same class of problems. More generally, our scheme will work (under mild technical assumptions) for the recursions wherein the distance between iterates for two instantiations of the algorithm with the same random inputs, but with two different initial conditions, remains bounded by a function of the initial conditions.

Finally, we note that for recursive algorithms the idea of using projection as a way of forcing boundedness is not new. The difference in our approach is that the use of scaling is only a method of proof, and the objective is to establish the boundedness of the original iteration without altering the iterates by forcing them to be bounded.

2. Boundedness lemmas. The results in this paper will be divided into two parts: the boundedness lemmas and the convergence analysis of appropriately scaled synchronous and asynchronous iterations. The boundedness lemmas are given in the present section, and rely on the nonexpansiveness property of the concerned map with respect to some norm $\|\cdot\|_p$, $p \in (1, \infty]$, for the synchronous case, and the sup-norm for the asynchronous case. The convergence of the scaled iteration is analyzed in the next section.

For a set $\mathcal{A}$ of $\Re^n$, we denote by $\partial\mathcal{A}$ and $\bar{\mathcal{A}}$ the boundary and closure of $\mathcal{A}$, respectively (i.e., $\bar{\mathcal{A}} = \mathcal{A} \cup \partial\mathcal{A}$). We introduce via scaling a map that "projects" any point onto a bounded and open set $\mathcal{B}$ that contains the origin. This is done each time the point leaves a given set $\mathcal{C}$ that contains $\mathcal{B}$. The map is defined as follows.

DEFINITION 2.1. *Let $\mathcal{B}$ be an open and bounded subset of $\Re^n$ containing the origin, and let $\mathcal{C}$ be a subset of $\Re^n$ that contains $\mathcal{B}$. We define the mapping $\Pi_{\mathcal{B},\mathcal{C}} : \Re^n \mapsto \bar{\mathcal{B}}$ by*

$$\Pi_{\mathcal{B},\mathcal{C}}(x) = \gamma_{\mathcal{B},\mathcal{C}}(x) \cdot x,$$

where $\gamma_{\mathcal{B},\mathcal{C}} : \mathbb{R}^n \to (0, 1]$ is given by

$$\gamma_{\mathcal{B},\mathcal{C}}(x) = \begin{cases} 1 & \text{if } x \in \mathcal{C}, \\ \max\{\beta > 0 : \beta x \in \bar{\mathcal{B}}\} & \text{if } x \notin \mathcal{C}. \end{cases}$$

Since $\bar{\mathcal{B}}$ is compact, it can be seen that $\Pi_{\mathcal{B},\mathcal{C}}$ is well defined as a real-valued function. If $\mathcal{B}$ is an open ball with respect to the Euclidean norm centered at the origin, the map $\Pi_{\mathcal{B},\mathcal{C}}$ is like a projection on $\mathcal{B}$, but the decision to project depends on whether the point is outside the larger set $\mathcal{C}$.

Our first result is inspired by a lemma of Jaakola, Jordan, and Singh (1994), which guarantees convergence of an iteration as long as a scaled version converges. Their lemma uses a strong homogeneity assumption, which is unnecessary for our purposes.

LEMMA 2.1. *Let $\mathcal{B}$ be an open and bounded subset of $\Re^n$ containing the origin, and let $\mathcal{C}$ be a subset of $\Re^n$ that contains $\mathcal{B}$. Consider the algorithm*

$$x^{k+1} = G^k(x^k, \xi^k), \tag{3}$$

where we assume the following:

1. $\{\xi^k\}$ *is a sequence in a measurable space* $(\Omega, \mathcal{F})$.
2. G^k *is nonexpansive in* x *with respect to some norm* $\|\cdot\|$ *for every* $\xi \in \Omega$*:*

$$\|G^k(x,\xi) - G^k(y,\xi)\| \le \|x - y\| \qquad \forall\ x, y, \xi.$$

3. *The sequence* $\{\tilde{x}^k\}$ *generated by the scaled iteration*

$$\tilde{x}^{k+1} = G^k\big(\Pi_{\mathcal{B},\mathcal{C}}(\tilde{x}^k), \xi^k\big), \qquad \tilde{x}^0 = x^0,$$

converges to some vector $x^* \in \mathcal{B}$.

Then $\{x^k\}$ *is bounded.*

Proof. Since $\mathcal{B}$ is open, there exists a large enough $\overline{k}$ such that $\tilde{x}^k \in \mathcal{B}$ for $k \ge \overline{k}$. In other words, there exists a large enough $\overline{k}$ such that

$$\gamma_{\mathcal{B},\mathcal{C}}(\tilde{x}^k) = 1 \qquad \forall\ k \ge \overline{k}, \tag{4}$$

and hence

$$\tilde{x}^{k+1} = G^k(\tilde{x}^k, \xi^k) \qquad \forall\ k \ge \overline{k}. \tag{5}$$

Therefore, for $k \ge \overline{k}$,

$$\|x^{k+1} - \tilde{x}^{k+1}\| = \|G^k(x^k, \xi^k) - G^k(\tilde{x}^k, \xi^k)\| \le \|x^k - \tilde{x}^k\| \le \cdots \le \|x^{\overline{k}} - \tilde{x}^{\overline{k}}\|.$$

Since $\{\tilde{x}^k\}$ is bounded, it follows that $\{x^k\}$ is bounded. □

3. Analysis of the scaled iteration. Our objective is to apply Lemma 2.1 to the synchronous and asynchronous algorithms given by (1) and (2). To this end, we will first establish the convergence of scaled versions of iterations (1) and (2) by using ODE-type arguments and conclude boundedness of the unscaled versions. However, the scaling (i.e., the sets $\mathcal{B}$ and $\mathcal{C}$ in Lemma 2.1) must be chosen so that we can find a limiting ODE that is easily analyzed. In particular, if the scaling is not done appropriately, the scaled iteration might not converge. The iterates could, for example, keep hitting the boundary of $\mathcal{B}$ infinitely often and thus never converge, or the scaling could generate additional fixed points at the boundary that the iterates might converge to.

Given an ODE $\dot{x} = h(x)$ in $\Re^n$ with a global asymptotically stable equilibrium point x^*, a smooth Lyapunov function $V : \Re^n \mapsto \Re$ is a continuously differentiable function satisfying $V(x^*) = 0$, $V(x) > 0$ for all $x \ne x^*$, and such that the inner product of its gradient $\nabla V(x)$ and $h(x)$ is negative for all $x \ne x^*$. A necessary and sufficient condition for x^* to be a global asymptotically stable equilibrium point is the existence of a corresponding Lyapunov function (see Yoshizawa (1966)). Using some smoothing techniques, Wilson showed that the Lyapunov function can be taken to be smooth (in fact, infinitely differentiable; see Theorem 3.2 in Wilson (1969)). The following lemma will be useful to us.

LEMMA 3.1. *Let* $\dot{x} = h(x)$ *be an ODE with a global asymptotically stable equilibrium point* x^*. *Let* V *be a smooth Lyapunov function for the ODE. For any* $R > 0$, *there is a* $C > 0$ *such that the closed ball* $\bar{B}(x^*, R)$ *of radius* R *centered at* x^* *is in the interior of the level set* $L = \{x \in \Re^n : V(x) \le C\}$.

Proof. Consider the closure $\bar{B}(x^*, R)$ of $B(x^*, R)$. Since V is continuous and $\bar{B}(x^*, R)$ is compact, the maximum of V over $\bar{B}(x^*, R)$ is attained. Let $\overline{C} = \max_{x \in \bar{B}(x^*, R)} V(x)$. Any level set of the form $L = \{x \in \Re^n : V(x) \le C\}$, where $C > \overline{C}$, contains $\bar{B}(x^*, R)$ in its interior. □

3.1. Analysis of the scaled iteration-synchronous case. The scaled version of the synchronous algorithm of (1) is given by

$$\begin{aligned} \tilde{x}^{k+1} &= x^k + \gamma(k)\big(h(x^k) + w^k\big), \\ x^{k+1} &= \Pi_{\mathcal{B},\mathcal{C}}(\tilde{x}^{k+1}). \end{aligned} \tag{6}$$

We first show, under appropriate conditions, that this iteration converges w.p.1 to the unique equilibrium point of an appropriate ODE. The scaled iteration (6) can be written as

$$x^{k+1} = x^k + \gamma(k)\big(h(x^k) + w^k\big) + g^k, \tag{7}$$

where

$$g^k = \Pi_{\mathcal{B},\mathcal{C}}\big[x^k + \gamma(k)\big(h(x^k) + w^k\big)\big] - \big[x^k + \gamma(k)\big(h(x^k) + w^k\big)\big]. \tag{8}$$

We formally state our assumptions, which for completeness include some of our earlier assertions on the existence of a global asymptotically stable equilibrium x^*, the choice of the sets $\mathcal{B}$ and $\mathcal{C}$, etc., as follows.

ASSUMPTION 3.1. *The stepsizes $\gamma(k)$ satisfy*

$$0 < \gamma(k) \to 0, \qquad \sum_{k=0}^{\infty} \gamma(k) = \infty.$$

ASSUMPTION 3.2.

1. *There exists D such that $\|w^k\| \le D$ for all k.*
2. $\lim_{k\to\infty} \sum_{m=k}^{m_T(k)} \gamma(m) w^m = 0$ *for all T, where*

$$m_T(k) = \min\left\{ m \ge k : \sum_{l=k}^{m} \gamma(l) \ge T \right\}.$$

3. *h is Lipschitz continuous; i.e., for some $L > 0$,*

$$\|h(x) - h(y)\| \le L\|x - y\|.$$

4. *The ODE $\dot{x} = h(x)$ has a globally asymptotically stable equilibrium point x^*.*

Remark 3.1. Note that the boundedness condition in Assumption 3.2 is for the rescaled iterations, *not* for the original iterations. For the applications we have in mind, $||w^k||$ will be bounded by an affine function of $||x^k||$ and therefore will be bounded whenever the latter is. But the latter is bounded, by construction, for the rescaled iterations, and thus Assumption 3.2.1 is satisfied. It is being neither assumed nor implied a priori that the noise sequence $\{w^k\}$ in the original iterations is bounded; this will, in fact, be a consequence of our stability result. More generally, it will suffice to have $||w^k||$ bounded by a continuous function of x^k.

Remark 3.2. This remark concerns Assumption 3.2. The important thing to note here is that we are imposing this assumption on the *projected* algorithm, for which the boundedness of iterates is true by construction, not for the original scheme, whose stability we intend to prove.

In our analysis, we will use Lemma 2.1 with $\mathcal{B} = B(0, R)$ and $\mathcal{C} = \{x \in \Re^n : V(x) < C\}$, where $R > \|x^*\|$, V is a smooth Lyapunov function for the ODE $\dot{x} = h(x)$, and the constant C is large enough so that $\mathcal{C}$ contains $B(0, \bar{R})$ for some $\bar{R} > R$. Note

that the vector field of the ODE is transversal to the level sets of V, implying that if $x \in \partial\mathcal{C}$, then $(x + \Delta h(x)) \in \mathcal{C}$ for small enough $\Delta > 0$. This motivates the choice of the scaling sets $\mathcal{B}$ and $\mathcal{C}$ above. Intuitively, if the stepsize is small enough, we can think of the algorithm as starting at the boundary of $\mathcal{B}$ and moving around initially in $\mathcal{C}$. As it approaches the boundary of $\mathcal{C}$, it gets pushed back to the interior of $\mathcal{C}$, thanks to the fact that the vector field of the ODE on the boundary points inward and in spite of the noise term.

In order to proceed with our convergence analysis, we need to define piecewise linear or piecewise constant interpolated processes based on the iterates $\{x^k\}$. Let

$$t_k = \sum_{m=0}^{k-1} \gamma(m), \qquad k \geq 1,$$

with $t_0 = 0$. Let

$$\tilde{x}^{k+1} = x^k + \gamma(k)\big(h(x^k) + w^k\big), \qquad k \geq 0,$$

$$X_l(t) = \begin{cases} x^k & \text{for } t = t_k, \\ \big(1 - \frac{t-t_k}{\gamma(k)}\big)x^k + \frac{t-t_k}{\gamma(k)}\tilde{x}^{k+1} & \text{for } t \in [t_k, t_{k+1}), \end{cases}$$

$$X_c(t) = x^k, \qquad k \geq 0, \qquad \text{for } t \in [t_k, t_{k+1}),$$

$$G_c(t) = \sum_{m=0}^{k-1} g^m \qquad \text{for } t \in [t_k, t_{k+1}),$$

$$W_l(t) = \begin{cases} \sum_{m=0}^{k-1} \gamma(m) w^m & \text{for } t = t_k, \\ \big(1 - \frac{t-t_k}{\gamma(k)}\big)W_l(t_k) + \frac{t-t_k}{\gamma(k)} W_l(t_{k+1}) & \text{for } t \in [t_k, t_{k+1}). \end{cases}$$

Thus $X_l(\cdot)$ is right-continuous with left limits (r.c.l.l., for short); that is, $X_l(t^+) = \lim_{\delta \downarrow 0} X_l(t + \delta)$ and $X_l(t^-) = \lim_{\delta \downarrow 0} X_l(t - \delta)$ are well defined, with $X_l(t) = X_l(t^+)$. In fact, $X_l(\cdot)$ is piecewise linear and continuous everywhere, except at times t_k for which $g^k \neq 0$, where it has a jump discontinuity. Define the left-shifted versions of these processes as follows, for $t \geq 0$:

$$X_l^k(t) = X_l(t + t_k),$$

$$W_l^k(t) = W_l(t + t_k) - W_l(t_k),$$

$$X_c^k(t) = X_c(t + t_k),$$

$$G_c^k(t) = G_c(t + t_k) - G_c(t_k).$$

Then it is easy to see that for $t \geq -t_k$

$$\begin{aligned} X_l^k(t) &= X_l^k(0) + \int_0^t h\big(X_c^k(\tau)\big)d\tau + W_l^k(t) + G_c^k(t) \\ &= X_l^k(0) + \int_0^t h\big(X_l^k(\tau)\big)d\tau + W_l^k(t) + G_c^k(t) + e^k(t), \end{aligned}$$

where

$$e^k(t) = \int_0^t h\big(X_c^k(\tau)\big)d\tau - \int_0^t h\big(X_l^k(\tau)\big)d\tau.$$

By Assumption 3.2, $\{W_l^k(\cdot)\}$ converges to zero uniformly on finite intervals as $k \to \infty$. We show next that $\{e^k(\cdot)\}$ and $\{G_c^k(\cdot)\}$ behave analogously.

LEMMA 3.2. *For any $T > 0$, $\sup_{t\in[0,T]} \|e^k(t)\| \to 0$ as $k \to \infty$.*

Proof. By Assumptions 3.2,

$$\|e^k(t)\| \le \int_0^t \big\|h\big(X_c^k(\tau)\big) - h\big(X_l^k(\tau)\big)\big\|d\tau \le L\int_0^t \big\|X_c^k(\tau) - X_l^k(\tau)\big\|d\tau.$$

Letting

$$m_T(k) = \min\left\{m \ge k \ : \ \sum_{l=k}^{m}\gamma(m) \ge T\right\},$$

we have

$$\begin{aligned}
\sup_{t\in[0,T]} \|e^k(t)\| &\le L\int_0^T \big\|X_c^k(\tau) - X_l^k(\tau)\big\|d\tau \\
&\le \sum_{m=k}^{m_T(k)} \gamma(m)L \sup_{\tau\in[t_m,t_{m+1})} \big\|X_c^k(\tau) - X_l^k(\tau)\big\| \\
&\le \sum_{m=k}^{m_T(k)} \gamma(m)L(t_{m+1} - t_m)\big\|h(x^m) + w^m\big\| \\
&\le \sum_{m=k}^{m_T(k)} \gamma^2(m)LD',
\end{aligned}$$

where

$$D' = D + \sup_{x\in C}\|h(x)\|,$$

and the second inequality is a consequence of the definitions of $X_l^k(\cdot)$ and $X_c^k(\cdot)$. By Assumption 3.1, we have $\sum_{m=k}^{m_T(k)} \gamma^2(m) \to 0$ as $k \to \infty$, implying the result. ∎

To analyze the r.c.l.l. processes $X_l^k(\cdot)$, $G_c^k(\cdot)$, we recall from Billingsley (1968) the space $D([0,T];\Re^n)$ of r.c.l.l. functions from $[0,T]$ to $\Re^n$ (where $T > 0$), equipped with the Skorohod topology. This topology is defined so that $f^k(\cdot) \to f(\cdot)$ in $D([0,T];\Re^n)$ if and only if there exist continuous, nondecreasing, onto functions $\lambda^k : [0,T] \to [0,T]$ such that $f^k(\lambda^k(t)) \to f(t)$ and $\lambda^k(t) \to t$, uniformly on $[0,T]$. We denote by $D([0,\infty);\Re^n)$ the space of r.c.l.l. functions from $[0,\infty)$ to $\Re^n$, defined such that $f^k(\cdot) \to f(\cdot)$ in $D([0,\infty);\Re^n)$ if and only if their respective restrictions to $[0,T]$ converge in $D([0,T];\Re^n)$ for every $T > 0$. Both $D([0,T];\Re^n)$ and $D([0,\infty);\Re^n)$ are separable and metrizable with a complete metric.

We recall from Billingsley (1968, p. 118) the following characterization of relative compactness in $D([0,T];\Re^n)$: a set $A \subset D([0,T];\Re^n)$ is relatively compact if and only if

$$\sup_{x(\cdot)\in A}\sup_{t\in[0,T]}\|x(t)\| < \infty \tag{9}$$

and

$$\lim_{\delta\to 0}\sup_{x(\cdot)\in A}\sup_{t_1\le t\le t_2,\, t_2-t_1\le\delta}\min\bigl\{\|x(t)-x(t_1)\|,\,\|x(t_2)-x(t)\|\bigr\}=0, \tag{10a}$$

$$\lim_{\delta\to 0}\sup_{x(\cdot)\in A}\sup_{t_1,t_2\in[0,\delta)}\|x(t_2)-x(t_1)\|=0, \tag{10b}$$

$$\lim_{\delta\to 0}\sup_{x(\cdot)\in A}\sup_{t_1,t_2\in[T-\delta,T)}\|x(t_2)-x(t_1)\|=0. \tag{10c}$$

This generalizes the well-known Arzelà–Ascoli theorem for $C([0,T];\Re^n)$, the space of continuous functions from $[0,T]$ to $\Re^n$ with the sup-norm.

LEMMA 3.3. *The sequences* $\{X_l^k(\cdot)\}$ *and* $\{G_c^k(\cdot)\}$ *are relatively compact in* $D([0,\infty);\Re^n)$.

Proof. It suffices to check the relative compactness of their restrictions to $[0,T]$ in $D([0,T];\Re^n)$ for arbitrary $T>0$. Let us fix $T>0$. Since $\{x^k\}$ and $\{g^k\}$ are bounded, so are the sequences $\{X_l^k(\cdot)\}$ and $\{G_c^k(\cdot)\}$. Thus (9) above holds. It is easy to see that (10a)–(10c) will follow if any two discontinuity points of $x(\cdot)\in A$ are separated by at least some $\Delta>0$. For the processes under consideration, discontinuities occur at some of the t_k's. Let there be a discontinuity at t_k for some k. Then $g^{k-1}\neq 0$ and $x^k\in\partial B$. Let

$$d=\min_{x\in\partial B,\, y\in\partial C}\|x-y\|>0,$$

and define

$$m(k)=\max\left\{j\ :\ \sum_{i=0}^{j}\gamma(k+i)\le\frac{d}{D'}\right\},$$

where D' is as before. We claim that $x^{k+1},x^{k+2},\ldots,x^{k+m(k)}$ are in the interior of C. To see this, notice that if

$$\gamma(k)<\frac{d}{D'},$$

then

$$\|\tilde{x}^{k+1}-x^k\|<d,$$

implying that $\tilde{x}^{k+1}$ is in the interior of C and thus $x^{k+1}=\tilde{x}^{k+1}$. Therefore, $g^k=0$, implying no discontinuity at t_{k+1}. Similarly, if

$$\sum_{i=0}^{j-1}\gamma(k+i)<\frac{d}{D'},$$

then x^{k+i} is in the interior of C for $i=1,\ldots,j$. This implies the claim that there are no discontinuities in the interval $[t_k,t_k+d/D')$. Let $\Delta=d/2D'$. □

Let $K=\{k\ :\ g^k=0\}$. Let $\{X_l^k(\cdot)\}$ and $\{G_c^k(\cdot)\}$ converge in $D([0,T];\Re^n)$ to some $X(\cdot)$ and $G(\cdot)$, respectively, along a subsequence of K. (From the above proof, it is easy to see that K will be infinite: once k is large enough so that $\gamma(k)<\frac{d}{D'}$, each k with $g^k\neq 0$ will lead to $g^{k+1}=0$.) Then the limits must satisfy

$$X(t)=X(0)+\int_0^t h\bigl(X(\tau)\bigr)d\tau+G(t).$$

Furthermore, from the nature of our notion of convergence in $D([0,T];\Re^n)$, it is clear that $G(\cdot)$ is piecewise constant r.c.l.l. with $G(0)=0$ and that any two discontinuities of $G(\cdot)$ (hence of $X(\cdot)$) are separated by at least Δ on the time axis. Recall that for $x(\cdot)\in D([0,T];\Re^n)$, $x(t^+)=\lim_{t<s\to t} x(s)$ and $x(t^-)=\lim_{t>s\to t} x(s)$.

LEMMA 3.4. *We have $G(\cdot)\equiv 0$, implying $\dot{X}(t)=h(X(t))$.*

Proof. Let

$$\tau=\inf\{t>0\ :\ X(t^+)\neq X(t^-)\}.$$

By the right continuity at 0 and the fact that any two discontinuity points are separated by at least $\Delta>0$, it follows that $\tau>0$. Let $\|X(\tau^+)-X(\tau^-)\|=\delta>0$. Then, by our notion of convergence, we can find $\tau_k<\tau_k'$, $k\geq 0$, such that $\tau_k'-\tau_k\to 0$ and

$$\left\|X_l^{n(k)}(\tau_k')-X(\tau^+)\right\|\to 0, \tag{11}$$

$$\left\|X_l^{n(k)}(\tau_k)-X(\tau^-)\right\|\to 0. \tag{12}$$

Recall that $\|h(\cdot)\|$ is bounded on C and that $e^k(\cdot)$ and $W_l^k(\cdot)$ converge to 0 uniformly on compact sets. Also, any two discontinuities of $X_l^n(\cdot)$ must be at least Δ apart. Thus, for sufficiently large k, there must exist a $\hat{\tau}_k\in[\tau_k,\tau_k']$ such that

$$\left\|X_l^{n(k)}(\hat{\tau}_k)-X_l^{n(k)}(\hat{\tau}_k^-)\right\|\geq\frac{\delta}{2}.$$

But then $X_l^{n(k)}(\hat{\tau}_k^+)\in\partial B$, and $X_l^{n(k)}(\hat{\tau}_k^-)$ is not in the interior of C. Once again, using (11) and (12) and the fact that two discontinuities of $X_l^n(\cdot)$ must be at least Δ apart, we conclude that $X(\tau^+)\in\partial B$ and $X(\tau^-)\in\partial C$. But then $X(\cdot)$ satisfies $\dot{X}(t)=h(X(t))$ on $[0,\tau)$ (since $G(\cdot)\equiv 0$ on $[0,\tau)$), and therefore an interior trajectory of this ODE in C hits ∂C, a contradiction of our choice of C. (Since C is a level set of the Lyapunov function $V(\cdot)$, $h(\cdot)$ is transversal to ∂C everywhere and is directed towards the interior.) This contradiction proves that $G(\cdot)\equiv 0$. ∎

The preceding lemma allows us to prove the following proposition, the proof of which proceeds along standard lines; see, e.g., Kushner and Clark (1978), Benveniste, Metivier, and Priouret (1990).

PROPOSITION 3.1. *Let Assumptions* 3.1 *and* 3.2 *hold. The scaled synchronous algorithm* (6) *converges to x^*.*

3.2. Analysis of the scaled iteration-asynchronous case. The scaled version of the asynchronous algorithm of (2) is given by

$$\begin{aligned}\tilde{x}_i^{k+1}&=x_i^k+\gamma\big(\nu(k,i)\big)\left(h_i(x^k)+w_i^k\right)I(i\in Y^k),\\ x^{k+1}&=\Pi_{\mathcal{B},\mathcal{C}}(\tilde{x}^{k+1}).\end{aligned} \tag{13}$$

We confine ourselves to nonexpansive mappings with respect to the sup-norm. We also impose a further assumption on the stepsize. In particular, we will use the following assumptions in place of Assumption 3.1. We use $[a]$ to denote the integer part of a real number a.

ASSUMPTION 3.3. *The stepsizes $\gamma(k)$ are eventually nonincreasing and satisfy*

$$0<\gamma(k)\to 0,\qquad \sum_{k=0}^{\infty}\gamma(k)=\infty.$$

In addition, for all $\beta \in (0,1)$,

$$\sup_k \frac{\gamma([k\beta])}{\gamma(k)} < \infty$$

and

$$\lim_{k\to\infty} \frac{\sum_{m=0}^{[k\overline{\beta}]} \gamma(m)}{\sum_{m=0}^{k} \gamma(m)} = 1, \qquad \textit{uniformly in } \overline{\beta} \in [\beta, 1].$$

ASSUMPTION 3.4. *There exists a* $\Gamma > 0$ *such that for all* i

$$\liminf_{k\to\infty} \frac{1}{k+1}\nu(k,i) \geq \Gamma.$$

Furthermore, for all $T > 0$, *the limit*

$$\lim_{n\to\infty} \frac{\sum_{k=\nu(n,i)}^{\nu(m_T(n),i)} \gamma(k)}{\sum_{k=\nu(n,j)}^{\nu(m_T(n),j)} \gamma(k)}$$

exists for all i, j.

Theorem 3.2 of Borkar (1998) implies that the above limit will in fact be 1, a fact we use later. In addition, we change Assumption 3.2 to the following.

ASSUMPTION 3.2′. *For* $T, m_T(k)$ *as before,*

$$\lim_{k\to\infty} \sum_{m=k}^{m_T(k)} \gamma(m) w^{l(m)} = 0,$$

where $\{l(m)\}$ *is any increasing sequence of nonnegative integers satisfying* $l(m) \geq m$ *for all* m.

Examples of stepsizes that satisfy Assumption 3.3 include $\gamma(k) = 1/k$, $\gamma(k) = 1/(k \log k)$, etc., for $k \geq 2$, with suitable modifications for $k = 0, 1$. The essential meaning of Assumption 3.4 is that all components are updated comparably often.

Under Assumptions 3.2′, 3.3, and 3.4, the analysis closely mimics that of the synchronous case, except that the ODE-based convergence analysis of Kushner and Clark (1978) and Benveniste, Metivier, and Priouret (1990) is replaced by the corresponding analysis of Borkar (1998). In order to avoid undue repetition, we shall provide only a brief sketch. The key result of Borkar (1998) that is used here is briefly described in the appendix.

The first simplifying assumption that we make is that Y^k is a singleton for all k; i.e., only one component is updated at a time. This is justified as in Borkar (1998), the idea being that one unfolds a single iteration that updates d components, $d \geq 2$, into d iterations, in which each iteration updates a single component. There is, however, a complication in that this artificially introduces bounded delays; that is, the update of the ith component at time $k+1$ may use the value of the jth component updated not at time k, but at time $k-m$ for some $m \leq n$. These delays can be handled as in Borkar (1998). For simplicity of exposition, we ignore the delays here.

Thus we have $Y^k = \{\phi^k\}$, where ϕ^k is the index of the component updated at time k, and the iteration (13) is written as

$$x^{k+1} = x^k + D^k(h(x^k) + w^k) + g^k,$$

where

$$D^k = \mathrm{diag}\big[\gamma(\nu(k,1))I(\phi^k = 1), \ldots, \gamma(\nu(k,n))I(\phi^k = n)\big]$$

and

$$g^k = \Pi_{\mathcal{B},\mathcal{C}}\big[x^k + D^k(h(x^k) + w^k)\big] - \big[x^k + D^k(h(x^k) + w^k)\big].$$

Let us denote

$$\bar{\mu}^k = \big[I(\phi^k = 1), \ldots, I(\phi^k = n)\big]$$

and set $\bar{\gamma}(m,j) = \gamma(\nu(m,j))$, $\hat{\gamma}(m) = \bar{\gamma}(m, \phi^m)$, $t_0 = 0$, and $t_k = \sum_{m=0}^{k-1} \hat{\gamma}(m)$, $k \geq 1$. Let us define piecewise linear and piecewise constant processes as follows:

$$\mu(t) = \bar{\mu}^k \qquad \text{for } t \in [t_k, t_{k+1}),$$

$$X_c(t) = x^k \qquad \text{for } t \in [t_k, t_{k+1}),$$

$$G_c(t) = \sum_{m=0}^{k-1} g^m \qquad \text{for } t \in [t_k, t_{k+1}),$$

$$X_l(t) = \begin{cases} x^k & \text{for } t = t_k, \\ \big(1 - \frac{t-t_k}{\gamma(k)}\big)x^k + \frac{t-t_k}{\gamma(k)}\tilde{x}^{k+1} & \text{for } t \in [t_k, t_{k+1}), \end{cases}$$

where

$$\tilde{x}^{k+1} = x^k + D^k\big(h(x^k) + w^k\big),$$

$$W_l(t) = \begin{cases} \sum_{m=0}^{k-1} D^m w^m & \text{for } t = t_k, \\ \big(1 - \frac{t-t_k}{\gamma(\nu(k,\phi^k))}\big)W_l(t_k) + \frac{t-t_k}{\gamma(\nu(k,\phi^k))}W_l(t_{k+1}) & \text{for } t \in [t_k, t_{k+1}). \end{cases}$$

Define the corresponding left-shifted processes as follows, for $t \geq 0$:

$$X_l^k(t) = X_l(t + t_k),$$

$$X_c^k(t) = X_c^k(t + t_k),$$

$$W_l^k(t) = W_l(t + t_k) - W_l(t_k),$$

$$G_c^k(t) = G_c(t + t_k) - G_c(t_k),$$

$$\mu^k(t) = \mu(t + t_k).$$

For an n-dimensional probability vector $p = [p_1, \ldots, p_n]$, let $\mathrm{diag}(p)$ denote the diagonal matrix whose ith diagonal entry is p_i. Then, letting μ^* denote the uniform probability vector $[1/n, \ldots, 1/n]$, we have, for $t \geq 0$,

$$X_l^k(t) = X_l^k(0) + \int_0^t \mathrm{diag}(\mu^*)h\big(X_l^k(\tau)\big)d\tau + W_l^k(t) + G_c^k(t) + e^k(t) + \eta^k(t),$$

$$\eta^k(t) = \int_0^t \big(\mathrm{diag}(\mu^k(\tau)) - \mathrm{diag}(\mu^*)\big)h\big(X_l^k(\tau)\big)d\tau,$$

$$e^k(t) = \int_0^t \mathrm{diag}\big(\mu^k(\tau)\big)\big(h\big(X_c^k(\tau)\big) - h\big(X_l^k(\tau)\big)\big)d\tau.$$

The convergence of $\{W_l^k(\cdot)\}$ to 0 follows from Assumption 3.2′. Convergence of $\{e^k(\cdot)\}$ to 0 follows along the lines of the preceding subsection. The proof of Lemma 3.3 now goes through as before, with $D' = \sup_{z\in\mathcal{C}} \max_i |h_i(z)| + D$. We also have the following.

LEMMA 3.5. *For each $T > 0$,*

$$\lim_{k\to\infty} \sup_{t\in[0,T]} ||\eta^k(t)|| = 0.$$

Proof. As before, one verifies that the set $\{\{X_l^k(t), t \in [0,T]\}, k \geq 1\}$ is relatively compact in $D([0,T];\Re^n)$. Thus one may drop to a subsequence of $\{k\}$, denoted by $\{k\}$ again by abuse of notation, such that $X_l^k(\cdot) \to Z(\cdot)$ for some $Z(\cdot) \in D([0,T];\Re^n)$. Since the map $x(\cdot) \in D([0,T];\Re^n) \to x(t) \in \Re^n$ for any $t \in [0,T]$ is continuous at $z(\cdot)$ if $z(\cdot)$ is continuous at t (see Billingsley (1968, p. 121)), and also any $x(\cdot) \in D([0,T];\Re^n)$ has at most countably many points of discontinuity (see Borkar (1998, p. 119)), it follows that $X_l^k(t) \to Z(t)$ for almost every $t \in [0,T]$. By the dominated convergence theorem, one then has

$$\lim_{k\to\infty} \int_0^t \big(\mathrm{diag}(\mu^k(\tau)) - \mathrm{diag}(\mu^*)\big)\big(h(X_l^k(\tau)) - h(Z(\tau))\big)d\tau = 0.$$

Since the left-hand side (L.H.S.) has a bounded derivative in t, it is equicontinuous. It is clearly bounded for each fixed t. Thus a straightforward application of the Arzelà–Ascoli theorem shows that the above convergence is uniform in $t \in [0,T]$. Therefore the claim would follow if we show that

$$\lim_{k\to\infty} \int_0^t \big(\mathrm{diag}(\mu^k(\tau)) - \mathrm{diag}(\mu^*)\big)h(Z(\tau))d\tau = 0,$$

uniformly in $[0,T]$. The uniformity of convergence over $[0,T]$ will follow as before from the Arzelà–Ascoli theorem if we prove pointwise convergence on $[0,T]$. In turn, the latter follows if we show that for each t

$$\lim_{k\to\infty} \int_0^t \big(\mathrm{diag}(\mu^k(\tau)) - \mathrm{diag}(\mu^*)\big)f(\tau)d\tau = 0$$

for any $f \in L_2([0,T];\Re^n)$. Consider $\mu^k(\cdot)$, $k \geq 1$, as elements of the space $\mathcal{U}$ of measurable maps from $[0,\infty)$ to the space of probability vectors in $\Re^n$, with the coarsest topology that renders continuous the maps $\mu(\cdot) \in \mathcal{U} \to \int_0^t \langle \mu(s), f(s)\rangle ds$ for all $t > 0$ and f as above. It is easy to deduce from the Banach–Alaoglu theorem that $\mathcal{U}$ is compact metrizable. Let $\bar{\mu}(\cdot)$ be any limit point of $\{\mu^k(\cdot)\}$ in $\mathcal{U}$ as $k \to \infty$. It follows from Theorem 3.2 of Borkar (1998) that $\bar{\mu} = \mu^*$. The claim follows. □

The proof of Lemma 3.4 now goes through as before. Thus the asynchronous iterates, suitably interpolated, track the ODE $\dot{x}(t) = (1/n)h(x(t))$, which has the same qualitative behavior as $\dot{x}(t) = h(x(t))$—the difference is a mere time scaling. As in Borkar (1998), we then obtain the following proposition.

PROPOSITION 3.2. *Let Assumptions* 3.2′, 3.3, *and* 3.4 *hold. The scaled asynchronous algorithm* (6) *converges to* x^*.

The only difference with Borkar (1998) will be that we are dealing with the projected algorithm here; therefore we have to allow for discontinuous trajectories. But this can be dealt with exactly as in the synchronous case.

4. Convergence theorems for stochastic approximation. Now we consider the situation in which $\{w^k\}$ is a random noise sequence. Specifically, we assume that it is adapted to a family of increasing σ-fields $\{\mathcal{F}^{k+1}\}$ to which $\{x^{k+1}\}$ is also adapted and satisfies

$$E[w^k/\mathcal{F}^k] = 0$$

for $k \geq 1$. We strengthen Assumption 3.1 to include

$$\sum_0^\infty \gamma(k)^2 < \infty,$$

which is a standard assumption in stochastic approximation theory. We further assume, in place of Assumptions 3.1 and 3.2′, that

$$E[||w^k||^2/\mathcal{F}^k] \leq H(x^k)$$

for some continuous $H(\cdot)$. Assumptions 3.3, 3.4 remain as before. We shall refer to the modified Assumptions 3.1, 3.2 as Assumptions 3.1(m), 3.2(m), respectively.

Note that the only use of Assumption 3.2 has been to ensure that there exists a $\Delta > 0$ such that consecutive jump times of $X_l(\cdot)$ are at least Δ apart. However, this Δ can depend on sample path in the present case without affecting the proof in any way. Since we are seeking almost sure convergence, it suffices to show the following.

LEMMA 4.1. *There exists w.p.*1 *a (possibly sample path dependent)* Δ *with the above property.*

Proof. Suppose that the claim is not true for some sample path. Let $\{t^{m(k)}\}$ denote the successive jump times, with $+\infty$ being a possible value for these. (In particular, $t^{m(k)} = \infty$ for $k > k_0$ if there are only k_0 jumps.) Then for the sample path under consideration, these are all finite, and moreover, there exist consecutive jump times $t^{m(k(l)+1)} > t^{m(k(l))}$ such that $t^{m(k(l)+1)} - t^{m(k(l))} \to 0$ as $l \to \infty$. Let $K = \sup_{x\in\mathcal{C}}||h(x)||$. Since the iterates move from $\partial\mathcal{B}$ to $\partial\mathcal{C}$ between $(t^{m(k(l))})^+$ and $(t^{m(k(l)+1)})^-$, we must have

$$\left\| \sum_{i=m(k(l))}^{m(k(l)+1)-1} \gamma(i) w^i \right\| \geq d - \left(t^{m(k(l)+1)} - t^{m(k(l))} \right) K \geq \frac{d}{2}$$

for l sufficiently large. Letting Ψ^l denote the L.H.S. above, it then follows that $\Psi^l \geq \frac{d}{2}$ infinitely often (i.o.). We shall prove that

$$P\left(\Psi^l \geq \frac{d}{2}, \text{ i.o.} \right) = 0,$$

which will imply the desired claim. By the Chebyshev inequality, we have

$$\sum_k P\left(\psi^k \geq \frac{d}{2} \right) \leq \sum_k \frac{4E[||\sum_{i=m(k)}^{m(k+1)-1} \gamma(i) w^i||^2 I(t^{m(k)} < \infty)]}{d^2}.$$

Summing over k, the R.H.S. sums to a quantity bounded by

$$\frac{4\sum_i \gamma(i)^2 E[||w^i||^2]}{d^2} \leq \frac{4(\sum_i \gamma(i)^2)\sup_{x\in\mathcal{C}}|H(x)|}{d^2} < \infty,$$

in view of our hypotheses on $\{w^k\}$. The claim follows from the Borel–Cantelli lemma. □

Our hypotheses also ensure that Assumption 3.2 holds a.s. To see this, let $M^k = \sum_{i=0}^k \gamma(i)w^i$ for $i \geq 0$. Then $(M^k, \mathcal{F}^{k+1})$ is a square-integrable martingale. Its quadratic variation process is

$$\sum_{i=0}^k \gamma(i)^2 \left(E\left[\frac{||w^i||^2}{\mathcal{F}^{i-1}}\right] - \left\| E\left[\frac{w^i}{\mathcal{F}^{i-1}}\right]\right\|^2 \right),$$

which is bounded by the finite quantity $2\sup_{x\in\mathcal{C}}|H(x)|\sum_i \gamma(i)^2$. By Theorem 3.3.4, p. 53, of Borkar (1995), $\{M^k\}$ converges a.s. It then follows that Assumption 3.2 holds a.s. Hence we have the following counterpart of Proposition 3.1.

LEMMA 4.2. *Under the above hypotheses, the scaled synchronous algorithm* (5) *converges to* x^* *a.s.*

For the asynchronous case, note that $(\sum_{m=0}^k \gamma(\nu(m,i))w_i^m, \mathcal{F}^k)$ is a (square-integrable) martingale for each i. Considerations similar to those above then lead to the following stochastic counterpart of Proposition 3.2.

LEMMA 4.3. *Under the above hypotheses, the scaled asynchronous algorithm converges to* x^* *a.s.*

We now specialize to algorithms of the form

$$x^{k+1} = x^k + \gamma(k)(F(x^k, \xi^k) - x^k)$$

in synchronous form and

$$x_i^{k+1} = x_i^k + \gamma(\nu(k,i))(F_i(x^k, \xi^k) - x_i^k)I(i \in Y^k)$$

in asynchronous form, where $\{\xi^k\}$ is an independently and identically distributed (i.i.d.) stochastic noise sequence taking values in some measurable space, and the function $F(\cdot,\cdot)$ is assumed to satisfy the nonexpansivity property:

$$||F(x,u) - F(y,u)||_p \leq ||x-y||_p$$

for some $p \in (0,\infty]$ and all x, y, u. Let $T(x) = E[F(x,\xi^k)]$. Then

$$||T(x) - T(y)||_p \leq ||x-y||_p.$$

The aim is to find a fixed point x^* of $T(\cdot)$, i.e., a point x^* satisfying $x^* = T(x^*)$, which we assume to exist uniquely. Define $h(x) = T(x) - x$ and $w^k = F(x^k,\xi^k) - T(x^k)$, which casts this algorithm into the form analyzed above. Note, in particular, that in view of our hypotheses on F, $E[||w^k||^2/\mathcal{F}^k] \leq c(||x^k||^2 + 1)$ for some $c > 0$. The foregoing then leads to the following.

PROPOSITION 4.1. *Let* $\{x^k\}$ *be generated by the synchronous stochastic approximation algorithm* (1). *Let Assumptions* 3.1(m) *and* 3.2(m) *hold. Then the sequence* $\{x^k\}$ *converges to* x^* *w.p.*1.

Proof. The theorem is an application of Lemmas 2.1 and 4.2, the global asymptotic stability of the equilibrium x^* for the ODE $\dot{x}(t) = T(x(t)) - x(t)$ being proved in Borkar and Soumyanath (1997). □

PROPOSITION 4.2. *Let $\{x^k\}$ be generated by the asynchronous version of the above algorithm. Let Assumptions* 3.1(m)*,* 3.2(m)*,* 3.3*, and* 3.4 *hold with the modifications stated above. Then the sequence $\{x^k\}$ converges to x^* w.p.*1.

Proof. The theorem is an application of Lemmas 2.1 and 4.3, the global asymptotic stability of the ODE $\dot{x}(t) = (1/n)(T(x(t)) - x(t))$ being ensured as before by observing that the scalar $1/n$ on its R.H.S. represents a mere time scaling. □

5. Analysis of *Q*-learning algorithms. The convergence theorems above are directly applicable to the analysis of Q-learning algorithms for discounted and SSP dynamic programming problems. As discussed in Bertsekas (2001, Vol. 1), discounted cost problems can be formulated as SSP problems. We will therefore restrict ourselves to SSP problems. Here we have a controlled discrete-time dynamic system where at state i the use of a control u specifies the transition probability $p_{ij}(u)$ to the next state j. There are a finite number of states. At state i, the control u is constrained to take values from a given finite control set $U(i)$. The cost of using u at state i and moving to state j is denoted by $g(i, u, j)$. We assume that there is a special cost-free termination state 0. Once the system reaches that state, it remains there at no further cost; that is, $p_{00}(u) = 1$ for all u. We denote by $1, \ldots, n$ the states other than the termination state 0.

The total expected cost associated with an initial state i and a policy $\pi = \{\mu_0, \mu_1, \ldots\}$, where each μ_k maps states i into controls $\mu_k(i) \in U(i)$, is

$$J_\pi(i) = \lim_{N\to\infty} E\left\{\sum_{k=0}^{N} g\big(x_k, \mu_k(x_k), x_{k+1}\big) \;\Big|\; x_0 = i\right\}.$$

Note that the discounted cost problem with discount factor $\alpha \in (0, 1)$ and states $i = 1, \ldots, n$ is obtained as the special case of an SSP problem, where $p_{i0}(u) = 1 - \alpha$ and $g(i, u, 0) = 0$ for all $i = 1, \ldots, n$ and $u \in U(i)$.

A stationary policy is a policy of the form $\pi = \{\mu, \mu, \ldots\}$, and its corresponding cost function is denoted by $J_\mu(i)$. We call a stationary policy π *proper* if there exists an integer m such that

$$\max_{i=1,\ldots,n} P\{x_m \neq 0 \mid x_0 = i, \pi\} < 1,$$

and call π *improper* otherwise. We assume the following.

ASSUMPTION 5.1. *There exists at least one proper policy.*

ASSUMPTION 5.2. *Every improper policy results in infinite expected cost from at least one initial state.*

These assumptions, introduced by Bertsekas and Tsitsiklis (1991), have become standard in the analysis of SSP problems and are sufficient to show the validity of the major types of dynamic programming results. For example, the value iteration method converges to the optimal cost function J^*, which is the unique solution of Bellman's equation

$$J^*(i) = \min_{u\in U(i)} \sum_{j=0}^{n} p_{ij}(u)\big(g(i, u, j) + J^*(j)\big), \qquad i = 1, \ldots, n,$$

$$J^*(0) = 0.$$

Q-learning algorithms update estimates of the Q-factors, defined for all pairs (i,u) by

$$Q^*(i,u) = \sum_{j=0}^{n} p_{ij}(u)\big(g(i,u,j) + J^*(j)\big).$$

From this definition and Bellman's equation, we see that the Q-factors are the unique solution of the following system of equations:

$$Q(i,u) = \sum_{j=0}^{n} p_{ij}(u)\left(g(i,u,j) + \min_{v\in U(j)} Q(j,v)\right), \qquad i = 1,\dots,n,\ u \in U(i),$$

$$Q(0,u) = 0,$$

which may be viewed as Bellman's equation for Q-factors.

Let us generically denote by Q the vector of Q-factors. The synchronous version of Q-learning is given by

$$Q^{k+1} = Q^k + \gamma(k)\big(F(Q^k,\xi^k) - Q^k\big), \tag{14}$$

where $\{\xi^k\}$ is a sequence of independent vector-valued random variables taking the values $0,1,\dots,n$, with probabilities Prob $(\xi^k_{iu} = j) = p_{ij}(u)$ for all k,

$$F(Q,\xi)(i,u) = g(i,u,\xi_{iu}) + \min_{v\in U(\xi_{iu})} Q(\xi_{iu},v).$$

The initial condition is assumed to satisfy $Q^0(0,u) = 0$, which ensures that $Q^k(0,u) = 0$ for all k. Also, for $i = 0$, $\xi_{iu} = 0$ w.p.1. Thus $g(i,u,\xi_{iu}) = g(0,u,0) = 0$ (because 0 is a cost-free state) and $Q(\xi_{iu},u) = Q(0,u) = 0$ for all u. Thus $F(Q,\xi)(0,u) = 0$ for all u. In fact, this permits us to consider the iteration of $Q^k(i,u)$ for $1 \le i \le n$ alone, which we denote again by Q^k by abuse of notation. Define

$$T(Q)(i,u) = \sum_{j=1}^{n} p_{ij}(u)F(Q,j)$$

and

$$w^k = F(Q^k,\xi^k) - T(Q^k).$$

Assumption 3.2 applies to the stepsize $\gamma(k)$ and the noise w^k for the *rescaled* iterates.

The following two properties of the mapping T are significant for our purposes:

1. T is nonexpansive with respect to the sup-norm.
2. The unique fixed point Q^* of the mapping T is a global asymptotically stable equilibrium of the ODE $\dot{Q} = T(Q) - Q$.

Property 1 follows from the nonexpansiveness of F, which can be verified by noting that for all $Q_1, Q_2 \in \mathbb{R}^{n+m}$ we have

$$\begin{aligned} F(Q_1,\xi)(i,u) - F(Q_2,\xi)(i,u) &= \min_{u'} Q_1(\xi_{iu},u') - \min_{u'} Q_2(\xi_{iu},u') \\ &\le Q_1(\xi_{iu},u_2) - Q_2(\xi_{iu},u_2) \\ &\le \max_{(i,u)} |Q_1(i,u) - Q_2(i,u)| \\ &\le \|Q_1 - Q_2\|_\infty, \end{aligned}$$

where u_2 achieves the minimum in $\min_{u'} Q_2(\xi_{iu}, u')$. A symmetric argument shows that

$$F(Q_2, \xi)(i,u) - F(Q_1, \xi)(i,u) \le \|Q_1 - Q_2\|_\infty.$$

Property 2 follows from the analysis of Bellman's equation for SSP problems (see e.g., Bertsekas (2001, Vol. 2)), and from the analysis of ODE maps involving nonexpansive mappings in Borkar and Soumyanath (1997). Using the facts that Q^* is the unique fixed point of T and that T is nonexpansive, it follows that any solution trajectory $Q(t)$ converges to Q^*. Moreover, the analysis in Borkar and Soumyanath (1997) implies that $\|Q(t) - Q^*\|_\infty$ is nonincreasing, establishing that Q^* is a global asymptotically stable equilibrium point for the ODE.

The mapping F, in addition to being nonexpansive, satisfies

$$E\big[F(Q^k, \xi^k) \mid \mathcal{F}^k\big] = T(Q^k),$$

where

$$\mathcal{F}^k = \sigma(x^k, \ldots, x^0, \xi^{k-1}, \ldots, \xi^0).$$

The properties above are sufficient to show that all of the assumptions of Proposition 4.1 are satisfied, thus implying the following convergence result.

PROPOSITION 5.1. *The sequence* $\{Q^k\}$ *generated by the synchronous* Q*-learning iteration* (14) *converges to* Q^* *w.p.*1.

5.1. Analysis of the SSP asynchronous Q-learning. The asynchronous version of (14) is what is usually referred to as the Q-learning algorithm. It is written as

$$Q^{k+1}(i,u) = Q^k(i,u) + \gamma\big(\nu(k,\phi^k)\big)\big(F(Q^k,\xi^k)(i,u) - Q^k(i,u)\big)I\big((i,u) = \phi^k\big), \tag{15}$$

where $\{\xi^k\}$ is as defined above and $\{\phi^k\}$ is a random process. Again we impose Assumption 3.2′ on the stepsize, and we assume in addition that

1.
$$\liminf_{k\to\infty} \frac{1}{k+1}\nu(k,i,a) \ge \Delta \qquad \text{for some } \Delta > 0.$$

Furthermore, for all $T > 0$, the limit

$$\lim_{n\to\infty} \frac{\sum_{k=\nu(n,i,a)}^{\nu(m_T(n),i,a)} \gamma(k)}{\sum_{k=\nu(n,j,b)}^{\nu(m_T(n),j,b)} \gamma(k)}$$

exists w.p.1 for all i, j, a, b.

2. $\{\gamma(k)\}$ is as in Assumption 3.3.

Again the mapping F satisfies

$$E\big[F(Q^k, \xi^k) \mid \mathcal{F}^k\big] = T(Q^k),$$

with

$$\mathcal{F}^k = \sigma(x^k, \ldots, x^0, \xi^{k-1}, \ldots, \xi^0, \phi^k, \ldots, \phi^0).$$

Similarly, the assumptions of Proposition 4.2 are satisfied, and we have the following.

PROPOSITION 5.2. *The sequence* $\{Q^k\}$ *generated by the asynchronous* Q*-learning iteration* (15) *converges to* Q^* *w.p.*1.

As already mentioned, the case in which more than one component is updated at a time can be reduced to the one above, modulo bounded delays, which can be separately taken care of as in Borkar (1998).

Remark 5.1. The usual formalism for Q-learning algorithms (see, e.g., Bertsekas and Tsitsiklis (1996)) presupposes the availability of a simulation device that generates independent random variables $\{\xi^k\}$ as above. Alternatively, one may consider it as an on-line scheme where the samples are generated by a single simulation or actual run $\{X^k\}$ of the controlled Markov chain with the control process $\{Z^k\}$. Then it is asynchronous, with $\phi^k = (X^k, Z^k)$. The above framework still applies if we use the representation $X^{k+1} = f(X^k, Z^k, \xi^k)$, where $\{\xi^k\}$ are i.i.d. and f is a suitable map. Such a representation is always possible (albeit on a possibly augmented probability space) by the stochastic realization theoretic results of Borkar (1993). See Kifer (1986) for the uncontrolled case.

6. Some extensions. This section points out some important extensions of the preceding analysis. The first is an extension of Lemma 2.1. It is possible to replace the assumption of nonexpansivity with respect to a norm there by nonexpansivity with respect to the span seminorm $||\cdot||_s$, defined by

$$||x||_s = \max_{i=1,\ldots,n} x_i - \min_{i=1,\ldots,n} x_i,$$

where $x_1, \ldots, x_n$ are the components of x. In this case, however, a weaker boundedness result is obtained, which is the subject of the following lemma. This lemma is used crucially in our companion paper on Q-learning in average cost control (Abounadi, Bertsekas, and Borkar (2001)).

LEMMA 6.1. *Let* $\mathcal{B}$ *be an open and bounded subset of* $\Re^n$ *containing the origin, and let* $\mathcal{C}$ *be a subset of* $\Re^n$ *that contains* $\mathcal{B}$*. Consider the algorithm*

$$x^{k+1} = G^k(x^k, \xi^k),$$

where we assume the following:

1. $\{\xi^k\}$ *is a sequence in a measurable space* $(\Omega, \mathcal{F})$.
2. G^k *is nonexpansive in* x *with respect to the span seminorm; i.e., for every* $\xi \in \Omega$,

$$||G^k(x,\xi) - G^k(y,\xi)||_s \le ||x - y||_s \qquad \forall\ x, y, \xi.$$

3. *The sequence* $\{\tilde{x}^k\}$ *generated by the scaled iteration*

$$\tilde{x}^{k+1} = G^k(\Pi_{\mathcal{B},\mathcal{C}}(\tilde{x}^k), \xi^k), \qquad \tilde{x}^0 = x^0,$$

converges to some vector $x^* \in \mathcal{B}$.

Then $\{||x^k||_s\}$ *remain bounded.*

Proof. The proof is identical to that of Lemma 2.1. □

The second extension relates to the Q-learning schemes described above. One can also allow for random costs under mild technical conditions. Thus, let a real or simulated transition from i to j under control u at time k lead to a random cost ζ_{iuj}^{k+1}. We suppose that $E[\zeta_{iuj}^{k+1} \mid \mathcal{F}^{k+1}] = g(i,u,j)$ and $E[(\zeta_{iuj}^{k+1})^2 \mid \mathcal{F}^{k+1}] \le M$ w.p.1 for some constant $M < \infty$. (Compare with Remark 3.2.) Then the foregoing analysis

goes through exactly as before with one modification: the "martingale difference" sequence w^k gets replaced by $\hat{w}^k$, defined as follows: its (iu)th component is $\hat{w}^k_{iu} = w^k_{iu} + \zeta^{k+1}_{iu\xi^k_{iu}} - g(i,u,\xi^k_{iu})$, where w^k_{iu} is the (iu)th component of w^k. Note that $\{\hat{w}^k\}$ is also a martingale difference sequence. An example is the case in which $\zeta^{k+1}_{iuj} = g(i,u,j)+\psi^{k+1}_{iuj}$, where $\{\psi^n_{iuj}\}$ are i.i.d. zero mean, bounded variance random variables representing additive noise.

7. Conclusions. In this paper we have studied the convergence of synchronous and asynchronous algorithms involving nonexpansive maps and additive deterministic or stochastic noise. We have used the ODE approach, but we have dispensed with the restrictive boundedness assumption on the generated iterates that this approach requires. The nonexpansiveness property ensures that the distance between the iterates of two instantiations of the algorithm, driven by the same noise sequence and differing only in the initial conditions, remains bounded. In fact, our arguments will work for any algorithm for which this is true, and the associated ODE has a globally asymptotically stable equilibrium, under mild technical conditions on noise as above. As a special case of our analysis, we have discussed Q-learning algorithms for SSP problems, and we have refined the assumptions under which convergence can be proved. Our results used Lemma 2.1 for the boundedness argument. We can likewise use Lemma 6.1 to prove boundedness for certain Q-learning algorithms for the average cost dynamic programming problem. The analysis of these algorithms requires considerable additional machinery and is given separately in a companion paper (Abounadi, Bertsekas, and Borkar (2001)).

Appendix. Here we briefly recall the main results of Borkar (1998) that are used in the paper. Let $F(\cdot,\cdot) = [F_1(\cdot,\cdot),\ldots,F_d(\cdot,\cdot)]^T : \mathcal{R}^d \times \mathcal{R}^m \to \mathcal{R}^d$ be Lipschitz in its first argument uniformly w.r.t. the second. Consider the stochastic approximation algorithm of the form

$$x^{k+1} = x^k + \gamma(k)F(x^k,\xi^k), \quad k \geq 0,$$

for $x^k = [x^k_1,\ldots,x^k_d]$. Let $h(x) = E[F(x,\xi^1)]$. We assume that the ODE $\dot{x}(t) = h(x(t))$ has a globally asymptotically stable equilibrium x^*. The asynchronous version of this algorithm is given by

$$x^{k+1}_i = x^k_i + \gamma\big(\nu(k,i)\big)I(i \in Y^k)F_i(x_1^{k-\tau_{1i}(k)},\ldots,x_d^{k-\tau_{di}(k)},\xi^k), \quad 1 \leq i \leq d,$$

for $k \geq 0$, where

(1) $\{Y^k\}$ is a set-valued random process taking values in the subsets of the set $\{1,\ldots,d\}$, representing the components that do get updated at time k.
(2) $\{\tau_{ij}(k), 1 \leq i,j \leq d, k \geq 0\}$ are bounded random delays. One usually takes $\tau_{ii}(k) = 0$ for all i, though this is not necessary. (Borkar (1998) also relaxes the boundedness condition on delays to a conditional moment bound.)
(3) $\nu(k,i) = \sum_{m=0}^{k} I(i \in Y^m)$ denotes the number of times component i gets updated until time k.

Let Assumptions 3.1–3.4 hold. The main result of Borkar (1998) is the following.

THEOREM A.1. *If $\{x^k\}$ remain w.p.1 bounded, they converge to x^* w.p.1.*

We shall briefly describe what the proof entails, using the notation of section 3.2 above. The intuition behind why the bounded delays don't affect the asymptotics is simple. Recall that the passage from the discrete iteration to an interpolated "approximation to ODE" involves the time scaling $k \to t^k$. This scaling shrinks the

time axis more and more as k increases, because $t^{k+1} - t^k \to 0$. If K denotes a bound on the delays, the intervals $[k, k+1, \ldots, k+K]$ map to $[t^k, t^k + \sum_{m=k}^{k+K-1} \gamma(m)]$, which become smaller and smaller as k increases, because of which the delays as seen by the ODE approximation on the rescaled time become smaller and smaller, becoming asymptotically negligible. This intuition can be made precise quite easily. In fact, it simply contributes one additional asymptotically negligible error term to the usual ODE analysis of stochastic approximations. See Lemma 3.3 of Borkar (1998) for details.

The harder problem is to deal with the Y^k's, i.e., with the fact that not all components are getting updated at each step. As in section 3.2 above, one has $X_l^k(t) = X_l^k(0) + \int_0^t \mathrm{diag}(\mu^k(\tau))h(X_l^k(\tau))d\tau$ + error terms, the latter going to zero w.p.1 as $k \to \infty$. View $\mu^k(\cdot)$ as elements of the space of measurable maps $[0, \infty) \to \{d$-dimensional probability vectors$\}$, with the coarsest topology that renders continuous the maps $\mu(\cdot) \to \int_0^T \langle \mu(t), g(t) \rangle dt$ for any $T > 0$ and any $g : [0, T] \to \mathcal{R}^d$ that satisfies $\int_0^T ||g(t)||^2 dt < \infty$. (Recall Lemma 3.5 above.) This is a compact metrizable topology. Let $\mu^k(\cdot)$ converge along a subsequence to some $\hat{\mu}(\cdot)$ in this topology. Then the limiting trajectory of $X_l^k(\cdot)$ along this subsequence will satisfy the nonautonomous ODE

$$\dot{x}(t) = \mathrm{diag}(\hat{\mu}(t))h(x(t)).$$

The additional conditions on $\gamma(k)$ stipulated in Assumptions 3.3 and 3.4 are required to further ensure that $\hat{\mu}(t)$ in fact equals μ^* for almost every t. See Borkar (1998) for details.

One can, in fact, work with the nonautonomous ODE itself to draw the same conclusions by using Lemma 2.4 of Borkar (1998), the only requirement being that the components of $\hat{\mu}(t)$ remain uniformly bounded away from zero from below for almost every t. This is a weaker version of the statement "all components get updated comparably often." Unfortunately, no simple transparent sufficient condition to ensure this (short of Assumptions 3.3, 3.4) seems available.

Acknowledgment. Thanks are due to John Tsitsiklis, whose suggestions resulted in important simplifications of the lemmas in section 2.

REFERENCES

J. Abounadi, D. P. Bertsekas, and V. S. Borkar (2001), *Learning algorithms for Markov decision processes with average cost*, SIAM J. Control Optim., 40, pp. 681–698.

D. P. Bertsekas and J. N. Tsitsiklis (1989), *Parallel and Distributed Computation: Numerical Methods*, Prentice–Hall, Englewood Cliffs, NJ.

D. P. Bertsekas and J. N. Tsitsiklis (1991), *An analysis of stochastic shortest path problems*, Math. Oper. Res., 16, pp. 580–595.

D. P. Bertsekas and J. N. Tsitsiklis (1996), *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA.

D. P. Bertsekas (2001), *Dynamic Programming and Optimal Control*, 2nd ed., Athena Scientific, Belmont, MA.

A. Benveniste, M. Metivier, and P. Priouret (1990), *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, New York.

P. Billingsley (1968), *Convergence of Probability Measures*, John Wiley, New York.

V. S. Borkar (1993), *White noise representations in stochastic realization theory*, SIAM J. Control Optim., 31, pp. 1093–1102.

V. S. Borkar (1995), *Probability Theory: An Advanced Course*, Springer-Verlag, New York.

V. S. Borkar (1998), *Asynchronous stochastic approximations*, SIAM J. Control Optim., 36, pp. 840–851. Correction note in ibid, 38 (2000), pp. 662–663.

V. S. Borkar and S. P. Meyn (2000), *The O.D.E. method for convergence of stochastic approximation and reinforcement learning*, SIAM J. Control Optim., 38, pp. 447–469.

V. S. Borkar and K. Soumyanath (1997), *An analog parallel scheme for fixed point computation—Part* I*: Theory*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 44, pp. 351–355.

S. Csibi (1975), *Learning under computational constraints from weakly dependent samples*, Prob. Control Inform. Theory, 4, pp. 3–21.

L. Gerencsér (1992), *Rate of convergence of recursive estimators*, SIAM J. Control Optim., 30, pp. 1200–1227.

T. Jaakola, M. I. Jordan, and S. P. Singh (1994), *On the convergence of stochastic iterative dynamic programming algorithms*, Neural Computation, 6, pp. 1185–1201.

Y. Kifer (1986), *Ergodic Theory of Random Transformations*, Birkhäuser Boston, Cambridge, MA.

H. J. Kushner and D. S. Clark (1978), *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, New York.

H. J. Kushner and G. Yin (1997), *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, New York.

L. Ljung (1977), *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, 22, pp. 551–575.

K. Soumyanath and V. S. Borkar (1999), *An analog scheme for fixed point computation—Part* II*: Applications*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 46, pp. 442–451.

P. Tseng, D. P. Bertsekas, and J. N. Tsitsiklis (1990), *Partially asynchronous parallel algorithms for network flow and other problems*, SIAM J. Control Optim., 28, pp. 678–710.

J. N. Tsitsiklis (1994), *Asynchronous stochastic approximation and Q-learning*, Machine Learning, 16, pp. 185–202.

C. J. C. H. Watkins (1989), *Learning from delayed rewards*, Ph.D. thesis, Cambridge University, Cambridge, England.

C. J. C. H. Watkins and P. Dayan (1992), *Q-learning*, Machine Learning, 8, pp. 279–292.

F. W. Wilson (1969), *Smoothing derivatives of functions and applications*, Trans. Amer. Math. Soc., 139, pp. 413–428.

T. Yoshizawa (1966). *Stability Theory by Lyapunov's Second Method*, Mathematical Society of Japan, Tokyo.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 1, pp. 23–59

IDENTIFIABILITY AND WELL-POSEDNESS OF SHAPING-FILTER PARAMETERIZATIONS: A GLOBAL ANALYSIS APPROACH*

CHRISTOPHER I. BYRNES†, PER ENQVIST‡, AND ANDERS LINDQUIST‡

Abstract. In this paper, we study the well-posedness of the problems of determining shaping filters from combinations of finite windows of cepstral coefficients, covariance lags, or Markov parameters. For example, we determine whether there exists a shaping filter with a prescribed window of Markov parameters and a prescribed window of covariance lags. We show that several such problems are well-posed in the sense of Hadamard; that is, one can prove existence, uniqueness (identifiability), and continuous dependence of the model on the measurements. Our starting point is the global analysis of linear systems, where one studies an entire class of systems or models as a whole, and where one views measurements, such as covariance lags and cepstral coefficients or Markov parameters, from data as functions on the entire class. This enables one to pose such problems in a way that tools from calculus, optimization, geometry, and modern nonlinear analysis can be used to give a rigorous answer to such problems in an algorithm-independent fashion. In this language, we prove that a window of cepstral coefficients and a window of covariance coefficients yield a bona fide coordinate system on the space of shaping filters, thereby establishing existence, uniqueness, and smooth dependence of the model parameters on the measurements from data.

Key words. identifiability, parameterization, well-posedness, foliations, Carathéodory extension, spectral estimation, cepstrum

AMS subject classifications. 30E05, 42A15, 58C99, 93B29, 93E12

PII. S0363012900383077

1. Introduction.

It is common to model a (real, zero-mean) stationary process $\{y(t) \mid t \in \mathbb{Z}\}$ as a convolution

$$y(t) = \sum_{k=-\infty}^{t} w_{t-k} u_k$$

of an excitation signal $\{u(t) \mid t \in \mathbb{Z}\}$, which is a white noise, i.e., $\mathrm{E}\{u(t)u(s)\} = \delta_{ts}$, where δ_{ts} is one if $t = s$ and zero otherwise. In the language of systems and control, under suitable finiteness conditions this amounts to passing the white noise u through a linear filter with the transfer function $w(z)$ having the Laurent expansion

$$w(z) = \sum_{k=0}^{\infty} w_k z^{-k} \tag{1.1}$$

for all $z \geq 1$, thus obtaining the process y as the output, as depicted in Figure 1. In addition, we assume that $w_0 \neq 0$ and that $w(z)$ is a rational function, the latter assumption being the finiteness condition required in systems and control theory. Such a filter will be called a *shaping filter*, and the coefficients $w_0, w_1, w_2, \ldots$ will be called the *Markov parameters*.

*Received by the editors December 20, 2000; accepted for publication (in revised form) September 25, 2001; published electronically March 27, 2002. This research was supported in part by grants from AFOSR, TFR, the Göran Gustafsson Foundation, and Southwestern Bell.
http://www.siam.org/journals/sicon/41-1/38307.html

†Department of Systems Science and Mathematics, Washington University, St. Louis, MO 63130 (chrisbyrnes@seas.wustl.edu).

‡Division of Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology, 100 44 Stockholm, Sweden (pere@math.kth.se, alq@math.kth.se).

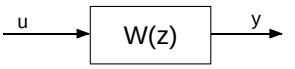


FIG. 1. *Representing a signal as the output of a black box.*

Clearly, any shaping filter must be *stable* in the sense that $w(z)$ has all of its poles in the open unit disc. To begin, we also assume that all zeros are located in the open unit disc. Such a shaping filter will be called a *minimum-phase shaping filter*.

Then the stationary stochastic process y has a rational spectral density

$$\Phi(e^{i\theta}) = |w(e^{i\theta})|^2,$$

which is positive for all θ. It is well known that the spectral density has a Fourier expansion

$$\Phi(e^{i\theta}) = r_0 + 2\sum_{k=1}^{\infty} r_k \cos k\theta,$$

where the Fourier coefficients

$$r_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\theta}\Phi(e^{i\theta})d\theta \tag{1.2}$$

are the covariance lags $r_k = \mathrm{E}\{y(t+k)y(t)\}$.

The spectral density $\Phi(z)$ is analytic in an annulus containing the unit circle and has there the representation

$$\Phi(z) = f(z) + f(z^{-1}),$$

where f is a rational function with all of its poles and zeros in the open unit disc. Hence, in particular, f is analytic outside the unit disc, and

$$f(z) = \frac{1}{2}r_0 + r_1 z^{-1} + r_2 z^{-2} + r_3 z^{-3} + \cdots . \tag{1.3}$$

Moreover,

$$\Phi(e^{i\theta}) = 2\mathrm{Re}\{f(e^{i\theta})\} > 0$$

for all θ, and, therefore, f is a real function which maps $\{|z| \geq 0\}$ into the right half-plane $\{\mathrm{Re}\, z > 0\}$; such a function is called *positive real*. For this to hold, the Toeplitz matrices

$$T_k = \begin{bmatrix} r_0 & r_1 & \cdots & r_k \\ r_1 & r_0 & \cdots & r_{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ r_k & r_{k-1} & \cdots & r_0 \end{bmatrix} \tag{1.4}$$

must be positive definite for $k = 0, 1, 2, \ldots$.

Another way of representing the distribution of the stationary process is via the so-called *cepstrum*

$$\log \Phi(e^{i\theta}) = c_0 + 2\sum_{k=1}^{\infty} c_k \cos k\theta. \tag{1.5}$$

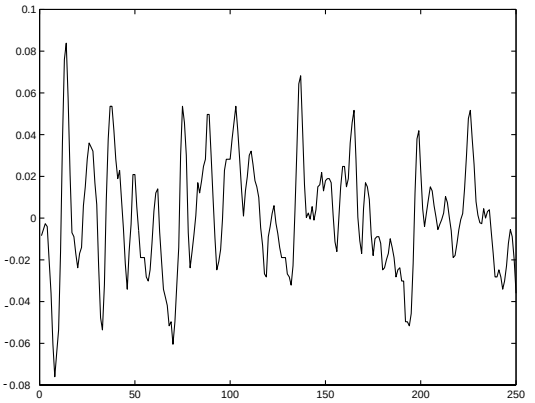


FIG. 2. *A frame of speech for the voiced nasal phoneme [ng].*

The Fourier coefficients

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\theta} \log \Phi(e^{i\theta}) d\theta \tag{1.6}$$

are known as the *cepstral coefficients*.

Finite windows of covariance lags and cepstral coefficients can be estimated from an observed data record

$$y_0, y_1, y_2, \ldots, y_N$$

of the process $\{y(t) \mid t \in \mathbb{Z}\}$. In fact, a limited number of covariance lags can be estimated via some ergodic estimate

$$r_k = \frac{1}{N+1-n} \sum_{t=0}^{N-n} y_{t+k} y_k. \tag{1.7}$$

However, we can only estimate

$$r_0, r_1, \ldots, r_n, \tag{1.8}$$

where $n << N$, with some precision. A complementary set of observables are given by the window

$$c_0, c_1, \ldots, c_n \tag{1.9}$$

of cepstral coefficients. One topic considered in this paper is to investigate the conditions under which these estimated coefficients can be used to determine minimum-phase shaping filters, i.e., to determine the identifiability of such shaping filters from covariance and cepstral windows.

As an example, to which we shall return several times in this paper, let us consider a 30 ms frame of speech from the voiced nasal phoneme [ng], depicted in Figure 2. Here $N = 250$, a typical sample length for a mobile telephone.

Figure 3 depicts a periodogram of this signal, i.e., a spectral estimate obtained by fast Fourier transform. This spectral estimate can be modeled as a smooth spectral envelope perturbed by contributions from an excitation signal. The spectral envelope

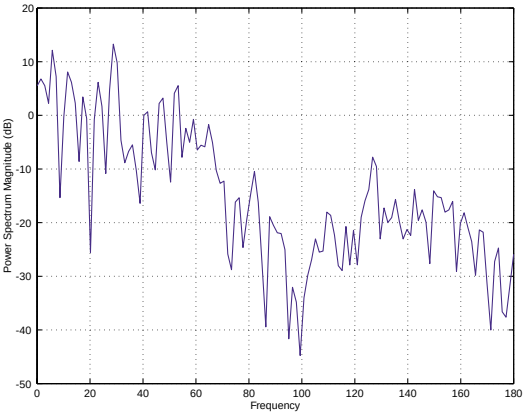


FIG. 3. *Periodogram for the voiced nasal phoneme [ng].*

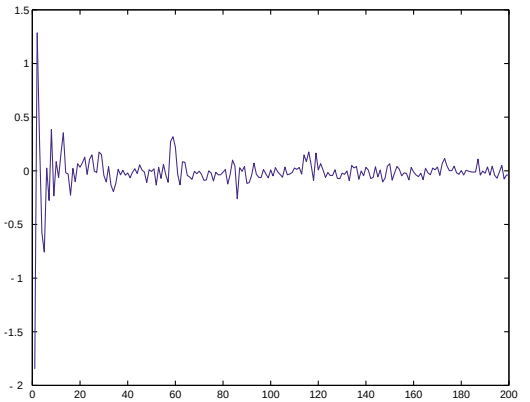


FIG. 4. *Cepstrum of voice speech signal.*

corresponds to the shaping of the vocal tract, which is described by the minimum-phase shaping filter.

As the Fourier transform of a convolution, the contributions of the shaping filter and the excitation signal to the spectral estimate are multiplicative. If we consider the logarithm of the spectral density Φ, the cepstrum, instead of Φ itself, the contribution of the excitation signal is additively superimposed on the that of the shaping filter.

Figure 4 shows the estimated cepstral coefficients of a frame of voiced speech. A contribution of the excitation signal is seen as spikes at multiples of the pitch period, corresponding to approximately $n_0 = 57$ in Figure 4. The spectral envelope can be estimated from a finite window

$$c_0, c_1, \dots, c_n \tag{1.10}$$

of cepstral coefficients, where $n < n_0$.

For minimum-phase shaping filters, the cepstral coefficients used in signal processing are closely related to the Markov parameters $w_0, w_1, w_2, \dots$ defined by (1.1). In more general systems problems, the minimum-phase requirement is relaxed to allow

σ to be an arbitrary (monic) polynomial. In this case, a record

$$w_0, w_1, \ldots, w_n \tag{1.11}$$

of Markov parameters are typically determined from the impulse response of an underlying system and not from data such as a finite time series, and for this reason Markov parameters can be quite useful in model reduction problems, starting from an underlying system. Nonetheless, for minimum-phase shaping filters, the cepstral coefficients used in signal processing are closely related to the Markov parameters of the shaping filter $w(z)$. Indeed, in section 6, we shall see that there is a one-to-one correspondence between windows of cepstral and Markov parameters of the same length.

In this paper, we are interested in the mathematical nature of the transformation of measurements, such as covariance lags and cepstral coefficients or Markov parameters, from data into the parameters of systems which produce such data. Our starting point will be the global analysis of linear systems, where one studies an entire class of systems or models as a whole and where one views measurements from data or model parameters as functions on the entire class. This point of view has been pioneered in [2, 4, 27, 16, 24]; see [5] for a survey. The central issue is whether the transformation from a set of measurements, viewed as functions, to a set of model parameters is well-posed, for example, in the sense of Hadamard. To be more precise, suppose the class of models is the class of (minimum-phase) shaping filters of bounded degree. This class can be viewed as a smooth manifold, for which any such shaping filter may be viewed as a point, and on which the coefficients of the numerator and denominator polynomials are a bona fide system of smooth coordinates on the global geometrization of this class of shaping filters. Matters being so, one can now ask, for example, whether a window of cepstral coefficients and a window of covariance coefficients also yield a bona fide coordinate system, so that, for example, the change of coordinates is a transformation which is smooth, one-to-one, onto, and with a smooth inverse. That is, the problem of passing from such data to models is indeed well-posed. Global analysis enables one to pose such problems in a way that tools from calculus, optimization, geometry, and modern nonlinear analysis can be used to give a rigorous answer to such problems.

In the next section, we shall review some of the basic spaces of systems we will use in our global analysis of certain transformations from data to models. In section 3, we will state our principal results, which we then prove in the following sections. These results focus on the identifiability of the models from collections of partial windows of covariance lags, cepstral coefficients, and Markov parameters and the questions of whether these parameters can be used to smoothly coordinatize spaces of shaping filters. For example, in section 4, a partial window of covariance lags and a partial window of cepstral coefficients are shown to jointly provide a system of local coordinates for shaping filters in the context of the geometry of certain foliations on the space of positive real functions.

In section 5, we prove that these are global coordinates, using methods from convex optimization theory. These schemes begin with an extension of the maximum entropy method, from the classical case of maximizing the zeroth cepstral gain to the problem of maximizing a "positive" linear combination of the entire partial cepstral window. This gives a new primal problem whose dual solves the rational covariance extension problem. In section 6, we provide a fairly complete local and global analysis of the use of a partial window of covariance lags and a partial window of cepstral

coefficients. In lieu of a convex optimization argument, we used an extension of the solution to the rational covariance extension problem and the Lefschetz fixed point theorem as a generalization of the Brower fixed point theorem for the spaces of Schur polynomials. We conclude the paper in section 7 with a discussion and illustrations of the applications of some of these constructions to speech synthesis.

2. Some geometric representations of classes of models. Suppose the positive real function f is given by

$$f(z) = \frac{1}{2}\frac{a(z)}{b(z)}, \tag{2.1}$$

where

$$\begin{aligned} a(z) &= a_0 z^n + a_1 z^{n-1} + \cdots + a_n, \\ b(z) &= b_0 z^n + b_1 z^{n-1} + \cdots + b_n \end{aligned}$$

are (real) polynomials of degree n. Clearly, a_0 and b_0 must have the same sign. We assume that they are both positive. Then, since

$$f(z) + f(z^{-1}) = w(z)w(z^{-1}),$$

we must have

$$w(z) = \frac{\sigma(z)}{a(z)}, \tag{2.2}$$

where

$$\sigma(z) = \sigma_0 z^n + \sigma_1 z^{n-1} + \cdots + \sigma_n$$

is the unique polynomial with all roots in the open unit disc satisfying

$$\sigma(z)\sigma(z^{-1}) = \frac{1}{2}[a(z)b(z^{-1}) + a(z^{-1})b(z)] \tag{2.3}$$

and $\sigma_0 > 0$. We shall denote the class of such polynomials by $\hat{\mathcal{S}}_n$, and we shall denote the n-dimensional submanifold of monic (Schur) polynomials in $\hat{\mathcal{S}}_n$ by $\mathcal{S}_n$. Now, in order for f to be positive real, the pseudopolynomial

$$a(z)b(z^{-1}) + a(z^{-1})b(z)$$

must be positive on the unit circle, and $a(z)$ must belong to $\hat{\mathcal{S}}_n$. Then $b(z)$ also must belong to $\hat{\mathcal{S}}_n$.

Clearly, it is no restriction to take $a \in \mathcal{S}_n$ in (2.3). For each such $a(z)$, let

$$S(a)v = a(z)v(z^{-1}) + a(z^{-1})v(z)$$

define an operator $S(a) : V_n \to \mathcal{Z}_n$ from the vector space V_n of polynomials having degree less than or equal to n into the vector space $\mathcal{Z}_n$ of pseudopolynomials of degree at most n. Then (2.3) may be written as

$$S(a)b = 2\sigma\sigma^*, \tag{2.4}$$

where $\sigma^*(z) := \sigma(z^{-1})$. Now it is well known that $S(a)$ is bijective when $a \in \mathcal{S}_n$ (see, e.g., [8, p. 760]), and hence (2.4) establishes a one-to-one correspondence between f and w. We may normalize this relation by taking either $b(z)$ or $\sigma(z)$, but not both, in $\mathcal{S}_n$.

The normalization $b_0 = 1$ corresponds to taking $r_0 = 1$ in (1.3). We denote by $\mathcal{P}_n$ the set of all $(a, b) \in \mathcal{S}_n \times \mathcal{S}_n$ such that (2.1) is positive real. We know [13] that $\mathcal{P}_n$ is a smooth, connected, real manifold of dimension $2n$ and that it is diffeomorphic to $\mathbb{R}^{2n}$.

Choosing instead the normalization $\sigma \in \mathcal{S}_n$, corresponding to setting $w_0 = 1$ in (2.2) and $c_0 = 0$ in (1.5), we obtain an alternative coordinatization of $\mathcal{P}_n$ in terms of (a, σ). In fact, for each $(a, \sigma) \in \mathcal{S}_n \times \mathcal{S}_n$, we obtain the corresponding $(a, b) \in \mathcal{P}_n$ by dividing $b = 2S(a)^{-1}(\sigma\sigma^*)$ by b_0, thus normalizing it to form a monic b. This is a diffeomorphism, establishing that $\mathcal{P}_n$ is diffeomorphic to $\mathcal{S}_n \times \mathcal{S}_n$. In fact, the inverse of this coordinate transformation is the stable spectral factorization of $\frac{1}{2}S(a)b$ followed by the normalization of $\sigma(z)$. Since $\mathcal{S}_n$ is diffeomorphic to $\mathbb{R}^n$ (see Appendix A), spectral factorization gives an alternative method of exhibiting a diffeomorphism between $\mathcal{P}_n$ and $\mathbb{R}^{2n}$.

We shall generally use (a, σ)-coordinates to describe the geometry of $\mathcal{P}_n$. This normalizes the cepstral window (1.10) and the Markov window (1.11), fixing c_0 at zero and w_0 at one. However, a covariance window which is normalized in (a, b)-coordinates will not be normalized in (a, σ)-coordinates, and hence, to avoid increasing the dimension of the problem, we shall need to consider instead the *normalized covariance lags*

$$r_k = \frac{\int_{-\pi}^{\pi} e^{ik\theta}\Phi(e^{i\theta})d\theta}{\int_{-\pi}^{\pi} \Phi(e^{i\theta})d\theta}, \quad k = 1, 2, \ldots, n, \tag{2.5}$$

when working in (a, σ)-coordinates. In fact, in all of these descriptions, the polynomials $a(z)$, $b(z)$, and $\sigma(z)$ are monic. Working with unnormalized covariance lags (1.2), as we shall occasionally do, requires an extra parameter, bringing the number of coordinates to $2n + 1$.

There are several other spaces of models which we will need in this analysis. We denote by $\mathcal{P}_n^*$ the (dense) open subspace of $\mathcal{P}_n$ consisting of those pairs (a, σ) of polynomials which are coprime. Following the arguments in Appendix A, we see that $\mathcal{P}_n$ is diffeomorphic to the space of coprime pairs of real monic polynomials of degree n with poles and zeros in $\mathbb{C}$, first studied in [4] using the notation $\mathrm{Rat}(n)$. The space $\mathrm{Rat}(n)$ is a $2n$-dimensional manifold with $n + 1$ path-connected components, some of which have a rather complicated topology (see [4, 34, 37, 6]).

We shall also need to study the space Π_n of real, monic, degree n-polynomials, which is, of course, diffeomorphic to $\mathbb{R}^n$. Our interest in this space comes from the Markov expansion (1.11), where we take σ to be in Π_n and a to be in $\mathcal{S}_n$. Consequently, we allow (a, σ) to vary over the larger space

$$\mathcal{Q}_n := \mathcal{S}_n \times \Pi_n.$$

We shall also need to consider the space $\mathcal{Q}_n^*$, the (dense) open subspace of $\mathcal{Q}_n$ consisting of those pairs (a, σ) of polynomials which are coprime.

3. Main results. Our first results show that it is possible to parameterize minimum-phase shaping filters in terms of a window of cepstral coefficients and a

window of covariance lags, both of which can be estimated from data. It is tempting, of course, to argue the plausibility of this result by counting parameters. This method typically works only when there is a rigorous way to compute the dimension of some geometric object—in this case, the smooth $2n$-dimensional manifold $\mathcal{P}_n$. In this setting, the implicit function theorem enables one to compute dimensions by computing the rank of certain Jacobian matrices or, equivalently, the linear independence of differentials. The following theorem is proved in section 4 (see Remark 4.7).

THEOREM 3.1. *The normalized covariance lags $r_1, r_2, \dots, r_n$ and the cepstral coefficients $c_1, c_2, \dots, c_n$ form a bona fide smooth coordinate system on the open subset $\mathcal{P}_n^*$ of $\mathcal{P}_n$; i.e., the map from $\mathcal{P}_n^*$ to $\mathbb{R}^{2n}$ with components $(r_1, r_2, \dots, r_n, c_1, c_2, \dots, c_n)$ has an everywhere invertible Jacobian matrix.*

Accordingly, when viewed as functions on $\mathcal{P}_n^*$, $(r_1, r_2, \dots, r_n, c_1, c_2, \dots, c_n)$ form local coordinates for the space $\mathcal{P}_n^*$ of pole-zero filters of degree n. At this point, one might hope to be able to use a global inverse function theorem, such as Hadamard's theorem, to show that these data define a global coordinate system. In part because of the complicated topology of $\mathcal{P}_n^*$, this is not possible, and instead we use a convex optimization scheme to conclude one of the important features of a global inverse function theorem. Indeed, the very nontrivial consequence of our next observation, to be proved in section 5, is that there is a one-to-one correspondence between the $2n$ coefficients $r_1, r_2, \dots, r_n, c_1, c_2, \dots, c_n$ of the minimum-phase shaping filter (2.2) and the $2n$ coefficients $a_1, a_2, \dots, a_n, \sigma_1, \sigma_2, \dots, \sigma_n$ of the denominator and numerator polynomials of (2.2), provided the degree of w is *exactly* n.

THEOREM 3.2. *Each shaping filter in $\mathcal{P}_n^*$ determines and is uniquely determined by its window $r_1, r_2, \dots, r_n$ of normalized covariance lags and its window $c_1, c_2, \dots, c_n$ of cepstral coefficients.*

As we have indicated, uniqueness follows from the remarkable fact that such a modeling filter arises as the minimum of a (strictly) convex optimization problem (see section 5). This optimization problem has, of course, antecedents in the literature, beginning with maximum entropy methods. Recall that linear predictive coding (LPC) is the most common method for determining shaping filters in signal processing. Given the window of (unnormalized) covariance data

$$r_0, r_1, \dots, r_n \tag{3.1}$$

with a positive definite Toeplitz matrix T_n, find the (unnormalized) shaping filter $w(z)$ and the corresponding spectral density

$$\Phi(e^{i\theta}) = |w(e^{i\theta})|^2,$$

which maximizes the entropy gain

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \log \Phi(e^{i\theta}) d\theta, \tag{3.2}$$

subject to the covariance-matching condition

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\theta}\Phi(e^{i\theta}) d\theta = r_k, \quad k = 0, 1, \dots, n. \tag{3.3}$$

For this reason, the LPC filter is often called the *maximum entropy filter*.

Now observe that the entropy gain (3.2) is precisely the zeroth cepstral coefficient

$$c_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log \Phi(e^{i\theta}) d\theta.$$

However, in cepstral analysis, one is interested not only in c_0 but in a finite window

$$c_0, c_1, \dots, c_n \tag{3.4}$$

of cepstral coefficients. It is natural, therefore, to maximize instead some (positive) linear combination

$$p_0c_0 + p_1c_1 + \cdots + p_nc_n \tag{3.5}$$

of the cepstral coefficients in the window (3.4). In view of (1.6), this may be written as a generalized entropy gain

$$\mathbb{I}_P(f) = \frac{1}{2\pi}\int_{-\pi}^{\pi} P(e^{i\theta}) \log \Phi(e^{i\theta}) d\theta, \tag{3.6}$$

where P is the symmetric pseudopolynomial

$$P(z) = p_0 + \frac{1}{2}p_1(z + z^{-1}) + \cdots + \frac{1}{2}p_n(z^n + z^{-n}), \tag{3.7}$$

and f is the positive real part (1.3) of Φ. We shall say that $P \in \mathcal{D}$ if P is nonnegative on the unit circle and $P \in \mathcal{D}_+$ if it is positive there. We note that the covariance matching condition (3.3) becomes

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\theta}\Phi(e^{i\theta}) d\theta = r_k, \quad k = 0, 1, \dots, n, \tag{3.8}$$

in terms of $\Phi(e^{i\theta}) = |w(e^{i\theta})|^2$.

Indeed, in section 5, we show that the problem of maximizing (3.5) subject to (3.8) has a finite solution only if the pseudopolynomial (3.7) belongs to $\mathcal{D}$. Indeed, if $P \in \mathcal{D}_+$, there is a unique solution Φ, and this solution has the form

$$\Phi(z) = \frac{P(z)}{Q(z)},$$

where

$$Q(z) = q_0 + \tfrac{1}{2}q_1(z + z^{-1}) + \cdots + \tfrac{1}{2}q_n(z^n + z^{-n})$$

belongs to $\mathcal{D}_+$.

In particular, we see that if we take P to be

$$P(z) = \sigma(z)\sigma(z^{-1})$$

and let $a(z)$ be the unique stable polynomial satisfying

$$Q(z) = a(z)a(z^{-1}),$$

then we have also determined the unique shaping filter (2.2) that matches the covariance data (3.1). Hence we have an alternative proof of the following result, first appearing in [11].

THEOREM 3.3. *Let $r_0, r_1, \dots, r_n$ be a partial covariance sequence, i.e., real numbers such that the Toeplitz matrix* (1.4) *is positive definite. Then, to any stable polynomial*

$$\sigma(z) = z^n + \sigma_1 z^{n-1} + \cdots + \sigma_{n-1}z + \sigma_n$$

of degree n, there corresponds a unique real stable polynomial

$$a(z) = a_0 z^n + a_1 z^{n-1} + \cdots + a_{n-1} z + a_n$$

of degree n such that

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\theta} \left| \frac{\sigma(e^{i\theta})}{a(e^{i\theta})} \right|^2 d\theta = r_k, \quad k = 0, 1, \ldots, n. \tag{3.9}$$

Theorem 3.3 was conjectured by Georgiou [21] as a solution to the partial covariance extension problem posed by Kalman [25]. Georgiou had already established the existence part, but a complete proof of the conjecture was given much later in [11]. Similarly, in [11], we also showed the following theorem.

THEOREM 3.4. *The normalized covariance lags $r_1, r_2, \ldots, r_n$ and the zero coefficients $\sigma_1, \sigma_2, \ldots, \sigma_n$ form a bona fide smooth coordinate system on the open manifold $\mathcal{P}_n$; i.e., the map from $\mathcal{P}_n$ to $\mathbb{R}^{2n}$ with components $(r_1, r_2, \ldots, r_n, \sigma_1, \sigma_2, \ldots, \sigma_n)$ has an everywhere invertible Jacobian matrix.*

In section 6, we derive the following results for coordinatization by covariance data and Markov parameters.

THEOREM 3.5. *The normalized covariance lags $r_1, r_2, \ldots, r_n$ and the normalized Markov parameters $w_1, w_2, \cdots, w_n$ form a bona fide smooth coordinate system on $\mathcal{Q}_n^*$; i.e., the map from $\mathcal{Q}_n^*$ to $\mathbb{R}^{2n}$ with components $(r_1, r_2, \ldots, r_n, w_1, w_2, \ldots, w_n)$ has an everywhere invertible Jacobian matrix. For each choice of a covariance window and a Markov window, there exists exactly one shaping filter matching these windows.*

The last statement of this theorem is related to a class of results found in the literature on Q-Markov covers (see, e.g., [31, 29, 1]). Allowing windows of Markov parameters for which $w_0 = 0$, as in the literature cited above, would only add filters $w(z)$, which can be recovered from those of Theorem 3.5 by multiplying $w(z)$ by some power of z^{-1}.

4. Global analysis on $\mathcal{P}_n$. We choose to represent minimum-phase shaping filters (2.2) by a pair $(a, \sigma) \in \mathcal{S}_n \times \mathcal{S}_n$. This imposes the normalization discussed in section 2. There is a geometric manifestation of the fact that (a, σ) are smooth coordinates on $\mathcal{P}_n$, which we will use to show that the cepstral and covariance windows also form bona fide coordinate systems. First note that tangent vectors to $\mathcal{P}_n$ at (a, σ) may be represented as a perturbation $(a + \epsilon u, \sigma + \epsilon v)$, where u, v are polynomials of degree less than or equal to $n - 1$. If, as before, we denote the real vector space of polynomials of degree less than or equal to d by V_d, then the tangent space to $\mathcal{P}_n$ at a point (a, σ) is canonically isomorphic to $V_{n-1} \times V_{n-1}$.

Now, for $a \in \mathcal{S}_n$, define $\mathcal{P}_n(a)$ to be the space of all points in $\mathcal{P}_n$ with the polynomial a fixed. If we define $\mathcal{P}_n(\sigma)$ analogously, then $\mathcal{P}_n(a)$ and $\mathcal{P}_n(\sigma)$ are real, smooth, connected n-manifolds. In fact, both are clearly diffeomorphic to $\mathcal{S}_n$ and hence to $\mathbb{R}^n$ [7] (see also Appendix A). The tangent space to the submanifold $\mathcal{P}_n(a)$ at a point (a, σ) is, therefore,

$$T_{(a,\sigma)}\mathcal{P}_n(a) = \{(u, v) \in V_{n-1} \times V_{n-1} \mid u = 0\}.$$

Similarly, the tangent space to $\mathcal{P}_n(\sigma)$ is given by

$$T_{(a,\sigma)}\mathcal{P}_n(\sigma) = \{(u, v) \in V_{n-1} \times V_{n-1} \mid v = 0\}.$$

Now the n-manifolds $\{\mathcal{P}_n(a) \mid a \in \mathcal{S}_n\}$ form the leaves of a foliation of $\mathcal{P}_n$, as do the n-manifolds $\{\mathcal{P}_n(\sigma) \mid \sigma \in \mathcal{S}_n\}$. Moreover, these two foliations are complementary in

the sense that if a leaf of one intersects a leaf of the other, the tangent spaces intersect in just $(0,0)$. This transversality property is equivalent to the fact that the functions (a,σ) form a local system of coordinates.

We now turn to the cepstral functions and the covariance functions. Let $g : \mathcal{P}_n \to \mathbb{R}^n$ be the map which sends (a,σ) to the vector $c \in \mathbb{R}^n$ with components

$$c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\theta}\log|w(e^{i\theta})|^2 d\theta, \quad k = 1,2,\ldots,n, \tag{4.1}$$

and let $\mathcal{C}_n := g(\mathcal{P}_n)$. Moreover, for each $c \in \mathcal{C}_n$, define the subset

$$\mathcal{P}_n(c) = g^{-1}(c).$$

We wish to show that $\mathcal{P}_n(c)$ is a smooth submanifold of dimension n. To this end, we will need to compute the Jacobian matrix of g, evaluated at tangent vectors to a point $(a,\sigma) \in \mathcal{P}_n$.

Thus, for each component

$$g_k(a,\sigma) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\theta}\log\left|\frac{\sigma(e^{i\theta})}{a(e^{i\theta})}\right|^2 d\theta$$

of g, we form the directional derivative

$$D_{(u,v)}g_k(a,\sigma) = \lim_{\epsilon\to 0}\frac{1}{\epsilon}\left[g_k(a+\epsilon u,\sigma+\epsilon v) - g_k(a,\sigma)\right]$$

in the direction $(u,v) \in V_{n-1}\times V_{n-1}$. A straightforward calculation yields

$$D_{(u,v)}g_k(a,\sigma) = \frac{1}{2\pi}\int_{-\pi}^{\pi} 2\mathrm{Re}\left\{\frac{v(e^{i\theta})}{\sigma(e^{i\theta})} - \frac{u(e^{i\theta})}{a(e^{i\theta})}\right\} e^{ik\theta}d\theta \tag{4.2}$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\frac{S(\sigma)v}{\sigma\sigma^*} - \frac{S(a)u}{aa^*}\right] e^{ik\theta}d\theta. \tag{4.3}$$

Now, for any $\varphi \in \mathcal{S}_n$, define the linear map $G_\varphi : V_{n-1} \to \mathbb{R}^n$ by

$$G_\varphi u = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{S(\varphi)u}{\varphi\varphi^*}\begin{bmatrix} e^{i\theta} \\ e^{i2\theta} \\ \vdots \\ e^{in\theta}\end{bmatrix} d\theta.$$

Then the kernel of the Jacobian of g at (a,σ) is given by

$$\ker \mathrm{Jac}(g)|_{(a,\sigma)} = \{(u,v) \mid G_\sigma v = G_a u\}. \tag{4.4}$$

LEMMA 4.1. *The linear map G_φ is a bijection.*

Proof. Suppose that $G_\varphi u = 0$. Then

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{S(\varphi)u}{\varphi\varphi^*}e^{ik\theta}d\theta = 0 \tag{4.5}$$

for $k = 1,2,\ldots,n$. By symmetry this also holds for $k = -1,-2,\ldots,-n$. Moreover, since

$$\frac{S(\varphi)u}{\varphi\varphi^*}(z) = \frac{u(z)}{\varphi(z)} + \frac{u(z^{-1})}{\varphi(z^{-1})}$$

and $\frac{u(z)}{\varphi(z)}$ is strictly proper and analytic for $|z| \geq 1$, (4.5) holds for $k = 0$ also so that integration against $\frac{S(\varphi)u}{\varphi\varphi^*}$ annihilates all trigonometric pseudopolynomials of degree at most n. In particular, we obtain

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{S(\varphi)u}{\varphi}\right|^2 d\theta = 0,$$

which in turn yields $S(\varphi)u = 0$. But $S(\varphi)$ is nonsingular, and hence $u = 0$, establishing injectivity of G_φ. However, since the range and domain of G_φ are the same dimension, namely n, the map is also surjective. □

PROPOSITION 4.2. *For each $c \in \mathcal{C}_n$, the space $\mathcal{P}_n(c)$ is a smooth n-manifold. The tangent space $T_{(a,\sigma)}\mathcal{P}_n(c)$ at (a,σ) consists of precisely all $(u,v) \in V_{n-1} \times V_{n-1}$ such that*

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{S(\sigma)v}{\sigma\sigma^*}e^{ik\theta}d\theta = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{S(a)u}{aa^*}e^{ik\theta}d\theta \tag{4.6}$$

for $k = 0, 1, \dots, n$.

Proof. The tangent vectors of $\mathcal{P}_n(c)$ at (a,σ) are precisely the vectors in the null space of the Jacobian of g at (a,σ), as computed above. Consequently, by (4.4), (4.6) holds for $k = 1, 2, \dots, n$. However, as pointed out in the proof of Lemma 4.1, (4.5) holds for $k = 0$, and hence (4.6) holds for $k = 0$ also. Moreover, by (4.4) and Lemma 4.1, the tangent space has dimension n. Therefore, the rank of $\mathrm{Jac}(g)|_{(a,\sigma)}$ is full, and the rest of the claim follows from the implicit function theorem. □

Because the rank of $\mathrm{Jac}(g)|_{(a,\sigma)}$ is everywhere n, the connected components of the submanifolds $\mathcal{P}_n(c)$ form the leaves of a foliation of $\mathcal{P}_n$. However, according to Lemma C.1, the submanifolds $\mathcal{P}_n(c)$ are themselves connected.

PROPOSITION 4.3. *The n-manifolds $\{\mathcal{P}_n(c) \mid c \in \mathcal{C}_n\}$ are connected and hence form the leaves of a foliation of $\mathcal{P}_n$.*

As an example of the more involved calculation we shall next undertake with the covariance window, we note a simple consequence of the results proven so far.

COROLLARY 4.4. *The foliations $\{\mathcal{P}_n(a) \mid a \in \mathcal{S}_n\}$ and $\{\mathcal{P}_n(c) \mid c \in \mathcal{C}_n\}$ are complementary; i.e., any intersecting pair of leaves, with one leaf from each foliation, intersects transversely. Moreover, any intersecting pair of leaves intersects in at most one point.*

Proof. Setting $u = 0$ in (4.4), we obtain $G_\sigma v = 0$. Hence, by Lemma 4.1, $v = 0$ so that the foliations are transverse. If a leaf $\mathcal{P}_n(a)$ intersects a leaf $\mathcal{P}_n(c)$ at a point (a,σ), then the a-coordinates, and hence the roots of a, are known. According to Appendix B, the value of the cepstral coefficients coincides with the difference of the Newton sums of the powers of the roots of a and the roots of σ. Therefore, the Newton sums of the powers of the roots of σ are known, and, therefore, by the Newton identities, so is σ. □

A similar statement for the foliation $\{\mathcal{P}_n(\sigma) \mid \sigma \in \mathcal{S}_n\}$ can be proved by the mirror image of this proof and will be omitted.

Next, let $f : \mathcal{P}_n \to \mathbb{R}^n$ be the map which sends (a,σ) to the vector $r \in \mathbb{R}^n$ of normalized covariance lags with components

$$r_k = \frac{\int_{-\pi}^{\pi} e^{ik\theta}\left|w(e^{i\theta})\right|^2 d\theta}{\int_{-\pi}^{\pi}\left|w(e^{i\theta})\right|^2 d\theta}, \quad k = 1, 2, \dots, n, \tag{4.7}$$

and let $\mathcal{R}_n := f(\mathcal{P}_n)$. Of course, any $r \in \mathcal{R}_n$ satisfies the positivity condition

$$T_n = \begin{bmatrix} 1 & r_1 & \cdots & r_n \\ r_1 & 1 & \cdots & r_{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ r_n & r_{n-1} & \cdots & 1 \end{bmatrix} > 0.$$

Now, for each $r \in \mathcal{R}_n$, we want to show that

$$\mathcal{P}_n(r) = f^{-1}(r)$$

is a smooth manifold of dimension n. To this end, note that the function $f : \mathcal{P}_n \to \mathbb{R}^n$ has the components

$$f_k(a,\sigma) = \frac{h_k(a,\sigma)}{h_0(a,\sigma)},$$

where

$$h_k(a,\sigma) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\theta}\left|\frac{\sigma(e^{i\theta})}{a(e^{i\theta})}\right|^2 d\theta, \quad k = 0,1,2,\ldots,n.$$

Clearly, $h_0(a,\sigma) > 0$ for all $(a,\sigma) \in \mathcal{P}_n$.

A straightforward calculation shows that the directional derivative of f at $(a,\sigma) \in \mathcal{P}_n$ in the direction $(u,v) \in V_{n-1} \times V_{n-1}$ is

$$D_{(u,v)}f_k(a,\sigma) = \frac{1}{h_0(a,\sigma)}D_{(u,v)}h_k(a,\sigma) - \frac{h_k(a,\sigma)}{h_0(a,\sigma)^2}D_{(u,v)}h_0(a,\sigma), \tag{4.8}$$

where

$$D_{(u,v)}h_k(a,\sigma) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\frac{S(\sigma)v}{aa^*} - \frac{S(a)u}{aa^*}\frac{\sigma\sigma^*}{aa^*}\right]e^{ik\theta}d\theta. \tag{4.9}$$

Therefore, defining

$$\varphi(a,\sigma;u,v) := D_{(u,v)}\log h_0(a,\sigma) = \frac{D_{(u,v)}h_0(a,\sigma)}{h_0(a,\sigma)},$$

the kernel of the Jacobian of f at (a,σ) consists of those $(u,v) \in V_{n-1} \times V_{n-1}$ for which

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{S(\sigma)v}{aa^*}e^{ik\theta}d\theta = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{S(a)u}{aa^*}\frac{\sigma\sigma^*}{aa^*}e^{ik\theta}d\theta + \varphi(a,\sigma;u,v)\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\sigma\sigma^*}{aa^*}e^{ik\theta}d\theta$$

for $k = 0,1,\ldots,n$. In fact, this equation holds trivially for $k = 0$, and so, to simplify the notation in what follows, we add this equation.

PROPOSITION 4.5. *The space $\mathcal{P}_n(r)$ is a smooth, connected, n-manifold, and its tangent space $T_{(a,\sigma)}\mathcal{P}_n(r)$ consists of those $(u,v) \in V_{n-1} \times V_{n-1}$ for which*

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{S(\sigma)v}{aa^*}e^{ik\theta}d\theta = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{S(a)u}{aa^*}\frac{\sigma\sigma^*}{aa^*}e^{ik\theta}d\theta + \frac{\varphi}{2\pi}\int_{-\pi}^{\pi}\frac{\sigma\sigma^*}{aa^*}e^{ik\theta}d\theta \tag{4.10}$$

for $k = 0, 1, \dots, n$, where

$$\varphi = \frac{1}{h_0(a,\sigma)} \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\frac{S(\sigma)v}{aa^*} - \frac{S(a)u}{aa^*} \frac{\sigma\sigma^*}{aa^*} \right] d\theta. \tag{4.11}$$

The n-manifolds $\{\mathcal{P}_n(r) \mid r \in \mathcal{R}_n\}$ form the leaves of a foliation of $\mathcal{P}_n$.

Proof. The tangent space $T_{(a,\sigma)}\mathcal{P}_n(r)$ is the kernel of the Jacobian of f and is hence given by (4.10). Defining $p \in V_n$ as

$$p(z) := u(z) + \frac{1}{2}\varphi a(z),$$

these tangent equations may also be written as

$$Fp = Hv,$$

where $F : V_n \to \mathbb{R}^{n+1}$ and $H : V_{n-1} \to \mathbb{R}^{n+1}$ are the linear operators

$$Fp = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{S(a)p}{aa^*} \frac{\sigma\sigma^*}{aa^*} \begin{bmatrix} 1 \\ e^{i\theta} \\ \vdots \\ e^{in\theta} \end{bmatrix} d\theta, \qquad Hv = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{S(\sigma)v}{aa^*} \begin{bmatrix} 1 \\ e^{i\theta} \\ \vdots \\ e^{in\theta} \end{bmatrix} d\theta. \tag{4.12}$$

To see this, note that

$$\frac{1}{2} \frac{S(a)a}{aa^*} = 1.$$

Now the linear map F is nonsingular. In fact, supposing that $Fp = 0$ and, as in the proof of Lemma 4.1, taking the appropriate linear combination, we obtain

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{|S(a)p|^2}{aa^*} \frac{\sigma\sigma^*}{aa^*} d\theta = 0,$$

which holds if and only if $S(a)p = 0$. But since a is a Schur polynomial, $S(a)$ is nonsingular, and hence $Fp = 0$ if and only if $p = 0$. Since the range and the domain of F have the same dimension, F is nonsingular, as claimed. Then, since the leading term of the n-polynomial

$$p = F^{-1}Hu$$

is precisely $\frac{1}{2}\varphi$, φ is a linear function of u. This defines a linear map $L : V_{n-1} \to V_{n-1}$, which sends v to $u := p - \frac{1}{2}\varphi a$ so that $T_{(a,\sigma)}\mathcal{P}_n(r)$ consists of those (u, v) such that $u = Lv$. This establishes that $T_{(a,\sigma)}\mathcal{P}_n(r)$ is n-dimensional and that $\mathcal{P}_n(r)$ is an n-manifold. Since the rank of $\mathrm{Jac}(f)|_{(a,\sigma)}$ is full, smoothness follows from the implicit function theorem. The connectedness of $\mathcal{P}_n(r)$ was proven in [7]. Since the rank of $\mathrm{Jac}(f)|_{(a,\sigma)}$ is everywhere n, the connected submanifolds $\mathcal{P}_n(r)$ form the leaves of a foliation of $\mathcal{P}_n$. □

The relation between the foliations $\{\mathcal{P}_n(r) \mid r \in \mathcal{R}_n\}$ and $\{\mathcal{P}_n(c) \mid c \in \mathcal{C}_n\}$ is indeed interesting.

THEOREM 4.6. *For each $(a, \sigma) \in \mathcal{P}_n(r) \cap \mathcal{P}_n(c)$, the dimension of*

$$\Theta := T_{(a,\sigma)}\mathcal{P}_n(r) \cap T_{(a,\sigma)}\mathcal{P}_n(c)$$

equals the degree of the greatest common divisor of the polynomials $a(z)$ and $\sigma(z)$.

Proof. Any $(u, v) \in \Theta$ satisfies both (4.6) and (4.10). Taking the linear combinations of these equations corresponding to the coefficients of $\sigma\sigma^*$ and aa^*, respectively, we obtain

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} S(\sigma)v d\theta = \frac{1}{2\pi}\int_{-\pi}^{\pi} S(a)u\frac{\sigma\sigma^*}{aa^*}d\theta,$$
$$\frac{1}{2\pi}\int_{-\pi}^{\pi} S(\sigma)v d\theta = \frac{1}{2\pi}\int_{-\pi}^{\pi} S(a)u\frac{\sigma\sigma^*}{aa^*}d\theta + \varphi\|\sigma\|^2,$$

demonstrating that φ must be equal to zero. With $\varphi = 0$, (4.6) and (4.10) become

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{S(\sigma)v}{\sigma\sigma^*}e^{ik\theta}d\theta = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{S(a)u}{aa^*}e^{ik\theta}d\theta, \quad k = 0, 1, \dots, n, \tag{4.13}$$

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{S(\sigma)v}{aa^*}e^{ik\theta}d\theta = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{S(a)u}{aa^*}\frac{\sigma\sigma^*}{aa^*}e^{ik\theta}d\theta, \quad k = 0, 1, \dots, n. \tag{4.14}$$

Taking the appropriate linear combinations of (4.13) and (4.14), respectively, we obtain

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|S(\sigma)v|^2}{\sigma\sigma^*}d\theta = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{[S(a)u][S(\sigma)v]}{aa^*}d\theta,$$
$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{[S(\sigma)v][S(a)u]}{aa^*}d\theta = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|S(a)u|^2}{aa^*}\frac{\sigma\sigma^*}{aa^*}d\theta.$$

Now, setting

$$f_1 := \frac{S(\sigma)v}{\sigma^*} \quad \text{and} \quad f_2 := \frac{\sigma S(a)u}{aa^*},$$

these equations can be written as

$$\|f_1\|^2 = \langle f_1, f_2\rangle \quad \text{and} \quad \langle f_1, f_2\rangle = \|f_2\|^2$$

in the inner product and norm of $L^2[-\pi, \pi]$. Using the parallelogram law yields

$$\|f_1 - f_2\|^2 = \|f_1\|^2 + \|f_2\|^2 - 2\langle f_1, f_2\rangle = 0,$$

which in turn implies that $f_1 = f_2$. Therefore,

$$\frac{S(\sigma)v}{\sigma\sigma^*} = \frac{S(a)u}{aa^*}$$

on the unit circle or, equivalently,

$$\operatorname{Re}\left\{\frac{v}{\sigma}\right\} = \operatorname{Re}\left\{\frac{u}{a}\right\}. \tag{4.15}$$

However, since these are harmonic functions, (4.15) must hold in the whole complex plane. In particular, as $a(z)$ and $\sigma(z)$ are real polynomials, this becomes

$$\frac{v}{\sigma} = \frac{u}{a} \tag{4.16}$$

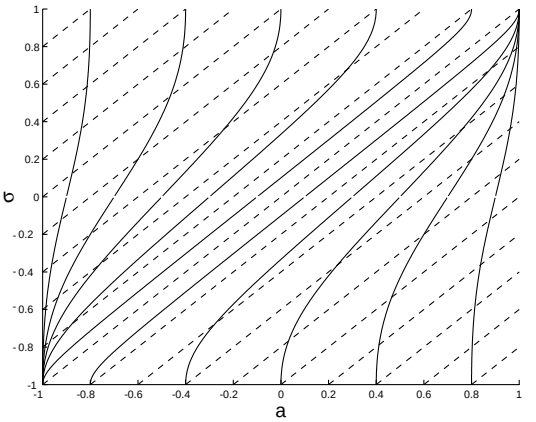


FIG. 5. *Cepstral (dotted line) and covariance (solid line) matching foliations of* $\mathcal{P}_1$.

on the real line. However, these functions are analytic outside the unit disc, and so, by the identity theorem, (4.16) is valid in the whole complex plane. Clearly, $u = v = 0$ satisfy (4.16), but let us see if there are nontrivial $u(z)$ and $v(z)$ of degree $n-1$. If so, (4.16) can also be written as

$$\frac{v}{u} = \frac{\sigma}{a},$$

which, of course, has no solution if $a(z)$ and $\sigma(z)$ are coprime. If $a(z)$ and $\sigma(z)$ have a greatest common factor of degree d, $u(z)$ and $v(z)$ could be polynomials of degree less than or equal to $n-1$ and have an arbitrary common factor of degree $d-1$, hence defining a vector space of dimension d, as claimed. □

Remark 4.7. It follows from Theorem 4.6 that the foliations $\{\mathcal{P}_n(r) \mid r \in \mathcal{R}_n\}$ and $\{\mathcal{P}_n(c) \mid c \in \mathcal{C}_n\}$ are complementary at any point $(a, \sigma) \in \mathcal{P}_n$, where a and σ are coprime, as illustrated in Figure 5 for $n = 1$. From this it follows that the kernels of $\mathrm{Jac}(g)|_{(a,\sigma)}$ and $\mathrm{Jac}(f)|_{(a,\sigma)}$ are complementary at any point (a, σ) in $\mathcal{P}_n^*$. In particular, the Jacobian of the joint map $(a, \sigma) \to (r_1, r_2, \dots, r_n, c_1, c_2, \dots, c_n)$ has full rank, and, by the inverse function theorem, the joint map forms a smooth local coordinate system on $\mathcal{P}_n^*$. This proves Theorem 3.1.

5. Identifiability of shaping filters from cepstral and covariance windows. In this section, we shall show that the window of n cepstral coefficients and the window of n normalized covariance lags do indeed determine the (normalized) shaping filter which generates these data, provided the filter has degree n, thus proving Theorem 3.2. As a preliminary to this argument, however, we want to return to the generalization of the maximum entropy integral in terms of "positive" linear combinations of the entire cepstral window. Not only is this an appealing idea, but it also turns out to give a novel derivation of a result which is of independent interest in itself, a solution of the rational covariance extension problem. We now formalize our analysis of this generalized maximum entropy problem.

THEOREM 5.1. *If the pseudopolynomial* (3.7) *belongs to* $\mathcal{D}_+$*, the problem to maximize* (3.5) *subject to* (3.8) *has a unique solution* Φ*, and this solution has the form*

$$\Phi(z) = \frac{P(z)}{Q(z)}, \tag{5.1}$$

where

$$Q(z) = q_0 + \frac{1}{2}q_1(z + z^{-1}) + \cdots + \frac{1}{2}q_n(z^n + z^{-n}) \tag{5.2}$$

also belongs to $\mathcal{D}_+$.

It turns out that the algorithm needed to determine Q is precisely the convex optimization algorithm presented in [12]. In fact, the algorithm is based on the dual problem, in the sense of mathematical programming, of the problem to maximize (3.6) subject to (3.8). More precisely, let $\mathcal{F}_+$ be the set of bounded positive real functions

$$f(z) = \frac{1}{2}f_0 + f_1 z^{-1} + f_2 z^{-2} + \cdots$$

such that $\Phi(e^{i\theta}) := 2\mathrm{Re}\{f(e^{i\theta})\}$ is bounded away from zero, and consider the (primal) problem to maximize the generalized entropy (3.6) over $\mathcal{F}_+$, i.e.,

$$\max_{f\in\mathcal{F}_+} \mathbb{I}_P(f),$$

subject to (3.8). Then duality theory amounts to forming the Lagrangian

$$\begin{aligned} L(f,q) &= \mathbb{I}_P(f) + \sum_{k=0}^{n} q_k \left[r_k - \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\theta}\Phi(e^{i\theta})d\theta \right] \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} P(e^{i\theta})\log\Phi(e^{i\theta})d\theta + r'q - \frac{1}{2\pi}\int_{-\pi}^{\pi} Q(e^{i\theta})\Phi(e^{i\theta})d\theta \end{aligned} \tag{5.3}$$

and determining the Lagrange multipliers $q \in \mathbb{R}^{n+1}$ by minimizing the dual functional

$$\psi(q) := \sup_{f\in\mathcal{F}_+} L(f,q).$$

Clearly, $\psi(q) < \infty$ only if both P and Q belong to $\mathcal{D}$. If the function $f \mapsto L(f,q)$ has a maximum in the open region $\mathcal{F}_+$, then

$$\frac{\partial L}{\partial f_k} = 0, \quad k = 0, 1, 2, \ldots,$$

in the maximizing point. This stationarity condition becomes

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\theta}\left[P(e^{i\theta})\Phi(e^{i\theta})^{-1} - Q(e^{i\theta})\right]d\theta = 0, \quad k = 0, 1, 2, \ldots,$$

which is satisfied if and only if (5.1) or, equivalently,

$$f_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\theta}\frac{P(e^{i\theta})}{Q(e^{i\theta})}d\theta \tag{5.4}$$

holds. Inserting this into (5.3) yields the dual functional

$$\psi(q) = \mathbb{J}_P(q) + \frac{1}{2\pi}\int_{-\pi}^{\pi} P(e^{i\theta})[\log P(e^{i\theta}) - 1]d\theta \tag{5.5}$$

for all $P, Q \in \mathcal{D}$, where

$$\mathbb{J}_P(q) = r_0 q_0 + r_1 q_1 + \cdots + r_n q_n - \frac{1}{2\pi}\int_{-\pi}^{\pi} P(e^{i\theta})\log Q(e^{i\theta})d\theta. \tag{5.6}$$

Since the last term in (5.5) does not depend on q, we shall call the optimization problem

$$\min_{Q\in\mathcal{D}} \mathbb{J}_P(Q) \tag{5.7}$$

the *dual problem.* The functional (5.6) is strictly convex, and, therefore, the minimum is unique, provided one exists. This is precisely the optimization problem considered in [12], where the following theorem was proven.

THEOREM 5.2. *The dual problem has a unique solution, and it belongs to* $\mathcal{D}_+$.

Since thus $\mathbb{J}_P$ takes its minimum in an interior point,

$$\frac{\partial \mathbb{J}_P}{\partial q_k} = r_k - \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\theta}\frac{P(e^{i\theta})}{Q(e^{i\theta})}d\theta \tag{5.8}$$

equals zero there for $k = 0, 1, \dots, n$. This stationarity condition is precisely the covariance matching condition. The dual problem is easily solved by Newton's method [12, 14]. The statement of Theorem 5.2 is nontrivial. In fact, the proof [12] relies on the fact that the gradient (5.8) tends to infinity as Q tends to the boundary of $\mathcal{D}$.

Proof of Theorem 5.1. Let $\hat{Q} \in \mathcal{D}_+$ be the unique solution to the dual problem (5.7), let $\hat{q} \in \mathbb{R}^{n+1}$ be the corresponding vector of coefficients, and let

$$\hat{f}_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\theta}\frac{P(e^{i\theta})}{\hat{Q}(e^{i\theta})}d\theta.$$

Clearly, $\hat{f} \in \mathcal{F}_+$. Since the gradient (5.8) is zero for $Q = \hat{Q}$, the covariance matching condition (3.8) is fulfilled for $f = \hat{f}$, and, therefore, $\mathbb{I}_P(\hat{f}) = L(\hat{f}, \hat{q})$. But, by the construction above,

$$L(\hat{f}, \hat{q}) = \sup_{f\in\mathcal{F}_+} L(f, \hat{q}) \geq L(f, \hat{q})$$

for all $f \in \mathcal{F}_+$. Then, for any $f \in \mathcal{F}_+$ which satisfies the covariance matching condition (3.8),

$$\mathbb{I}_P(f) = L(f, \hat{q}) \leq \mathbb{I}_P(\hat{f}),$$

which establishes the optimality of $\hat{f}$. □

This analysis motivates the construction of a functional which will be the key in establishing uniqueness of minimum-phase shaping filters having prescribed windows $r_0, r_1, \dots, r_n$ and $c_1, c_2, \dots, c_n$ of covariance lags and cepstral coefficients, respectively. More precisely, consider the (primal) problem of finding a spectral density

$$\Phi(e^{i\theta}) = f_0 + 2\sum_{k=1}^{\infty} f_k \cos k\theta,$$

which minimizes

$$\mathbb{I}(f) = \sum_{k=1}^{n}\left|\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\theta}\log\Phi(e^{i\theta})d\theta - c_k\right| - \frac{1}{2\pi}\int_{-\pi}^{\pi}\log\Phi(e^{i\theta})d\theta \tag{5.9}$$

subject to the covariance-lag matching condition

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\theta}\Phi(e^{i\theta})d\theta = r_k, \quad k = 0, 1, \dots, n. \tag{5.10}$$

The objective function (5.9) is the (ℓ_1) "cepstral error" minus the entropy gain. As discussed in section 3, the entropy gain is precisely what is maximized in the LPC solution, and it is identical to the zeroth cepstral coefficient corresponding to Φ. This term compensates for the absence of a zeroth term in the cepstral error.

To obtain a suitable dual problem, we reformulate the *primal problem* to minimize

$$\sum_{k=0}^{n} \epsilon_k - \frac{1}{2\pi} \int_{-\pi}^{\pi} \log \Phi(e^{i\theta}) d\theta$$

subject to the covariance matching condition (5.10) and

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\theta} \log \Phi(e^{i\theta}) d\theta - c_k - \epsilon_k \leq 0, \quad k = 1, 2, \ldots, n, \tag{5.11}$$

$$-\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\theta} \log \Phi(e^{i\theta}) d\theta + c_k - \epsilon_k \leq 0, \quad k = 1, 2, \ldots, n. \tag{5.12}$$

Taking $q_0, q_1, \ldots, q_n$ to be the Lagrange multipliers for the constraints (5.10) and $\lambda_1, \lambda_2, \ldots, \lambda_n$ and $\mu_1, \mu_2, \ldots, \mu_n$ to be nonnegative Lagrange multipliers for the sets of constraints (5.11) and (5.12), respectively, we obtain the Lagrangian

$$\begin{aligned} L(f, \epsilon, q, \lambda, \mu) = \sum_{k=1}^{n} \epsilon_k &- \frac{1}{2\pi} \int_{-\pi}^{\pi} \log \Phi(e^{i\theta}) d\theta \\ &+ \sum_{k=0}^{n} q_k \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\theta} \Phi(e^{i\theta}) d\theta - r_k \right] \\ &+ \sum_{k=1}^{n} \lambda_k \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\theta} \log \Phi(e^{i\theta}) d\theta - c_k - \epsilon_k \right] \\ &- \sum_{k=1}^{n} \mu_k \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\theta} \log \Phi(e^{i\theta}) d\theta - c_k + \epsilon_k \right]. \end{aligned}$$

Now, setting

$$p_0 = 1, \qquad p_k := \mu_k - \lambda_k, \quad k = 1, 2, \ldots, n, \tag{5.13}$$

we can write this in the more compact form

$$\begin{aligned} L(f, \epsilon, q, \lambda, \mu) = &\sum_{k=1}^{n} (1 - \lambda_k - \mu_k) \epsilon_k \\ &+ c_1 p_1 + c_2 p_2 + \cdots + c_n p_n - r_0 q_0 - r_1 q_1 - \cdots - r_n q_n \\ &+ \frac{1}{2\pi} \int_{-\pi}^{\pi} Q(e^{i\theta}) \Phi(e^{i\theta}) d\theta - \frac{1}{2\pi} \int_{-\pi}^{\pi} P(e^{i\theta}) \log \Phi(e^{i\theta}) d\theta, \end{aligned}$$

which clearly can have a finite minimum only for those values of the Lagrange multipliers for which both P and Q belong to $\mathcal{D}$ and $\lambda_k + \mu_k \leq 1$ for $k = 1, 2, \ldots, n$. For such Lagrange multipliers, if the function $(f, \epsilon) \to L(f, \epsilon, q, \lambda, \mu)$ has a minimum, then

$$\frac{\partial L}{\partial f_k} = 0, \quad k = 0, 1, 2, \ldots, \tag{5.14}$$

and

$$(1-\lambda_k-\mu_k)\epsilon_k=0, \quad k=1,2,\dots,n, \tag{5.15}$$

in the minimizing point. The stationarity condition (5.14) becomes

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\theta}\left[P(e^{i\theta})\Phi(e^{i\theta})^{-1}-Q(e^{i\theta})\right]d\theta=0, \quad k=0,1,2,\dots,$$

or, equivalently,

$$\Phi(z)=\frac{P(z)}{Q(z)},$$

which, inserted together with (5.15) into the Lagrangian with P given by (5.13), yields the dual functional

$$\inf_{(f,\epsilon)\in\mathcal{F}_+\times\mathbb{R}^+} L(f,\epsilon,q,\lambda,\mu)=\mathbb{J}(P,Q)+1,$$

where the functional

$$\begin{aligned}\mathbb{J}(P,Q)&=c_1p_1+c_2p_2+\cdots+c_np_n-r_0q_0-r_1q_1-\cdots-r_nq_n\\ &\quad-\frac{1}{2\pi}\int_{-\pi}^{\pi}P(e^{i\theta})\log\frac{P(e^{i\theta})}{Q(e^{i\theta})}d\theta\end{aligned} \tag{5.16}$$

is concave but not necessarily strictly concave.

THEOREM 5.3. *The dual problem to maximize $\mathbb{J}(P,Q)$ over all $(P,Q)\in\mathcal{D}\times\mathcal{D}$ such that $p_0=1$ has a solution $(\hat{P},\hat{Q})$, and, for any such solution, $\hat{Q}\in\mathcal{D}_+$, and*

$$\Phi(z)=\frac{\hat{P}(z)}{\hat{Q}(z)} \tag{5.17}$$

satisfies the covariance matching condition (5.10). *If, in addition, $\hat{P}\in\mathcal{D}_+$, then* (5.17) *is a solution of the primal problem with $\epsilon_1=\epsilon_2=\cdots=\epsilon_n=0$, i.e., there is both covariance matching and cepstral matching. A maximizing point $(\hat{P},\hat{Q})\in\mathcal{D}_+\times\mathcal{D}_+$ is unique if and only if $\hat{P}$ and $\hat{Q}$ are coprime.*

Proof. It can be shown along the same lines as in [12] that the functional $\mathbb{J}$ has compact sublevel sets in $\mathcal{D}\times\mathcal{D}$. Hence $\mathbb{J}$ has a maximal point $(\hat{P},\hat{Q})$ there. The boundary of $\mathcal{D}\times\mathcal{D}$ consists of those points where either $\hat{P}$ or $\hat{Q}$ or both have zeros on the unit circle. Now a straightforward calculation shows that

$$\frac{\partial\mathbb{J}}{\partial q_k}=\frac{1}{2\pi}\int_{-\pi}^{\pi}e^{ik\theta}\frac{P(e^{i\theta})}{Q(e^{i\theta})}d\theta-r_k, \quad k=0,1,\dots,n, \tag{5.18}$$

$$\frac{\partial\mathbb{J}}{\partial p_k}=c_k-\frac{1}{2\pi}\int_{-\pi}^{\pi}e^{ik\theta}\log\frac{P(e^{i\theta})}{Q(e^{i\theta})}d\theta, \quad k=1,2,\dots,n. \tag{5.19}$$

From this and the argument in [12], it can be shown that the gradient (5.18) becomes infinite when Q lies on the boundary and hence that $\hat{Q}\in\mathcal{D}_+$. Therefore, since the functional $\mathbb{J}$ is concave, (5.18) must be zero at $(\hat{P},\hat{Q})$, and hence (5.17) satisfies the covariance matching condition (5.10).

Next suppose that $\hat{P} \in \mathcal{D}_+$. Then (5.19) must also be zero at $(\hat{P}, \hat{Q})$, and hence there is also cepstral matching. For any $f \in \mathcal{F}_+$ satisfying (5.10) and $\epsilon > 0$,

$$\mathbb{I}(f) \geq L(f, \epsilon, \hat{q}, \hat{\lambda}, \hat{\mu}) \geq \mathbb{J}(\hat{P}, \hat{Q}) + 1,$$

where $\hat{q}$, $\hat{\lambda}$, and $\hat{\mu}$ are Lagrange multipliers corresponding to $(\hat{P}, \hat{Q})$. On the other hand, if $\hat{f}$ is the positive-real part of (5.17), then

$$\mathbb{I}(\hat{f}) = L(\hat{f}, 0, \hat{q}, \hat{\lambda}, \hat{\mu}) = \mathbb{J}(\hat{P}, \hat{Q}) + 1,$$

and hence $\hat{f}$ minimizes $\mathbb{I}$, and $\hat{\epsilon} = 0$, as claimed.

Clearly, the maximizing solution $(\hat{P}, \hat{Q})$ cannot be unique if $\hat{P}$ and $\hat{Q}$ are not coprime. Therefore, the last statement of the theorem would follow if we could show that $\mathbb{J}$ is *strictly* concave over some neighborhood of $\mathcal{D}_+ \times \mathcal{D}_+$ if $\hat{P}$ and $\hat{Q}$ are coprime. To this end, we consider the Hessian. Let

$$\delta\mathbb{J}(P, Q; \delta P, \delta Q) = \lim_{\epsilon \to 0} \frac{\mathbb{J}(P + \delta P, Q + \epsilon\delta Q) - \mathbb{J}(P, Q)}{\epsilon}$$

denote the directional derivative in the direction $(\delta P, \delta Q)$. The admissible directions $(\delta P, \delta Q)$ are symmetric pseudopolynomials such that $(P + \epsilon\delta P, Q + \epsilon\delta Q) \in \mathcal{D}_+ \times \mathcal{D}_+$ for sufficiently small $\epsilon > 0$. Since $p_0 = 1$, we must also have $\delta p_0 = 0$. It is straightforward to see that

$$\begin{aligned}\delta\mathbb{J}(P, Q; \delta P, \delta Q) = {} & c_1\delta p_1 + c_2\delta p_2 + \cdots + c_n\delta p_n - r_0\delta q_0 - r_1\delta q_1 - \cdots - r_n\delta q_n \\ & + \frac{1}{2\pi}\int_{-\pi}^{\pi} \delta Q(e^{i\theta}) \frac{P(e^{i\theta})}{Q(e^{i\theta})} d\theta - \frac{1}{2\pi}\int_{-\pi}^{\pi} \delta P(e^{i\theta}) \log \frac{P(e^{i\theta})}{Q(e^{i\theta})} d\theta,\end{aligned}$$

and hence second differentiation yields

$$\delta^2\mathbb{J}(P, Q; \delta P, \delta Q) = -\left\langle (P\delta Q - Q\delta P)^2, \frac{1}{PQ^2} \right\rangle \leq 0,$$

where equality holds if and only if $P\delta Q - Q\delta P = 0$, i.e., if and only if

$$\frac{\delta P}{\delta Q} = \frac{P}{Q}.$$

However, this is impossible if $\hat{P}$ and $\hat{Q}$ are to be coprime, since $p_0 = 1$ and $\delta p_0 = 0$. Consequently, $\mathbb{J}$ is strictly concave at $(\hat{P}, \hat{Q})$, as claimed. □

Now, given the minimizing pair of pseudopolynomials $(\hat{P}, \hat{Q})$ of Theorem 5.3, let $a(z)$ and $\sigma(z)$ be the normalized, polynomial spectral factors of $\hat{Q}$ and $\hat{P}$, respectively, i.e., the Schur polynomials satisfying

$$a(z)a(z^{-1}) = \frac{1}{a_0^2}\hat{Q}(z), \qquad \sigma(z)\sigma(z^{-1}) = \frac{1}{\sigma_0^2}\hat{P}(z),$$

where a_0^2 and σ_0^2 are the appropriate normalizing factors. Then Theorem 5.3 provides a procedure for determining, from a combined window $(r_0, r_1, \ldots, r_n, c_1, \ldots, c_n)$ of covariance lags and cepstral coefficients, a pair (a, σ), which is unique if and only if $a(z)$ and $\sigma(z)$ are coprime, i.e., $(a, \sigma) \in \mathcal{P}_n^*$, and a corresponding (unnormalized) shaping filter

$$w(z) = \frac{\sigma_0}{a_0}\frac{\sigma(z)}{a(z)}.$$

Therefore, in particular, we have proved Theorem 3.2. In fact, given any $(a, \sigma) \in \mathcal{P}_n^*$, a window $(r_1, \dots, r_n, c_1, \dots, c_n)$ is uniquely determined from (4.7) and (4.1). Conversely, given $(r_1, \dots, r_n, c_1, \dots, c_n)$, the optimization problem of Theorem 5.3 yields an $(a, \sigma) \in \mathcal{P}_n$, which matches this window and is unique if and only if $(a, \sigma) \in \mathcal{P}_n^*$.

6. The simultaneous partial realization problem. While the *stochastic realization problem* [25, 21, 26, 10, 30] amounts to determining shaping filters w having a fixed window of covariance lags $r_0, r_1, \dots, r_n$, the object of the *deterministic realization problem* (see, e.g., [3, 23]) is to find shaping filters w with a fixed window $w_0, w_1, \dots, w_n$ of Markov parameters (1.11). An important question is whether the two problems can be solved simultaneously so that both interpolation conditions are satisfied at the same time. This problem has been studied in the literature as the Q-Markov cover problem (see [31, 29, 1], where it has been used as a tool for performing model reduction).

This basic question will also be addressed in this section using geometric methods. Thus we would ask whether the two problems can be solved simultaneously and, if so, whether this solution is unique. We find a positive answer to the existence question in $\mathcal{Q}_n$ using fixed point methods. We also determine where these windows provide a bona fide set of smooth coordinates. Finally, we give a geometric proof of the uniqueness of the corresponding shaping filter, i.e., of identifiability of the shaping filter from covariance and Markov windows, providing an independent proof of a result which is basic to the existing theory of the Q-Markov cover problem. These results prove the assertions in Theorem 3.5. We also provide an independent proof of Theorem 3.4.

To address these issues, let $\psi : \mathcal{Q}_n \to \mathbb{R}^n$ be the map which sends (a, σ) to

$$w := \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix},$$

and let $\mathcal{W}_n := \psi(\mathcal{Q}_n)$. Given any $w \in \mathcal{W}_n$, define

$$\mathcal{Q}_n(w) := \psi^{-1}(w).$$

Now, multiplying (2.2) by $a(z)$ and identifying coefficients of nonnegative powers in z, we have

$$\begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \vdots \\ \sigma_n \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} + \begin{bmatrix} 1 & & & \\ w_1 & 1 & & \\ \vdots & \vdots & \ddots & \\ w_{n-1} & w_{n-2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}. \tag{6.1}$$

Identifying coefficients in negative powers of z yields the appropriate Hankel system. From (6.1) we see first that $\mathcal{W}_n = \mathbb{R}^n$. Second, given w, a can be chosen arbitrarily in $\mathcal{S}_n$. Hence, $\mathcal{Q}_n(w)$ is completely parameterized by $a \in \mathcal{S}_n$, and hence it is a connected n-manifold, diffeomorphic to $\mathbb{R}^n$. Its boundary is characterized by a having a root on the unit circle. Clearly, the closure $\overline{\mathcal{Q}_n(w)}$ is the graph of a continuous function $\gamma : \overline{\mathcal{S}_n} \to \Pi_n$, defined by (6.1). Although the manifold $\mathcal{Q}_n$ is not bounded, $\mathcal{Q}_n(w)$ is. Moreover, $\overline{\mathcal{Q}_n(w)}$ is homeomorphic to $\overline{\mathcal{S}_n}$, which is compact with a contractible interior (see Appendix A).

THEOREM 6.1. *Any continuous map* $T : \overline{\mathcal{S}_n} \to \overline{\mathcal{S}_n}$ *has a fixed point.*

Proof. We first note that $\overline{\mathcal{S}_n}$ is contained in the (Euclidean) space of real monic polynomials with roots in the open disc of radius $1 + \epsilon$ for any positive ϵ. As in Appendices A and C, the continuous retraction $r : \mathbb{D}_{1+\epsilon} \to \mathbb{D}$, defined by

$$r(x) = \begin{cases} x \text{ if } \|x\| \leq 1, \\ \frac{x}{\|x\|} \text{ if } \|x\| \geq 1, \end{cases}$$

induces a continuous retraction of $\mathcal{P}_{\mathbb{D}_{1+\epsilon}}(n) \to \overline{\mathcal{S}_n}$. In particular, $\overline{\mathcal{S}_n}$ is a Euclidean neighborhood retract, and, therefore, the Lefschetz fixed point theorem applies to continuous maps of $\overline{\mathcal{S}_n}$ to itself [18, p. 209]. The Lefschetz fixed point theorem asserts that a continuous map f from a space X to itself has a fixed point provided its Lefschetz number is nonzero. More precisely, to define the Lefschetz number, we need to introduce the homology (real) vector spaces $H_i(X; \mathbb{R})$, defined for each $i = 0, 1, 2, \ldots$. If X is a compact Euclidean neighborhood retract in $\mathbb{R}^N$, then each $H_i(X; \mathbb{R})$ is finite-dimensional and vanishes for $i > N$. In this case, the Lefschetz number of f, $\mathrm{Lef}(f)$, is defined as

$$\mathrm{Lef}(f) = \sum_{i=0}^{n} \mathrm{tr}(f_{*i}),$$

where (f_{*i}) is the linear transformation

$$f_{*i} : H_i(X; \mathbb{R}) \to H_i(X; \mathbb{R})$$

introduced by f. For $X = \overline{\mathcal{S}_n}$, we have

$$H_i(\overline{\mathcal{S}_n}, \mathbb{R}) = \{0\} \quad \text{for } i \geq 1$$

since $\overline{\mathcal{S}_n}$ is contractable. Moreover, since $\overline{\mathcal{S}_n}$ is therefore connected,

$$H_0(\overline{\mathcal{S}_n}, \mathbb{R}) \sim \mathbb{R},$$

and the map f_{*i} is the identity. In summary, $\mathrm{Lef}(f) = 1$, and the Lefschetz fixed point theorem therefore implies that f has a fixed point. ∎

Remark 6.2. One might hope that the Brower fixed point theorem would apply directly to $\overline{\mathcal{S}_n}$. Even in the case when $n = 2$, this does not work. In fact, the space $\overline{\mathcal{S}_2}$ is represented by a triangle in the plane, and its interior is a manifold with corners and not a disc. While in this simple case the closure of the Schur region is homeomorphic to a disc, a proof in arbitrary dimensions has not yet been formulated, but the current standard methods of the Lefschetz fixed point theorem apply readily.

The tangent space of $\mathcal{Q}_n(w)$ at (a, σ) is given by the following proposition.

PROPOSITION 6.3. *For each* $w \in \mathcal{W}_n$, *the space* $\mathcal{Q}_n(w)$ *is a smooth, connected* n*-manifold with the tangent space*

$$T_{(a,\sigma)}\mathcal{Q}_n(w) = \{(u, v) \in V_{n-1} \times V_{n-1} \mid av - \sigma u = \rho;\ \deg \rho \leq n - 1\} \tag{6.2}$$

at $(a, \sigma) \in \mathcal{Q}_n(w)$. *The* n*-manifolds* $\{\mathcal{Q}_n(w) \mid w \in \mathcal{W}_n\}$ *form the leaves of a foliation of* $\mathcal{Q}_n$.

Proof. We have already established that $\mathcal{Q}_n(w)$ is a connected n-manifold, diffeomorphic to $\mathbb{R}^n$. To prove that $T_{(a,\sigma)}\mathcal{Q}_n(w)$ is given by (6.2), observe that the

directional derivative

$$D_{(u,v)}\psi(a,\sigma) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\theta}\left[\frac{v}{a} - \frac{\sigma u}{a^2}\right] d\theta$$
$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\theta}\frac{av-\sigma u}{a^2} d\theta$$

is zero for $k = 0, 1, \dots, n$ if and only if the polynomial $\rho = av - \sigma u$ has degree at most $n-1$. In fact, $z^k\rho(z)/a(z)^2$ is analytic for $z \geq 1$ and strictly proper precisely when $\deg \rho < 2n - k$. Since the tangent space has dimension n, the rank of $\mathrm{Jac}(\psi)|_{(a,\sigma)}$ is everywhere n, and hence the connected submanifolds $\mathcal{Q}_n(w)$ form the leaves of a foliation of $\mathcal{Q}_n$. □

As pointed out in the introduction, for *minimum-phase* shaping filters, there is a close relation between the cepstral coefficients and the Markov parameters of the corresponding shaping filter w. To establish these relations, make a Laurent expansion of

$$\log \Phi(z) = \log w(z) + \log w(z^{-1}) \tag{6.3}$$

on a subset Ω of the complex plane, where Ω is the intersection between an annulus containing the unit circle but none of the zeros of $w(z)$ or $w(z^{-1})$ and a sector containing the positive-real axis. The purpose of the sector is to avoid circling the origin. Then the Laurent expansion obtained from the series expansions on the corresponding segment of the real line of $\log w(z)$ and $\log w(z^{-1})$ extends to all of Ω and hence, in particular, to the arc on the unit circle contained in Ω. Then, however, the uniqueness of the Fourier transform ensures that the Laurent expansion also holds there. From this we see

$$\begin{aligned}
c_0 &= 2\log w_0,\\
c_1 &= \frac{w_1}{w_0},\\
c_2 &= \frac{w_2}{w_0} - \frac{1}{2}\left(\frac{w_1}{w_0}\right)^2,\\
c_3 &= \frac{w_3}{w_0} - \frac{1}{2}\left(2\frac{w_1}{w_0}\frac{w_2}{w_0}\right) + \frac{1}{3}\left(\frac{w_1}{w_0}\right)^3.\\
&\vdots
\end{aligned}$$

Indeed, these equations form a triangular system, and hence the Markov parameters can also be obtained from the cepstral coefficients, and vice versa. Setting $w_0 = 1$, we obtain the usual normalization with $c_0 = 0$. Therefore, the nonempty submanifolds $\mathcal{Q}_n(w) \cap \mathcal{P}_n$ are precisely the leaves of the foliation $\{\mathcal{P}_n(c) \mid c \in \mathcal{C}_n\}$. In fact, let $\phi : \mathcal{P}_n \to \mathbb{R}^n$ be the restriction of ψ to $\mathcal{P}_n$, and define $\mathcal{P}_n(w) := \phi^{-1}(w)$ for each $w \in \mathcal{M}_n := \phi(\mathcal{P}_n)$. Then we have the following corollary.

COROLLARY 6.4. *The n-manifolds $\{\mathcal{P}_n(w) \mid w \in \mathcal{M}_n\}$ form the leaves of a foliation of $\mathcal{P}_n$, which is identical to $\{\mathcal{P}_n(c) \mid c \in \mathcal{C}_n\}$.*

In the present setting, however, we also consider nonminimum phase shaping filters, allowing σ to be an arbitrary real monic polynomial. Whereas in $\mathcal{P}_n$ there is a one-to-one correspondence between windows of cepstral coefficients and Markov parameters, this is no longer the case in $\mathcal{Q}_n$. The tangent vectors of $\mathcal{P}_n(w)$ at (a,σ)

do satisfy (4.6) of Proposition 4.2, but this does not extend to the situation where $\sigma(z)$ is no longer a Schur polynomial. Indeed, the first integral in (4.6) is not even defined when $\sigma(z)$ has a root on the unit circle. Nevertheless, we have the following lemma, which is all we need below.

LEMMA 6.5. *Any* $(u, v) \in T_{(a,\sigma)}\mathcal{Q}_n(w)$ *satisfies the equation*

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} S(\sigma)v\,d\theta = \frac{1}{2\pi}\int_{-\pi}^{\pi} S(a)u\frac{\sigma\sigma^*}{aa^*}d\theta. \tag{6.4}$$

Proof. By Proposition 6.3, the tangent space $T_{(a,\sigma)}\mathcal{Q}_n(w)$ consists of those (u, v) for which the polynomial $\rho := av - \sigma u$ has degree at most $n-1$. Since

$$v = \frac{\sigma u}{a} + \frac{\rho}{a},$$

we have

$$S(\sigma)v = S(a)u\frac{\sigma\sigma^*}{aa^*} + \sigma^*\frac{\rho}{a} + \sigma\left(\frac{\rho}{a}\right)^*. \tag{6.5}$$

However, ρ/a is strictly proper and analytic for $|z| \geq 1$, and hence it has a Laurent expansion

$$\frac{\rho}{a} = \alpha_1 z^{-1} + \alpha_2 z^{-2} + \cdots,$$

which is valid on the unit circle. Therefore,

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\sigma^*\frac{\rho}{a} + \sigma\left(\frac{\rho}{a}\right)^*\right]d\theta = 0, \tag{6.6}$$

and hence (6.4) follows. □

Next let $\phi : \mathcal{Q}_n \to \mathbb{R}^n$ be the map that sends (a, σ) to the vector $r \in \mathbb{R}^n$ of normalized covariance lags (4.7). Clearly, $\phi(\mathcal{Q}_n) = \mathcal{R}_n := f(\mathcal{P}_n)$. Given any $r \in \mathcal{R}_n$, define

$$\mathcal{Q}_n(r) := \phi^{-1}(r).$$

The following proposition is a $\mathcal{Q}_n$-version of Proposition 4.5, and the proof is the same mutatis mutandis.

PROPOSITION 6.6. *For each* $r \in \mathcal{R}_n$, $\mathcal{Q}_n(r)$ *is a smooth, connected manifold of dimension* n. *The tangent space* $T_{(a,\sigma)}\mathcal{Q}_n(r)$ *consists of those* $(u, v) \in V_{n-1} \times V_{n-1}$ *for which*

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{S(\sigma)v}{aa^*}e^{ik\theta}d\theta = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{S(a)u}{aa^*}\frac{\sigma\sigma^*}{aa^*}e^{ik\theta}d\theta + \frac{\varphi}{2\pi}\int_{-\pi}^{\pi}\frac{\sigma\sigma^*}{aa^*}e^{ik\theta}d\theta \tag{6.7}$$

for $k = 0, 1, \ldots, n$, *where*

$$\varphi = \frac{1}{h_0(a,\sigma)}\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\frac{S(\sigma)v}{aa^*} - \frac{S(a)u}{aa^*}\frac{\sigma\sigma^*}{aa^*}\right]d\theta. \tag{6.8}$$

The n*-manifolds* $\{\mathcal{Q}_n(r) \mid r \in \mathcal{R}_n\}$ *form the leaves of a foliation of* $\mathcal{Q}_n$.

In the case in which $a(z)$ and $\sigma(z)$ are coprime, we can now show that if the tangent spaces $T_{(a,\sigma)}\mathcal{Q}_n(w)$ and $T_{(a,\sigma)}\mathcal{Q}_n(r)$ do intersect, they intersect transversely.

PROPOSITION 6.7. *Suppose that the polynomials $a(z)$ and $\sigma(z)$ are coprime. Then*

$$T_{(a,\sigma)}\mathcal{Q}_n(w) \cap T_{(a,\sigma)}\mathcal{Q}_n(r) = 0 \tag{6.9}$$

for any $(a,\sigma) \in \mathcal{Q}_n(w) \cap \mathcal{Q}_n(r)$.

Proof. Suppose that $(u,v) \in T_{(a,\sigma)}\mathcal{Q}_n(w) \cap T_{(a,\sigma)}\mathcal{Q}_n(r)$. Then (u,v) satisfies (6.7) for $k = 0, 1, \dots, n$ and, by symmetry, also for $k = -1, -2, \dots, -n$. Taking the linear combination corresponding to the coefficients of aa^*, we obtain

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} S(\sigma)v\,d\theta = \frac{1}{2\pi}\int_{-\pi}^{\pi} S(a)u\frac{\sigma\sigma^*}{aa^*}d\theta + \varphi\|\sigma\|^2.$$

However, by Lemma 6.5, (u,v) also satisfies (6.4), and hence, since $\|\sigma\| > 0$, we must have $\varphi = 0$.

Consequently, $T_{(a,\sigma)}\mathcal{Q}_n(w) \cap T_{(a,\sigma)}\mathcal{Q}_n(r)$ consists of those $(u,v) \in V_{n-1} \times V_{n-1}$ which satisfy both

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{S(\sigma)v}{aa^*}e^{ik\theta}d\theta = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{S(a)u}{aa^*}\frac{\sigma\sigma^*}{aa^*}e^{ik\theta}d\theta, \quad k = 0, 1, \dots, n, \tag{6.10}$$

and

$$av - \sigma u = \rho, \quad \deg\rho \le n-1. \tag{6.11}$$

In view of (6.5), we have

$$\frac{S(\sigma)v}{aa^*} = \frac{S(a)u}{aa^*}\frac{\sigma\sigma^*}{aa^*} + \frac{S(a\sigma)\rho}{(aa^*)^2},$$

which, inserted into (6.10), yields

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{S(a\sigma)\rho}{(aa^*)^2}e^{ik\theta}d\theta = 0, \quad k = 0, 1, \dots, n. \tag{6.12}$$

Clearly, there is a decomposition

$$\frac{S(a\sigma)\rho}{(aa^*)^2} = \frac{d}{a^2} + \frac{d^*}{(a^*)^2} = \frac{S(a^2)d}{(aa^*)^2}, \tag{6.13}$$

where $d(z)$ is a real polynomial of degree at most $2n$. Since $a(z)$ has all of its roots in the open unit disc, there is also a Laurent expansion

$$\frac{d(z)}{a(z)^2} = \frac{1}{2}\beta_0 + \sum_{j=1}^{\infty}\beta_j z^{-j}$$

valid on the unit circle, having real coefficients $\beta_0, \beta_1, \beta_2, \dots$, in terms of which

$$\frac{S(a\sigma)\rho}{(aa^*)^2} = \sum_{j=-\infty}^{\infty}\beta_j e^{-ij\theta},$$

where $\beta_{-j} = \beta_j$ for all j. Inserting this into (6.12), we see that $\beta_0, \beta_1, \dots, \beta_n = 0$, and hence the polynomial $d(z)$ has degree at most $n-1$, precisely as $\rho(z)$ has.

Now from (6.13) we also have

$$S(a\sigma)\rho = S(a^2)d$$

or, equivalently,

$$a(\sigma^*\rho - a^*d)^* + a^*(\sigma^*\rho - a^*d) = 0.$$

Introducing the reversed polynomials $a_*(z) := z^n a(z^{-1})$ and $\sigma_*(z) := z^n\sigma(z^{-1})$, we may write this as

$$S(z^n a)(\sigma_*\rho - a_*d) = 0,$$

which is well defined since $\deg(\sigma_*\rho - a_*d) = 2n - 1 < \deg(z^n a)$. Then, since the polynomial $z^n a$ has all of its roots in the open unit disc, $\ker S(z^n a) = 0$, and hence

$$\sigma_*\rho = a_*d. \tag{6.14}$$

Now, if $\rho \neq 0$,

$$\frac{d_*}{\rho_*} = \frac{\sigma}{a},$$

where $\rho_*(z) := z^{n-1}\rho(z^{-1})$ and $d_*(z) := z^{n-1}d(z^{-1})$. But this is impossible when $a(z)$ and $\sigma(z)$ are coprime because the left member is a proper rational function of degree at most $n-1$, while the right member has degree n. Hence only $\rho = d = 0$ satisfies (6.14). However, for $\rho = 0$, (6.11) has only the solution $u = v = 0$, as claimed. In fact, if $v \neq 0$,

$$\frac{v}{u} = \frac{\sigma}{a},$$

which has no solution if $a(z)$ and $\sigma(z)$ are coprime. □

Just as in Remark 4.7, this establishes that the Jacobian of the joint map $(a, \sigma) \to (r_1, r_2, \ldots, r_n, w_1, w_2, \ldots, w_n)$ has full rank, and, by the inverse function theorem, the joint map forms a smooth local coordinate system on $\mathcal{Q}_n^*$. This proves the first statement of Theorem 3.5.

Figure 6 illustrates the fact that the covariance foliation and the Markov foliation are everywhere transverse. Also note that the shaded region in Figure 6 is identical to Figure 5, thus illustrating Corollary 6.4.

Figure 6 also suggests that each leaf of the Markov foliation meets each leaf of the covariance matching foliation, a fact that we shall now establish in a slightly generalized form. As above, $\overline{\mathcal{Q}_n(r)}$ and $\overline{\mathcal{Q}_n(w)}$ denote the closures of the submanifolds $\mathcal{Q}_n(r)$ and $\mathcal{Q}_n(w)$, respectively.

THEOREM 6.8. *The closure of every leaf of the Markov foliation intersects the closure of any leaf of the covariance matching foliation. Moreover, either the leaves themselves intersect, or every point of intersection is of the form (a, σ), where a has some roots on the unit circle and σ vanishes at each of these roots, while the ratio has the prescribed covariance and Markov windows.*

Proof. The basic space we work on is the product $\overline{\mathcal{S}_n} \times \Pi_n$. We have already seen that $\overline{\mathcal{Q}_n(w)}$ is the graph of a continuous function $\gamma : \overline{\mathcal{S}_n} \to \Pi_n$. We wish to exhibit $\overline{\mathcal{Q}_n(r)}$ as the graph of a continuous function $\delta : \Pi_n \to \overline{\mathcal{S}_n}$. Assuming this for the moment, we deduce from Theorem 6.1 that the continuous map

$$\delta \circ \gamma : \overline{\mathcal{S}_n} \to \overline{\mathcal{S}_n}$$

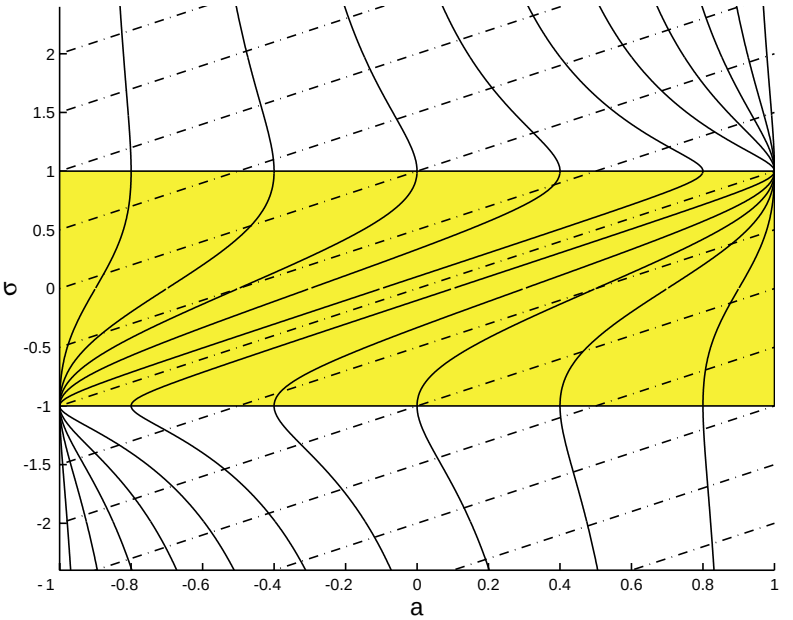


FIG. 6. *Markov (dotted line) and covariance (solid line) matching foliations of* $\mathcal{Q}_1$.

has a fixed point $\bar{a}$; i.e., $(\delta \circ \gamma)(\bar{a}) = \bar{a}$. If $\bar{\sigma} = \gamma(\bar{a})$, then $(\bar{a}, \bar{\sigma})$ is a point lying on both $\overline{\mathcal{Q}_n(w)}$ and $\overline{\mathcal{Q}_n(r)}$. To see this, note that $(\bar{a}, \bar{\sigma}) = (\bar{a}, \gamma(\bar{a}))$ by definition and that $(\bar{a}, \bar{\sigma}) = (\delta \circ \gamma(\bar{a}), \gamma(\bar{a}))$ by construction.

Therefore, it remains to construct δ. If σ is a Schur polynomial, then, according to Theorem 3.3, there exists a unique Schur polynomial a such that (a, σ) lies in $\mathcal{P}_n(r) \subset \mathcal{Q}_n(r)$. We shall write $\delta(\sigma) = a$. According to Theorem 3.4, δ is a smooth function on $\mathcal{S}_n$. Since this is crucial for what follows, we give an independent proof of Theorem 3.4, using the global analysis developed in section 4.

First, we note that the foliations $\{\mathcal{P}_n(r) \mid r \in \mathcal{R}_n\}$ and $\{\mathcal{P}_n(\sigma) \mid \sigma \in \mathcal{S}_n\}$ are complementary. To see this, we ask whether a tangent vector $(u, 0)$ to $\mathcal{P}_n(\sigma)$ at a point (a, σ) could also be tangent to the leaf $\mathcal{P}_n(r)$ through (a, σ). To this end, just as in the proof of Proposition 4.5, we first observe that (6.7) may be written as

$$Fp = Hv,$$

where $p(z) := u(z) + \frac{1}{2}\varphi a(z)$ and F, H are the linear maps (4.12). Then, substituting $(u, 0)$ into (6.7), we obtain $Fp = 0$. However, we also established in the proof of Proposition 4.5 that F is nonsingular, and hence $p = 0$, which, in turn, implies that $\varphi = 0$ and thus that $u = 0$.

Now consider the map $\eta : \mathcal{P}_n \to \mathcal{S}_n$ defined via $\eta(a, \sigma) = \sigma$. The kernel of the Jacobian of η at any point is the tangent space to $\mathcal{P}_n(\sigma)$ at that point. In particular, the kernel of the Jacobian of the map $\eta_r : \mathcal{P}_n(r) \to \mathcal{S}_n$ defined via $\eta_r(a, \sigma) = \sigma$ is zero at every point of $\mathcal{P}_n(r)$. According to Theorem 3.3, the map η_r has an inverse δ. Moreover, by the inverse function theorem, δ is smooth and hence continuous.

In [22], Georgiou proves that δ has a continuous extension to $\overline{\mathcal{S}_n}$ with a very interesting property. If σ has roots on the unit circle, $a = \delta(\sigma)$ may have roots on the unit circle, but σ must vanish at each of these roots, yielding a lower degree

ratio having the prescribed covariance window. Of course, Theorem 3.3 and the constructions in [11, 22] start with the pseudopolynomial

$$d(z, z^{-1}) = \sigma(z)\sigma(z^{-1})$$

rather than σ itself. Since d is taken to be an arbitrary pseudopolynomial of degree less than or equal to zero and nonnegative on the unit circle, the continuity of δ on $\overline{\mathcal{S}_n}$ is equivalent to the continuity of δ on the larger space Π_n. This enables us to form the continuous function $\delta \circ \gamma$ on $\overline{\mathcal{S}_n}$ and apply the Lefschetz fixed point theorem, yielding the statement of the theorem. □

Since, according to Theorem 6.8, any intersection between $\overline{\Omega_n(r)}$ and $\overline{\Omega_n(w)}$ on the boundary of Ω_n defines a pair (a, σ) of polynomials whose roots on the unit circle are common, after cancellation, $w(z) = \sigma(z)/a(z)$ has all of its poles in open unit disc and is thus a bona fide shaping filter. Consequently, Theorem 6.8 establishes the existence part of the last statement of Theorem 3.5. The uniqueness part follows from the following proposition.

PROPOSITION 6.9. *There is at most one shaping filter $w(z)$ having given windows $(1, w_1, \ldots, w_n)$ and $(1, r_1, \ldots, r_n)$ of normalized Markov parameters and normalized covariance lags, respectively.*

Proof. Let $w_1(z)$ and $w_2(z)$ be two shaping filters having the same window $(1, w_1, \ldots, w_n)$ of normalized Markov parameters. Then, if

$$w_1(z) = \frac{\sigma_1(z)}{a_1(z)}, \qquad w_2(z) = \frac{\sigma_2(z)}{a_2(z)},$$

where (a_1, σ_1) and (a_2, σ_2) are coprime pairs of monic polynomials, the degree of the polynomial

$$\rho := \sigma_1 a_2 - \sigma_2 a_1$$

is at most $n - 1$. In fact, the first n Markov parameters of

$$\frac{\sigma_1}{a_1} - \frac{\sigma_2}{a_2} = \frac{\rho}{a_1 a_2}$$

are zero.

Without restriction, we may order the shaping filters so that $\lambda_1 \geq \lambda_2$, where

$$\lambda_1 := \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{\sigma_1}{a_1} \right|^2 d\theta \right)^{-1}, \quad \lambda_2 := \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{\sigma_2}{a_2} \right|^2 d\theta \right)^{-1}.$$

Then, assuming that $w_1(z)$ and $w_2(z)$ also have the same normalized covariance lags $(1, r_1, \ldots, r_n)$, we have

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\theta} \Psi(e^{i\theta}) d\theta = 0, \quad 0, 1, \ldots, n,$$

where

$$\Psi := \lambda_1 \left| \frac{\sigma_1}{a_1} \right|^2 - \lambda_2 \left| \frac{\sigma_2}{a_2} \right|^2.$$

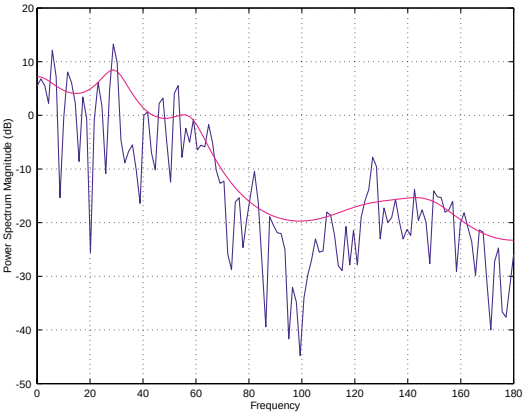


FIG. 7. *Spectral envelope of* 10*th order LPC filter.*

We want to show that $\rho = 0$. To this end, note that, in particular,

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} |a_2(e^{i\theta})|^2 \Psi(e^{i\theta}) d\theta = 0, \tag{6.15}$$

where

$$|a_2|^2\Psi = \lambda_1 \frac{|\rho|^2}{|a_1|^2} + \lambda_1 \frac{\sigma_2 \rho^*}{a_1^*} + \lambda_1 \frac{\sigma_2^* \rho}{a_1} + (\lambda_1 - \lambda_2)|\sigma_2|^2.$$

However, for the same reason as in (6.6),

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \left[\sigma_2 \left(\frac{\rho}{a_1}\right)^* + \sigma_2^* \left(\frac{\rho}{a_1}\right)\right] d\theta = 0,$$

and hence (6.15) can be written as

$$\left\| \frac{\rho}{a_1} \right\|^2 + \left(1 - \frac{\lambda_2}{\lambda_1}\right) \|\sigma_2\|^2 = 0.$$

Since $\|\sigma_2\| > 0$ and $1 - \lambda_2/\lambda_1 > 0$, this implies that $\lambda_1 = \lambda_2$ and $\rho = 0$. Hence $w_1 = w_2$, as claimed. □

7. Zero assignability vs. cepstral assignability. The theory derived in this paper was developed for dealing with problems encountered in applying Theorem 3.3 to the identification of speech segments. The maximum entropy solution described in section 3, often called the LPC method in the speech processing community, is a standard tool for representing the spectral envelope of speech signals [17]. Its popularity is mainly due to its low computation costs and nice matching of spectral peaks. The latter property is illustrated in Figure 7, which shows the periodogram of Figure 3 together with the spectral envelope determined by a tenth order LPC filter, based on ergodic estimates of $r_0, r_1, \ldots, r_{10}$ from the data in Figure 2.

However, it is well known that this estimate of the spectral envelope may not reproduce the notches of the spectrum very well, especially for nasal sounds, where the spectra have a deep valley because of the dead end formed by the mouth. This "flatness" of the spectral envelope, illustrated by Figure 7, is one of the shortcomings

of LPC filtering. It is due to the fact that the zeros of the modeling filter, being at the origin, are maximally removed from the unit circle, where the spectral density is evaluated. There is thus a need for introducing nontrivial zeros in the shaping filter.

By Theorem 3.3, to any Schur polynomial $\sigma(z)$, there is a unique shaping filter having $\sigma(z)$ as its numerator polynomial and matching the covariance window $r_0, r_1, \dots, r_n$ in the same way as the LPC filter. In fact, there is even a convex optimization procedure, based on (5.7), to determine this shaping filter. However, this does leave us with the problem of how to choose the zeros.

It is generally agreed that a finite window (1.9) of cepstral coefficients contains more information about the zeros than does a finite window (1.8) of covariance lags. In fact, differentiate the expansion

$$\log \frac{\sigma(z)}{a(z)} = \frac{c_0}{2} + \sum_{k=1}^{\infty} c_k z^{-k},$$

obtained from (6.3), with respect to z to obtain

$$\frac{\sigma'(z)a(z) - \sigma(z)a'(z)}{\sigma(z)a(z)} = -\sum_{k=1}^{\infty} k c_k z^{-k-1}. \tag{7.1}$$

Consequently, $\{-kc_k\}$ are the Markov parameters of a filter whose poles are the original poles and zeros. Therefore, modulo deciding which are which, both the poles and the zeros can be determined from a finite number of *exact* cepstral coefficients by solving a Hankel system. In so-called homomorphic prediction, e.g., the method of Shanks [35], the zeros are estimated according to these principles once the poles have been determined using LPC analysis. Indeed, it is well known [32] that the LPC envelope has a nonuniform spectral weighting and that it matches the peaks much more accurately than the valleys, i.e., giving much better estimates of poles than zeros. While, in theory, these methods provide estimates of a shaping filter, and hence of a spectral envelope, they do not achieve covariance matching and may produce shaping filters that are neither stable nor minimum-phase. Therefore, these ad hoc methods do not as such provide an alternative to an algorithm based on Theorem 3.3, but they could provide the required zero estimates.

In this context, we suggest an alternative method for estimating the zeros: Given estimates of spectral values of a periodogram at equidistant points on the unit circle,

$$\Phi(e^{i\theta_k}), \quad k = 1, 2, \dots, N, \tag{7.2}$$

find, by linear programming, pseudopolynomials P and Q which minimize

$$\max_k |Q(e^{i\theta_k})\hat{\Phi}(e^{i\theta_k}) - P(e^{i\theta_k})| \tag{7.3}$$

subject to the constraints that $|P(e^{i\theta_k}| \geq \epsilon$ and $|Q(e^{i\theta_k}| \geq \epsilon$ for some $\epsilon > 0$. Again, the shaping filter P/Q obtained in this way would have the same undesirable properties describe above, but we can use P as the pseudopolynomial required in the dual problem (5.7) to determine a new Q such that P/Q satisfies the covariance matching condition. In this procedure, the Q obtained via (7.3) can be used as an initial condition when applying Newton's method to solve the dual problem. For all the reasons described above, it is better to use a *cepstrally smoothed periodogram* in determining (7.2). Explicitly, the cepstral parameters are calculated from the data (1.9) using an

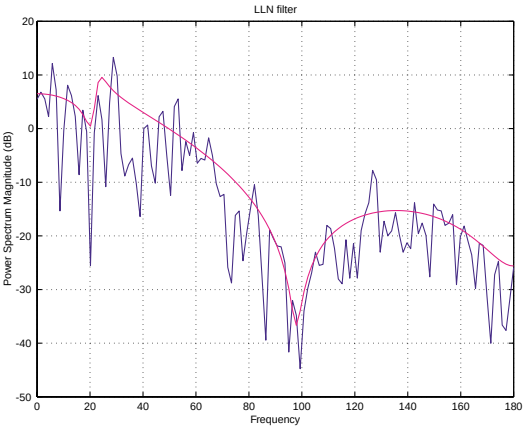


FIG. 8. *Spectral envelope of a* 6*th order LLN filter.*

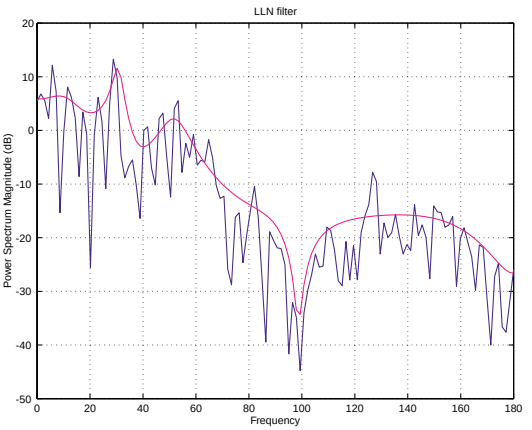


FIG. 9. *Spectral envelope of* 10*th order LLN filter.*

inverse discrete Fourier transform on the logarithm of the periodogram, after which the cepstral coefficients are windowed and inversely transformed [33, pp. 494–495]. As we have seen, the logarithm evens out the difference of energy in the valleys and the peaks and then treats valleys and peaks the same. In Figure 8, we show the spectral envelope of the signal in Figure 2 obtained from a sixth order shaping filter computed by this method. This spectral envelope should be compared with that of the tenth order LPC filter in Figure 7. Instead using a tenth order filter, we obtain the spectral envelope in Figure 9.

However, instead of matching covariance lags and zeros, we may match covariance lags and cepstral coefficients, thus applying an algorithm based on the dual problem to maximize (5.16) described in Theorem 5.3. The covariance and cepstrum interpolation problem is very appealing since both the covariances and the cepstral parameters can be estimated directly from data using ergodicity. Estimation of covariances is well analyzed (see e.g., the books [28, 36]), whereas the estimation of the cepstrum is a less studied problem. One method based on taking the discrete Fourier transform of the periodogram has been analyzed in, e.g., [20]. Using estimated covariance and cepstrum parameters, the filter depicted in Figure 10 was determined.

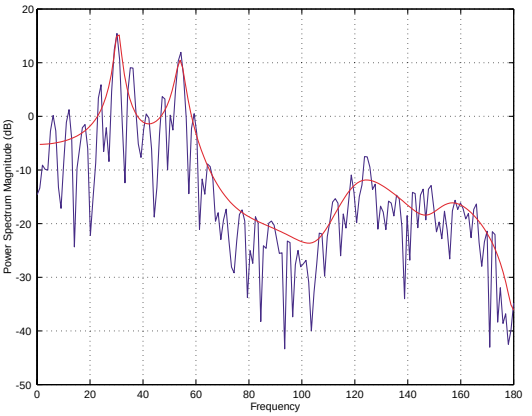


FIG. 10. *Spectral envelope of* 10*th order cepstral match filter.*

More specifically, Figure 10 shows the periodogram of a frame of speech for the phoneme [s] together with a tenth order spectral envelope produced by this method. In this case, $P \in \mathcal{D}_+$, so there is both covariance and cepstral matching. In general, however, this is not the case, as Theorem 5.3 states. This can be seen already in the case when $n = 1$. In Figure 5, the covariance matching foliation (straight lines) is depicted together with the cepstral matching foliation (curved). Clearly, a leaf in one foliation in general does not intersect all leaves in the other. Therefore, methods for determining approximate solutions in the interior $\mathcal{D}_+$ have been developed [19].

The problem that P may tend to the boundary of $\mathcal{D}$ led us to relax the stability constraint of the numerator polynomial σ and hence, in view of the bijection between cepstral and Markov parameters, prompted us to consider the simultaneous partial realization problem of section 7.

Appendix A. Divisors and polynomials. In global analysis, we shall also need to recognize spaces of real polynomials which are diffeomorphic to $\mathbb{R}^n$ as well as certain subsets of polynomials having certain properties, e.g., connectivity, in the relative topology. For this reason, we will adapt the standard treatment of divisors and elementary symmetric functions to the real case.

Let Ω be a self-conjugate, open subset of $\mathbb{C}$, which we take to be path-connected. For such an Ω we denote by $\mathcal{P}_\Omega(n)$ the space of real monic polynomials $p(z)$, of degree n, with all roots lying in Ω. Now the roots of any $p \in \mathcal{P}_\Omega(n)$ determine a self-conjugate, unordered n-tuple $(\lambda_1, \ldots, \lambda_n)$ of points $\lambda_i \in \mathbb{C}$, not necessarily distinct, known as a real *divisor* of degree n on Ω. We denote this divisor by D_p and refer to the space of such divisors as the *real symmetric product* $\Omega^{(n)}$ of Ω.

Alternatively, it is standard to construct the symmetric product $\Omega^{(n)}$ by letting the permutation group S_n on n-letters act on the ordinary Cartesian product Ω^n by permuting the coordinates of n-vectors with entries in Ω. The set of equivalence classes, or orbits of S_n, in the Cartesian product form the points in the symmetric product. In general, the real symmetric product $\Omega^{(n)}$ is always a smooth n-manifold; in fact, $\Omega^{(n)}$ is diffeomorphic to $\mathcal{P}_\Omega(n)$ using the identification

$$(\lambda_1, \ldots, \lambda_n) \to (p_1, \ldots, p_n),$$

where $p(z) = z^n + p_1 z^{n-1} + \cdots + p^n := \prod_{k=1}^n (z - \lambda_k)$. For example, we see that the real symmetric product $\Omega^{(n)}$ for $\Omega = \mathbb{C}$ is diffeomorphic to $\mathbb{R}^n$. For the unit disc,

$\mathbb{D}$, the real symmetric product is diffeomorphic to the space of real Schur polynomials, i.e., those real polynomials satisfying the Schur–Cohn conditions, while for the open left half-plane the real symmetric product is diffeomorphic to the space of those real monic polynomials satisfying the Routh–Hurwitz conditions. Each of these real symmetric products is in turn diffeomorphic with $\mathbb{R}^n$, although not via the standard correspondence given above. Indeed, if $\Omega \subset \mathbb{C}$ is a self-conjugate open subset of the Riemann sphere, with a simple, closed, rectifiable, orientable curve as boundary, then $\mathcal{P}_\Omega(\backslash)$ is diffeomorphic to $\mathbb{R}^n$. As noted in [7], this follows from the Riemann mapping theorem and the corresponding result for the open unit disc $\mathbb{D}$. For $\Omega = \mathbb{D}$ this may be explicitly represented using the real diffeomorphism T of $\mathbb{D}$ to $\mathbb{C}$, defined in polar coordinates via

$$T(r,\theta) = \left(\tan\frac{r\pi}{2}, \theta\right).$$

In general, the projection $P_n : \Omega^n \to \Omega^{(n)}$ is smooth, and any diffeomorphism $T : \Omega_1^{(n)} \to \Omega_2^{(n)}$ is induced by a unique S_n-invariant diffeomorphism $\tilde{T} : \Omega_1^n \to \Omega_2^n$. In particular, if $T : \Omega_1 \to \Omega_2$ is a diffeomorphism, then the induced map $\bar{T} : \Omega_1^{(n)} \to \Omega_2^{(n)}$ defined on divisors of degree n via

$$\bar{T}(\lambda_1, \dots, \lambda_n) = (T(\lambda_1), \dots, T(\lambda_n))$$

is a diffeomorphism. In particular, $\mathcal{P}_\mathbb{D}(n)$ is diffeomorphic with $\mathcal{P}_\mathbb{C}(n)$, which is diffeomorphic to $\mathbb{R}^n$.

Appendix B. Calculation of cepstral coefficients. Suppose

$$\Phi(e^{i\theta}) = \rho^2 \left| \frac{\sigma(e^{i\theta})}{a(e^{i\theta})} \right|^2,$$

where

$$a(z) = z^n + a_1 z^{n-1} + \cdots + a_n$$

and

$$\sigma(z) = z^n + \sigma_1 z^{n-1} + \cdots + \sigma_n$$

are Schur polynomials, i.e., have all of their roots in the open unit disc, and ρ is a real number. Then the cepstral coefficients, i.e., the Fourier coefficients in the expansion (1.5), are given by

$$\begin{aligned} c_0 &= 2\log\rho, \\ c_k &= \frac{1}{k}\{s_k(a) - s_k(\sigma)\}, \qquad k = 1, 2, 3, \dots, \end{aligned}$$

where

$$\begin{aligned} s_k(a) &= p_1^k + p_1^k + \cdots + p_n^k, \\ s_k(\sigma) &= z_1^k + z_1^k + \cdots + z_n^k, \end{aligned}$$

in which $p_1, p_2, \dots, p_n$ are the roots of $a(z)$ and $z_1, z_2, \dots, z_n$ are the roots of $\sigma(z)$.

For the case of maximal entropy (or LPC) filters, we have $z_i = 0$, and the above formula is well known. For pole-zero models, this formula is, to the best of our

knowledge, new but straightforward to derive using the basic algebraic properties of the logarithm.

Moreover, using Newton's identities [15, p. 5], one derives the following recursions:

$$s_k(a) = -ka_k - \sum_{j=1}^{k-1} a_{k-j} s_j(a),$$

$$s_k(\sigma) = -k\sigma_k - \sum_{j=1}^{k-1} \sigma_{k-j} s_j(\sigma).$$

These equations also hold for $k > n$ provided we set $a_k = 0$ and $\sigma_k = 0$ whenever $k > n$.

Appendix C. Connectivity of $\mathcal{P}_n(c)$. We also need to know about various coordinates on $\mathcal{P}_\Omega(n)$ and hence about C^∞ functions. If $f : \Omega^{(n)} \to \mathbb{R}$ is C^∞, then f lifts to a C^∞ function on Ω^n which is S_n-invariant, and, conversely, any C^∞ function on Ω^n which is S_n-invariant descends to a C^∞ function defined on $\Omega^{(n)}$. We denote the algebra of C^∞ functions on $\Omega^{(n)}$ by $\mathcal{C}^\infty[\Omega^{(n)}]$ and the algebra of S_n-invariant C^∞ functions on Ω^n by $\mathcal{C}^\infty[\Omega^n]^{S_n}$. In light of the remarks made above, $\mathcal{C}^\infty[\Omega^{(n)}]$ is canonically isomorphic to $\mathcal{C}^\infty[\Omega^n]^{S_n}$.

Whenever a real diffeomorphism M maps such a domain Ω_1 onto such a domain Ω_2, M commutes with the actions of S_n on Ω_1^n and on Ω_2^n, so that composition with M induces an isomorphism between $\mathcal{C}^\infty[\Omega_2^n]^{S_n}$ and $\mathcal{C}^\infty[\Omega_1^n]^{S_n}$ and hence between $\mathcal{C}^\infty[\Omega_2^{(n)}]$ and $\mathcal{C}^\infty[\Omega_1^{(n)}]$. Therefore, composition with M^{-1} will map generators of $\mathcal{C}^\infty[\Omega_1^n]^{S_n}$ to generators of $\mathcal{C}^\infty[\Omega_2^{(n)}]$.

As an example, consider $\Omega = \mathbb{C}$. Then the algebra of S_n-invariant real polynomials is generated by the coefficients p_i of the polynomials $p(z)$, treated as the points of the real symmetric product. We denote this by writing

$$\mathcal{C}^\infty[\mathbb{C}^{(n)}] = \mathcal{C}^\infty[p_1, \ldots, p_n].$$

Any diffeomorphism of $\mathbb{R}^n$ with itself will give another set of n generators, and, conversely, any other choice of n generators will define a diffeomorphism. Indeed, consider the self-conjugate polynomials in λ,

$$s_k(\lambda) = \lambda_1^k + \cdots + \lambda_n^k,$$

which are invariant under the action of S_n on the n-fold Cartesian product of $\mathbb{C}$. Each $s_k(\lambda)$ lies in $\mathcal{C}^\infty[\mathbb{C}^{(n)}]$ and is in fact a real polynomial in $(\lambda_1, \ldots, \lambda_n)$, as described by the Newton identities [15, p. 5]

$$s_k(\lambda) = -k\lambda_k - \sum_{j=1}^{k-1} \lambda_{k-j} s_j(\lambda),$$

where we set $a_k = 0$ and $\sigma_k = 0$ whenever $k > n$.

Conversely, the Newton identities also show that the λ_i are real polynomials in the s_k, and so we may write

$$\mathcal{C}^\infty[\mathbb{C}^{(n)}] = \mathcal{C}^\infty[\lambda_1, \ldots, \lambda_n].$$

To put this another way, the functions s_k form a system of smooth coordinates on the real Euclidean n-space, $\mathbb{C}^{(n)}$.

If

$$\tilde{c}_k(\tilde{T}(a), \tilde{T}(\sigma)) = c_k(\sigma, a),$$

then the functions $\tilde{c}_k$ form a set of generators for $\mathcal{C}^\infty[\mathbb{C}^{(n)}]$. In particular, in these coordinates, the sets are affine planes and are hence connected.

LEMMA C.1. *The submanifolds $\mathcal{P}_n(c)$ are connected.*

REFERENCES

[1] B. D. O. ANDERSON AND R. E. SKELTON, *q-Markov covariance equivalent realizations*, Internat. J. Control, 44 (1986), pp. 1477–1490.

[2] K. J. ÅSTRÖM AND T. SÖDERSTRÖM, *Uniqueness of the maximum likelihood estimates of the parameters of an ARMA model*, IEEE Trans. Automat. Control, 19 (1974), pp. 769–773.

[3] R. W. BROCKETT, *The geometry of the partial realization problem*, in Proceedings of the 17th IEEE Conference on Decision and Control, San Diego, CA, 1978, pp. 1048–1052.

[4] R. W. BROCKETT, *Some geometric questions in the theory of linear systems*, IEEE Trans. Automat. Control, 21 (1976), pp. 449–455.

[5] C. I. BYRNES, *On the global analysis of linear systems*, in Mathematical Control Theory, J. Baillieul and J. C. Willems, eds., Springer-Verlag, New York, 1999, pp. 99–139.

[6] C. I. BYRNES AND T. E. DUNCAN, *On certain topological invariants arising in system theory*, in New Directions in Applied Mathematics, P. Hilton and G. Young, eds., Springer-Verlag, New York, 1981, pp. 29–71.

[7] C. I. BYRNES AND A. LINDQUIST, *On the geometry of the Kimura-Georgiou parameterization of modelling filter*, Internat. J. Control, 50 (1989), pp. 2301–2312.

[8] C. I. BYRNES, A. LINDQUIST, AND Y. ZHOU, *On the nonlinear dynamics of fast filtering algorithms*, SIAM J. Control Optim., 32 (1994), pp. 744–789.

[9] C. I. BYRNES, T. T. GEORGIOU, AND A. LINDQUIST, *A generalized entropy criterion for Nevanlinna-Pick interpolation with degree constraint*, IEEE Trans. Automat. Control, 46 (2001), pp. 822–839.

[10] C. I. BYRNES AND A. LINDQUIST, *On the partial stochastic realization problem*, IEEE Trans. Automat. Control, 42 (1997), pp. 1049–1069.

[11] C. I. BYRNES, A. LINDQUIST, S. V. GUSEV, AND A. V. MATVEEV, *A complete parameterization of all positive rational extensions of a covariance sequence*, IEEE Trans. Automat. Control, 40 (1995), pp. 1841–1857.

[12] C. I. BYRNES, S. V. GUSEV, AND A. LINDQUIST, *A convex optimization approach to the rational covariance extension problem*, SIAM J. Control Optim., 37 (1998), pp. 211–229.

[13] C. I. BYRNES AND A. LINDQUIST, *On the duality between filtering and Nevanlinna–Pick interpolation*, SIAM J. Control Optim., 39 (2000), pp. 757–775.

[14] C. I. BYRNES, P. ENQVIST, AND A. LINDQUIST, *Cepstral coefficients, covariance lags, and pole-zero models for finite data strings*, IEEE Trans. Signal Process., 49 (2001), pp. 677–693.

[15] P. BORWIEN AND T. ERD'ELYI, *Polynomials and Polynomial Inequalities*, Springer-Verlag, New York, 1995.

[16] D. F. DELCHAMPS, *The Geometry of Space of Linear Systems with an Application to the Identification Problem*, Ph.D. thesis, Harvard University, Cambridge, MA, 1982.

[17] PH. DELSARTE, Y. GENIN, Y. KAMP, AND P. VAN DOOREN, *Speech modelling and the trigonometric moment problem*, Philips J. Res., 37 (1982), pp. 277–292.

[18] A. DOLD, *Lectures in Algebraic Topology*, Springer-Verlag, Berlin, 1972.

[19] P. ENQVIST, *Spectral Estimation by Geometric, Topological and Optimization Methods*, Ph.D. thesis, Division of Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology, Stockholm, Sweden, 2001.

[20] Y. EPHRAIM AND M. RAHIM, *On second-order statistics and linear estimation of cepstral coefficients*, IEEE Trans. Speech and Audio Processing, 7 (1999), pp. 162–176.

[21] T. T. GEORGIOU, *Realization of power spectra from partial covariance sequences*, IEEE Trans. Acoustics, Speech and Signal Processing, 35 (1987), pp. 438–449.

[22] T. T. GEORGIOU, *The interpolation problem with a degree constraint*, IEEE Trans. Automat. Control, 44 (1999), pp. 631–635.

[23] W. B. GRAGG AND A. LINDQUIST, *On the partial realization problem*, Linear Algebra Appl., 50 (1983), pp. 277–319.

[24] K. GLOVER AND J. C. WILLEMS, *Parameterizations of linear dynamical systems: Canonical forms and identifiability*, IEEE Trans. Automat. Control, 19 (1974), pp. 640–645.

[25] R. E. Kalman, *Realization of covariance sequences*, in Proceedings of the Toeplitz Memorial Conference, University of Tel Aviv, Tel Aviv, Israel, 1981, pp. 331–342.
[26] H. Kimura, *Positive partial realization of covariance sequences*, in Modelling, Identification and Robust Control, C. I. Byrnes and A. Lindquist, eds., North–Holland, Amsterdam, 1987, pp. 499–513.
[27] P. S. Krishnaprasad, *On the geometry of linear passive systems*, in Algebraic and Geometric Methods in Linear Systems Theory, Lectures in Appl. Math. 18, AMS, Providence, RI, 1980, pp. 253–276.
[28] G. M. Jenkins and D. G. Watts, *Spectral Analysis and Its Applications*, Holden–Day, San Francisco, CA, 1968.
[29] A. M. King, U. B. Desai, and R. E. Skelton, *A generalized approach to Q-Markov covariance equivalent realizations for discrete systems*, Automatica J. IFAC, 53 (1988), pp. 507–515.
[30] A. Lindquist and G. Picci, *Canonical correlation analysis, approximate covariance extension, and identification of stationary time series*, Automatica, 32 (1996), pp. 709–733.
[31] K. Liu and R. E. Skelton, *A new formulation of Q-Markov covariance equivalent realization*, Appl. Math. Comput., 53 (1993), pp. 83–95.
[32] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*, Springer-Verlag, Berlin, 1976.
[33] A. V. Oppenheim and R. W. Shafer, *Digital Signal Processing*, Prentice–Hall, London, 1975.
[34] G. Segal, *The topology of spaces of rational functions*, Acta Math., 143 (1979), pp. 39–72.
[35] J. L. Shanks, *Recursion filters for digital processing*, Geophysics, 32 (1967), pp. 33–51.
[36] T. Söderström and P. Stoica, *System Identification*, Prentice–Hall, Englewood Cliffs, NJ, 1989.
[37] A. Tannenbaum, *Invariance and System Theory: Algebraic and Geometric Aspects*, Springer-Verlag, Berlin, 1981.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 1, pp. 60–82

EFFICIENT CONSTRAINED MODEL PREDICTIVE CONTROL WITH ASYMPTOTIC OPTIMALITY*

MARK CANNON† AND BASIL KOUVARITAKIS†

Abstract. A computationally inexpensive model predictive control strategy for constrained linear systems is presented. We describe an efficiently computed suboptimal control law which is exponentially stabilizing in the presence of constraints and which converges asymptotically to the conditions for constrained optimality with respect to the receding horizon optimization. The free parameters in input predictions are adapted online on the basis of the gradient of the predicted performance index and the boundary of the admissible set for an autonomous prediction system. A differential description of the admissible set boundary enables efficient detection of active constraints. The approach is illustrated via simulation examples.

Key words. receding horizon control, constrained control, optimality conditions, asymptotic convergence

AMS subject classifications. 49N35, 90C46, 93D05, 65K10

PII. S0363012999358373

1. Introduction.

Model predictive control (MPC) is a feedback strategy in which a control law is computed by repeatedly minimizing a predicted performance index online. MPC algorithms have proved effective in a wide range of commercial applications [19] primarily because they enable constraints on inputs and states to be handled systematically. For the case of linear dynamics and quadratic cost considered in this paper, the online optimization subject to linear input/state constraints is a quadratic programming (QP) problem, which can be solved efficiently and reliably using commercially available software. However, the computational burden of QP can be prohibitive, particularly for high-dimensional systems or rapid sampling applications, and considerable research effort has recently been devoted to reducing this burden [20, 7, 8, 3, 22, 2]. The objective of this paper is to develop an efficient MPC strategy suitable for millisecond sample intervals and small or medium-scale problems (up to, say, 10 states/inputs).

The computational burden of MPC can be reduced significantly if the receding horizon optimization is replaced by an approximate problem of reduced complexity. In [3, 22, 13], the feasible sets for the optimization variables are approximated via ellipsoids, thus reducing the online optimization to a 2-norm minimization subject to a single quadratic constraint which can be solved extremely efficiently. However, [3] and [22] are necessarily suboptimal, and although suboptimality can be reduced to insignificant levels through scaling [13], the use of ellipsoidal constraint set approximations can lead to conservative stabilizable initial condition sets. Larger stabilizable sets are obtained with low-complexity polytopic sets in place of ellipsoids [16] but only at the expense of increased online computation, which then becomes a linear programming (LP) problem.

*Received by the editors June 25, 1999; accepted for publication (in revised form) October 18, 2001; published electronically March 27, 2002. This work was supported by the Engineering and Physical Sciences Research Council.
http://www.siam.org/journals/sicon/41-1/35837.html
†Department of Engineering Science, University of Oxford, Parks Road, Oxford, OX1 3PJ, UK (mark.cannon@eng.ox.ac.uk, basil.kouvaritakis@eng.ox.ac.uk).

An alternative approach computes the optimal solution of the MPC cost minimization using algorithms tailored to specific applications, thus obtaining computational savings over generic solvers. For example, [7] describes a customized MPC algorithm for paper machine cross-directional control which uses physical insights into the problem of controlling web-forming processes to initialize the solution of the receding horizon optimization. In the same application area, [20] derives an efficient QP solver by exploiting the banded structure and sparsity of data matrices, which results from retaining predicted states as well as inputs as variables in the receding horizon optimization problem. However, like [7], this approach is highly specialized since it introduces additional variables into the online QP, and, as a result, its computational advantages are limited to large-scale problems (100 decision variables or more).

Efficient MPC algorithms with wider applicability can be derived by customizing the online optimization using parametric programming methods. Parametric programming is concerned with finding the solution of an optimization problem which is a perturbation of another problem for which a solution is known. The technique is particularly useful in MPC because of the similarity between the receding horizon optimizations solved at successive sampling instants [8]; in fact, "warm starting" (where a previous solution is used to initialize the current optimization) is employed in many commercial MPC algorithms. In [2], parametric programming is used to compute offline the solution of the online MPC optimization as a piecewise linear state feedback law, the parameters of which are state-dependent. Although this approach provides an explicit expression for the state feedback law associated with MPC, its applicability is likely to be limited by the large computational burden concurrent with identifying the regions of state-space on which each of the possible combinations of constraints are active at the solution of the MPC optimization.

In this paper, we use a parametric programming technique based on a differential description of the boundary of the feasible set for the plant state and the free parameters in predicted inputs. This enables efficient online computation of the active constraints corresponding to the current plant state and prediction parameters and hence simplifies the problem of determining a feasible control trajectory and feasible update direction for prediction parameters. The characterization of the feasible set is obtained by expressing input predictions as the sum of a stabilizing unconstrained state feedback law and a linear expansion over a set of exponential basis functions. Predicted input trajectories are governed by an autonomous system with initial state given by the plant state augmented by prediction parameters, and the feasible set is therefore given by the maximal admissible set (as defined in [10]) for the autonomous prediction system state.

Starting from a feasible but suboptimal point computed offline, the proposed algorithm successively updates the prediction parameters on the basis of the gradient of the predicted cost subject to input/state constraints. The approach is similar to the strategy of "early termination" of an optimization routine at a feasible but suboptimal point, which is used extensively in commercial MPC algorithms to reduce online computational burden [19]. The stability properties of MPC algorithms employing early termination are well known: closed-loop stability is unaffected by the suboptimality of predicted input trajectories provided that future feasibility is ensured and the predicted cost decreases sufficiently rapidly along closed-loop trajectories [18, 21]. However, in this paper, we make further use of the analysis of the closed-loop convergence of the predicted cost in order to derive update laws for prediction parameters which steer the closed-loop system asymptotically to a point satisfying the conditions for constrained optimality.

The current paper extends the algorithm proposed in [4] in respect of both the detection of active constraints and the prediction parameter update law. Here we make explicit use of the convexity of the feasible set to derive a simpler algorithm for constraint detection based on the intersections of the boundary of this set with a sequence of linear subspaces. Furthermore, we characterize the prediction parameter update law in terms of a simplified subproblem and describe an efficient method for its solution.

The paper is organized as follows. Section 2 defines the control problem and gives the autonomous formulation of prediction dynamics. Sections 3 and 4 describe the feasible set and the method of online constraint detection. The prediction parameter update law and its closed-loop optimality properties are described in section 5, and simulation examples are presented in section 6.

2. Prediction dynamics. Consider the linear plant dynamics described by the model

$$\dot{x}(t) = A_m x(t) + B_m u(t) \quad \forall t \geq 0, \tag{1}$$

where $x \in \mathbb{R}^{n_x}$ and $u \in \mathbb{R}^{n_u}$ is the input. This paper is concerned with optimal regulation of (1) subject to input constraints of the form

$$u \in \mathcal{U},\ \mathcal{U} = \big\{u;\ \underline{U} \leq u(t) \leq \overline{U},\ \underline{U}' \leq \dot{u}(t) \leq \overline{U}' \quad \forall t \geq 0\big\}. \tag{2}$$

Note that the approach described below is also applicable to systems with polytopic state constraints and that tracking problems can be handled analogously. The control law is developed with the objective of approximating the solutions of the optimization problem

(3)
$$\begin{array}{ll} \underset{\hat{u}_t}{\text{minimize}} & J(t),\ \ J(t) = \displaystyle\int_0^\infty \big(\hat{x}_t^T(\tau) Q \hat{x}_t(\tau) + \hat{u}_t^T(\tau) R \hat{u}_t(\tau)\big)\, d\tau,\ \ Q \geq 0,\ \ R > 0, \\ \text{subject to} & \dot{\hat{x}}_t(\tau) = A_m \hat{x}_t(\tau) + B_m \hat{u}_t(\tau) \quad \forall \tau \geq 0,\ \ \hat{x}_t(0) = x(t), \\ & \hat{u}_t \in \mathcal{U}. \end{array}$$

Here $\hat{x}_t$ and $\hat{u}_t$ are predictions at time t of the plant state and input on the interval $[t, \infty)$, and to ensure the existence of a stabilizing solution to (3) for some nontrivial set of initial plant states, we assume that $(A_m, B_m, Q^{1/2})$ is controllable and observable. We define a continuous-time receding horizon control law by setting $u(t) = \hat{u}_t(0)$, where $\hat{u}_t$ is determined continuously in t on the basis of a gradient descent algorithm.

To make the constrained optimization (3) tractable, we restrict the predicted input $\hat{u}_t$ to a finite-dimensional class. In order to avoid numerical sensitivity in predictions, this class is chosen to include a control law, which stabilizes the plant in the absence of constraints. Furthermore, it is often desirable to include the optimal control for the unconstrained dynamics (1) in the prediction class since this necessarily becomes feasible with respect to (3) at some future time under any control law which asymptotically stabilizes the constrained plant. Input predictions are therefore specified as

$$\hat{u}_t(\tau) = K \hat{x}_t(\tau) + \Phi(\tau) c(t) \quad \forall \tau \geq 0, \tag{4}$$

where $\hat{u}_t = K\hat{x}_t$ is the solution of the following linear-quadratic problem: $\text{minimize}_{\hat{u}_t} J(t)$ subject to $\dot{\hat{x}}_t(\tau) = A_m \hat{x}_t(\tau) + B_m \hat{u}_t(\tau)$ for all $\tau \geq 0$, $\hat{x}_t(0) = x(t)$. Also, $c \in \mathbb{R}^{n_c}$

is a vector of prediction parameters, and $\Phi : \mathbb{R} \to \mathbb{R}^{n_u \times n_c}$ is a matrix of fixed basis functions.

The stability of the receding horizon control law proposed in this paper is based on a guarantee of the feasibility of (3) given past feasibility. This in turn is ensured if the input prediction class employed at each instant t contains the extension to time t of the prediction associated with the input implemented at a previous instant $t - \delta$, $\delta > 0$ since this satisfies constraints by assumption. Appendix A shows that this condition, applied in the limit as $\delta \to 0$ to avoid conflict with the rate constraints of (2), is satisfied by the linear expansion of (4) if and only if the basis functions Φ have the exponential form

$$\Phi(\tau) = \Phi(0)\mathrm{e}^{A_\phi \tau} \quad \forall \tau \geq 0, \tag{5}$$

for some $A_\phi \in \mathbb{R}^{n_c \times n_c}$.

The predictions implied by (4) and (5) are governed by an autonomous system of order $n = n_x + n_c$ and are given by

$$\begin{aligned} \hat{u}_t(\tau) &= G\mathrm{e}^{A\tau} z(t), \quad G = \begin{bmatrix} \Phi(0) & K \end{bmatrix}, \\ z(t) &= \begin{bmatrix} c(t) \\ x(t) \end{bmatrix}, \quad A = \begin{bmatrix} A_\phi & 0 \\ B_m\Phi(0) & A_m + B_m K \end{bmatrix}. \end{aligned} \tag{6}$$

The choice of A_ϕ affects both the limits on achievable performance and the set of stabilizable plant states under a receding horizon control law based on the predictions of (6). Here we simply note that A_ϕ must be strictly Hurwitz in order that the prediction dynamics are stable, and the eigenvalues of A_ϕ should be chosen so that the maximum distance between the feedback law $K\hat{x}_t$ for any initial plant state x and the function space spanned by the elements of Φ is in some sense minimized.

3. Admissible set. The (maximal) admissible set for the prediction system (6) under the constraint $\hat{u}_t \in \mathcal{U}$ is defined by

$$\Omega = \left\{ z;\ \underline{U} \leq G\mathrm{e}^{A\tau} z \leq \overline{U},\ \underline{U}' \leq GA\mathrm{e}^{A\tau} z \leq \overline{U}' \quad \forall \tau \geq 0 \right\}. \tag{7}$$

This set is central to the receding horizon control problem since it consists of all prediction system initial states for which the constraint $\hat{u}_t \in \mathcal{U}$ is satisfied. The constrained optimization of predicted performance at time t, therefore, corresponds to the minimization of $J(t)$ over the prediction parameters c in the intersection of Ω with the subspace $\left\{ z \in \mathbb{R}^n;\ [0 \quad \mathrm{I}_{n_x}] z = x(t) \right\}$ (where I_{n_x} denotes the identity in $\mathbb{R}^{n_x \times n_x}$). This section gives a parametric description of the admissible set boundary, denoted by $\partial\Omega$, in terms of the prediction times τ at which $\hat{u}_t(\tau)$ reaches constraints.

Clearly, (7) is the intersection of the admissible sets associated with each of the individual constraints $u \geq \underline{U}$, $u \leq \overline{U}$, $\dot{u} \geq \underline{U}'$, $\dot{u} \leq \overline{U}'$. For simplicity we consider only a single constraint on the ith element of u in this section and in section 4 and accordingly redefine Ω as

$$\Omega = \left\{ z;\ g\mathrm{e}^{A\tau} z \leq \overline{u} \quad \forall \tau \geq 0 \right\},$$

where g is the ith row of G and $\overline{u}$ is the ith element of $\overline{U}$. In addition, we assume that (g, A) is observable; this assumption involves no loss of generality since the following arguments apply to the observable subspace alone in the case that (g, A) contains unobservable modes.

Noting that Ω can be equivalently expressed in terms of the intersection

$$\Omega = \bigcap_{\tau\in[0,\infty)} \left\{z;\ g\mathrm{e}^{A\tau}z \le \overline{u}\right\},$$

the following properties are immediately obvious:

(P1) Ω is convex.

(P2) $z \in \partial\Omega$ only if $g\mathrm{e}^{A\tau}z = \overline{u}$ for some $\tau \ge 0$.

(P3) $(g\mathrm{e}^{A\tau})^T$ is normal to $\partial\Omega$ at a point $z \in \partial\Omega$ such that $g\mathrm{e}^{A\tau}z = \overline{u}$.

Property (P3) is the basis of the method described in section 5 of adapting the prediction parameters $c(t)$ so that the prediction system (6) converges asymptotically to the conditions for optimality with respect to (3). While constraints are inactive (i.e., while the initial prediction system state $z(t)$ lies in the interior of Ω), this approach adapts $c(t)$ via $\dot{c}(t)$ in the direction of the gradient of $J(t)$. Alternatively, whenever $z(t)$ lies in the boundary of Ω, satisfaction of the constraint $\hat{u}_t \in \mathcal{U}$ is ensured by specifying $\dot{c}(t)$ as a function of the gradient of $J(t)$ projected into the subspace of $\mathbb{R}^n$ which is tangent to $\partial\Omega$ at $z(t)$. The required projection is determined from (P3) and the prediction times τ corresponding to active constraints, namely the values of τ such that $g\mathrm{e}^{A\tau}z(t) = \overline{u}$.

Active constraints can be detected by tracking points $\xi(t)$ lying in the boundary of Ω and the corresponding values of $\tau(t)$ satisfying $g\mathrm{e}^{A\tau(t)}\xi(t) = \overline{u}$, where $\xi(t)$ is such that $z(t) \in \partial\Omega$ if and only if $z(t) = \xi(t)$. This approach, which is described in section 4, is based on (P2) and the additional property that Ω is positively invariant for the prediction dynamics (6). Specifically, the positive invariance of Ω implies that a point $z \in \partial\Omega$ such that $g\mathrm{e}^{A\tau}z = \overline{u}$ for some $\tau > 0$ must also satisfy $gA\mathrm{e}^{A\tau}z = 0$, and z, therefore, lies on a trajectory of (6) which is tangent to the hyperplane $\left\{z;\ gz = \overline{u}\right\}$. Defining $\mathcal{G}_1$ as the set of trajectories of (6) that are tangent to $\left\{z;\ gz = \overline{u}\right\}$ so that

$$\mathcal{G}_1 = \left\{z;\ g\mathrm{e}^{A\tau}z = \overline{u},\ gA\mathrm{e}^{A\tau}z = 0,\ \tau \ge 0\right\},$$

we therefore have the following additional property:

(P4) $\partial\Omega \subset \left\{z;\ gz = \overline{u}\right\} \cup \mathcal{G}_1$.

Property (P4) allows the active constraints associated with a given prediction system state $z(t)$ to be determined by checking whether $gz(t) = \overline{u}$ or $z(t) \in \mathcal{G}_1$.

4. Constraint detection. This section describes a mechanism for determining the set of active constraints as the prediction parameters are adapted in continuous time. As in section 2, for simplicity we consider here a single constraint in the form of an upper bound $\overline{u}$ on an element of u. The approach is based on tracking points in the state-space of the prediction system (6) lying in the set $\mathcal{G}_1$ or the hyperplane $\left\{z;\ gz = \overline{u}\right\}$. Before giving the details of the method, we first determine the smoothness properties of $\mathcal{G}_1$.

From the total derivatives of $g\mathrm{e}^{A\tau}z$ and $gA\mathrm{e}^{A\tau}z$, the tangent vector $(dz, d\tau)$ to $\mathcal{G}_1$ (considered as a hypersurface in $\mathbb{R}^n \times \mathbb{R}^+$) at $(z,\tau) \in \mathcal{G}_1$ satisfies

$$\begin{aligned} g\mathrm{e}^{A\tau}\,dz + gA\mathrm{e}^{A\tau}z\,d\tau = g\mathrm{e}^{A\tau}\,dz = 0,\\ gA\mathrm{e}^{A\tau}\,dz + gA^2\mathrm{e}^{A\tau}z\,d\tau = 0. \end{aligned}$$

For any $dz \in \mathbb{R}^n$, it is possible to find $d\tau$ satisfying the second equation whenever $gA^2\mathrm{e}^{A\tau}z \ne 0$. The first equation, therefore, implies that the normal vector to $\mathcal{G}_1$ (considered as a hypersurface in $\mathbb{R}^n$) is given by $(g\mathrm{e}^{A\tau})^T$ at any point $z \in \mathcal{G}_1$ such

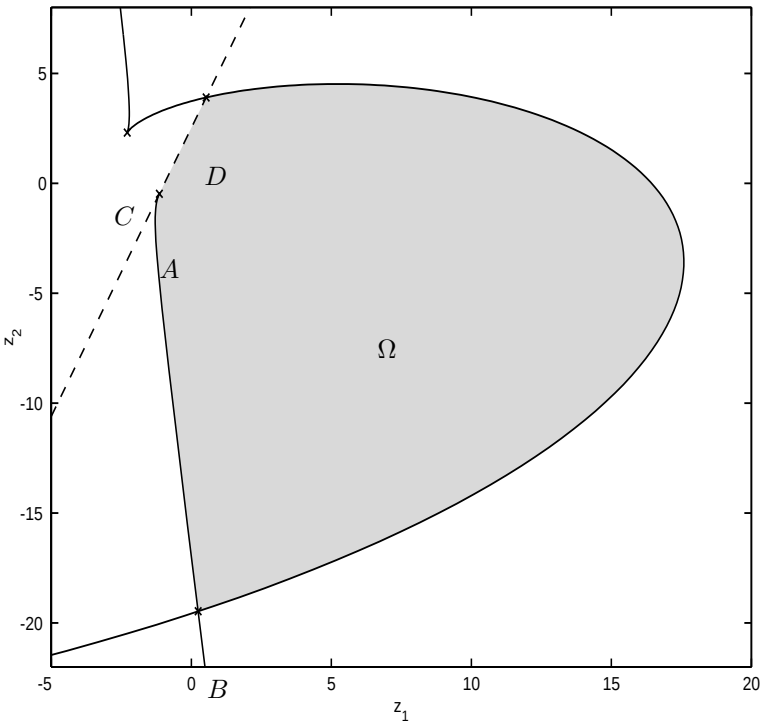


FIG. 1. *Example of the intersection of the admissible set Ω with the z_1, z_2-plane for a prediction system of order $n = 3$ with constraint $\overline{u} = 1$. The boundary $\partial\Omega$ consists of points lying in $\mathcal{G}_1$ (solid line) and the hyperplane $\{z;\ gz = \overline{u}\}$ (dashed line). Also shown are the following: A is an initial prediction system state $z \in \mathcal{G}_1$ such that $gz = \overline{u}$ and $gAz = 0$; B is a state $z \in \mathcal{G}_1$ satisfying $g\mathrm{e}^{A\tau_1}z = g\mathrm{e}^{A\tau_4}z = \overline{u}$ and $gA\mathrm{e}^{A\tau_1}z = gA\mathrm{e}^{A\tau_4}z = 0$ for some $\tau_1, \tau_4 > 0$; C is a state $z \in \mathcal{G}_2$ satisfying $g\mathrm{e}^{A\tau_2}z = \overline{u}$ and $gA\mathrm{e}^{A\tau_2}z = gA^2\mathrm{e}^{A\tau_2}z = 0$ for some $\tau_2 > 0$; D is a state $z \in \mathcal{G}_1$ satisfying $gz = g\mathrm{e}^{A\tau_3}z = \overline{u}$ and $gA\mathrm{e}^{A\tau_3}z = 0$ for some $\tau_3 > 0$; where $0 < \tau_1 < \tau_2 < \tau_3 < \tau_4$.*

that $gA^2\mathrm{e}^{A\tau}z \neq 0$, which is in agreement with properties (P3) and (P4). From the second equation, we have

$$\frac{\partial\tau}{\partial z} = -\frac{gA\mathrm{e}^{A\tau}}{gA^2\mathrm{e}^{A\tau}z},$$

and by the implicit function theorem, we have a value of τ such that $(z, \tau) \in \mathcal{G}_1$ is, therefore, a smooth (C^∞) function of z provided that $gA^2\mathrm{e}^{A\tau}z \neq 0$. Denoting $\mathcal{G}_2$ as the set

$$\mathcal{G}_2 = \big\{z;\ g\mathrm{e}^{A\tau}z = \overline{u},\ gA\mathrm{e}^{A\tau}z = 0,\ gA^2\mathrm{e}^{A\tau}z = 0,\ \tau \geq 0\big\},$$

it follows that $(g\mathrm{e}^{A\tau})^T$ is a smooth function of z for all $z \in \mathcal{G}_1 - \mathcal{G}_2$, and $\mathcal{G}_1$ is, therefore, a smooth $(n-1)$-dimensional hypersurface in $\mathbb{R}^n$ at every point $z \in \mathcal{G}_1 - \mathcal{G}_2$. A similar argument shows that $\mathcal{G}_2$ has dimension less than or equal to $n-2$, and $\mathcal{G}_2$, therefore, corresponds to a cusp in $\mathcal{G}_1$ in $\mathbb{R}^n$. This situation is illustrated graphically in Figure 1.

More generally, define $\mathcal{G}_k$ by

$$\mathcal{G}_k = \big\{z;\ g\mathrm{e}^{A\tau}z = \overline{u},\ gA\mathrm{e}^{A\tau}z = 0, \ldots,\ gA^k\mathrm{e}^{A\tau}z = 0,\ \tau \geq 0\big\}$$

for $k = 1, 2, \dots, n-1$. Then it is clear that

$$\mathcal{G}_1 \supset \mathcal{G}_2 \supset \cdots \supset \mathcal{G}_{n-1},$$

and Appendix B shows that $\mathcal{G}_k$ is locally a smooth $(n-k)$-dimensional hypersurface in $\mathbb{R}^n$ at every point $z \in \mathcal{G}_k - \mathcal{G}_{k+1}$ for $k = 1, 2, \dots, n-2$. Furthermore, $\mathcal{G}_{n-1}$, which consists of a single trajectory of (6), is everywhere smooth and of dimension 1.

The method of detecting active constraints is based on tracking the points of intersection with $\partial\Omega$ of a linear subspace $\mathcal{L}_1$, containing z, defined by

$$\mathcal{L}_1 = \{\xi;\ V_1(\xi - z) = 0\}.$$

Here $V_1 \in \mathbb{R}^{n-1\times n}$ can be chosen arbitrarily subject to the requirement that $\operatorname{rank}(V_1) = n-1$. The convexity of Ω ensures that a point $z \in \Omega$ lies in the boundary $\partial\Omega$ if and only if z coincides with the point of intersection of $\mathcal{L}_1$ with $\partial\Omega$. From property (P4), it follows that the active constraints can be determined from the intersection points $\mathcal{L}_1 \cap \mathcal{G}_1$ and $\mathcal{L}_1 \cap \{z;\ gz = \overline{u}\}$. Using convexity arguments, we show below that the intersection points $\mathcal{L}_1 \cap \{\mathcal{G}_1 - \mathcal{G}_2\}$ are continuous functions of z and can therefore be tracked in continuous time as z varies in $\Omega - \partial\Omega$. However, it is clear from the preceding discussion that $\mathcal{L}_1 \cap \mathcal{G}_1$ can be discontinuous whenever $\mathcal{L}_1 \cap \mathcal{G}_1 \in \mathcal{G}_2$, and for this reason it is also necessary to detect nonempty intersections of $\mathcal{L}_1$ with $\mathcal{G}_2$. The existence of points of intersection $\mathcal{L}_1 \cap \mathcal{G}_2$ can in turn be determined from the intersections of $\mathcal{G}_2$ with a linear subspace $\mathcal{L}_2$, where

$$\mathcal{L}_2 = \{\xi;\ V_2(\xi - z) = 0\}$$

and V_2 consists of (say) the first $n-2$ rows of V_1; and it can likewise be shown that $\mathcal{L}_2 \cap \mathcal{G}_2$ is continuous if $\mathcal{L}_2 \cap \mathcal{G}_2 \notin \mathcal{G}_3$. The remainder of this section, therefore, derives an algorithm for stable tracking of the intersections $\mathcal{L}_k \cap \mathcal{G}_k$, $k = 1, \dots, n-1$, as z varies in Ω, where $\mathcal{L}_k$ is defined as

$$\mathcal{L}_k = \{\xi;\ V_k(\xi - z) = 0\}$$

and $V_k \in \mathbb{R}^{n-k\times n}$ consists of the first $n-k$ rows of V_{k-1} for $k = 2, \dots, n-1$.

To determine the points of intersection of $\mathcal{L}_1$ with $\mathcal{G}_1$, let W_1 lie in the kernel of V_1. Then a point $\xi \in \mathcal{L}_1$ is a member of $\mathcal{G}_1$ if and only if

$$\xi = z + W_1\alpha \quad \begin{cases} g\mathrm{e}^{A\tau}W_1\alpha = \overline{u} - g\mathrm{e}^{A\tau}z, \\ gA\mathrm{e}^{A\tau}(z + W_1\alpha) = 0 \end{cases}$$

for some $\tau \geq 0$ and $\alpha \in \mathbb{R}$. If z lies in the interior of Ω, then $g\mathrm{e}^{A\tau}z < \overline{u}$ by definition, and it follows that $g\mathrm{e}^{A\tau}W_1 \neq 0$ for any τ corresponding to an intersection point $\mathcal{L}_1 \cap \mathcal{G}_1$. Therefore, $\xi \in \mathcal{L}_1 \cap \mathcal{G}_1$ if and only if $\xi = z + W_1(\overline{u} - g\mathrm{e}^{A\tau}z)/g\mathrm{e}^{A\tau}W_1$, where τ is a nonnegative solution of

$$gA\mathrm{e}^{A\tau}\left[z + W_1(\overline{u} - g\mathrm{e}^{A\tau}z)/g\mathrm{e}^{A\tau}W_1\right] = 0.$$

The following theorem generalizes this argument for the case of $k = 2, \dots, n-1$.

THEOREM 4.1. *Let the columns of W_k form a basis for the kernel of V_k, and, for given $\tau \in \mathbb{R}$, define M_k by*

$$M_k = \begin{bmatrix} g \\ gA \\ \vdots \\ gA^{k-1} \end{bmatrix} \mathrm{e}^{A\tau}.$$

Then, for any $z \in \Omega - \partial\Omega$, $\xi \in \mathcal{L}_k \cap \mathcal{G}_k$ if and only if $\xi = z + W_k\alpha$, where

$$\alpha = (M_k W_k)^{-1}(\overline{u}e_1 - M_k z), \quad e_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^T \in \mathbb{R}^k, \tag{8}$$

and τ is a nonnegative solution of

$$gA^k \mathrm{e}^{A\tau}\left[z + W_k(M_k W_k)^{-1}(\overline{u}e_1 - M_k z)\right] = 0. \tag{9}$$

Proof. At $\xi \in \mathcal{L}_k \cap \mathcal{G}_{k-1}$ we have $\xi = z + W_k\alpha$, and $M_k W_k \alpha = \overline{u}e_1 - M_k z$ for some $\tau \geq 0$. Since an intersection point $\mathcal{L}_k \cap \mathcal{G}_k$ necessarily lies in $\mathcal{G}_{k-1}$, a solution for α must exist. Furthermore, the assumption that $z \in \Omega - \partial\Omega$ implies that $M_k z \neq \overline{u}e_1$, and it follows that $\operatorname{rank}(M_k W_k) = k$. Therefore, every intersection point $\mathcal{L}_k \cap \mathcal{G}_k$ is given by $z + W_k\alpha$, where α is defined by (8) and $\tau \geq 0$ satisfies (9). ∎

We next determine the continuity properties of the intersections $\mathcal{L}_k \cap \mathcal{G}_k$ by considering the derivatives with respect to z of α in (8) and τ satisfying (9).

THEOREM 4.2. *For all $z \in \Omega - \partial\Omega$, an intersection point $\mathcal{L}_k \cap \mathcal{G}_k$ is a continuous function of z if and only if $\mathcal{L}_k \cap \mathcal{G}_k \notin \mathcal{G}_{k+1}$ for $k = 1, \ldots, n-2$; and the intersection $\mathcal{L}_{n-1} \cap \mathcal{G}_{n-1}$ is a continuous function of z.*

Proof. The derivatives with respect to z of α satisfying (8) and solutions τ to (9) are given by

$$\frac{\partial \alpha}{\partial z} = -(M_k W_k)^{-1} M_k, \tag{10a}$$

$$\frac{\partial \tau}{\partial z} = -\frac{gA^k \mathrm{e}^{A\tau}}{gA^{k+1}\mathrm{e}^{A\tau}(z + W_k\alpha)}\left[\mathrm{I} - W_k(M_k W_k)^{-1} M_k\right]. \tag{10b}$$

From Theorem 4.1, $M_k W_k$ in (10a) and (10b) has full rank if z lies in the interior of Ω. It follows that α and τ are continuous functions of z provided that $gA^{k+1}\mathrm{e}^{A\tau}(z + W_k\alpha) \neq 0$, which is equivalent to $\mathcal{L}_k \cap \mathcal{G}_k \notin \mathcal{G}_{k+1}$. In the special case of $k = n-1$, α and τ are continuous functions of z for all $z \in \Omega - \partial\Omega$ since $gA^n \mathrm{e}^{A\tau}(z + W_k\alpha)$ is necessarily nonzero due to the observability of (g, A). ∎

Remark 4.3. Theorem 4.2 shows that, for $k = 1, \ldots, n-2$, the intersection of $\mathcal{L}_k$ with $\mathcal{G}_k$ as z varies continuously in the interior of Ω can be discontinuous if and only if $\mathcal{L}_k \cap \mathcal{G}_k \in \mathcal{G}_{k+1}$. In fact, a nonempty intersection of $\mathcal{L}_k$ with $\mathcal{G}_{k+1}$ is precisely the condition under which new intersection points $\mathcal{G}_k \cap \mathcal{L}_k$ are created or existing intersections $\mathcal{G}_k \cap \mathcal{L}_k$ disappear since $\mathcal{G}_{k+1}$ constitutes a cusp in $\mathcal{G}_k$. However, a consequence of Theorem 4.2 is that changes in the number of intersections $\mathcal{L}_k \cap \mathcal{G}_k$ can be detected given a knowledge of the intersection points $\mathcal{L}_{k+1} \cap \mathcal{G}_{k+1}$: from the definition of $\mathcal{L}_{k+1}$, a point $z + W_{k+1}\alpha \in \mathcal{L}_{k+1} \cap \mathcal{G}_{k+1}$ is coincident with $\mathcal{L}_k$ if and only if $e_{k+1}^T \alpha = 0$ (where $e_k = [0 \quad 0 \quad \cdots \quad 1]^T \in \mathbb{R}^k$).

In order to track an intersection point $\mathcal{L}_k \cap \mathcal{G}_k$, it is sufficient to determine the corresponding prediction time τ satisfying (9) since $\xi \in \mathcal{L}_k \cap \mathcal{G}_k$ can then be computed using $\xi = z + W_k\alpha$ and (8). In the absence of disturbances, this could be done by solving (9) for $z = z(0)$ offline and integrating $\dot{\tau} = (\partial\tau/\partial z)\dot{z}$ online. However, this approach is likely to be numerically unstable, and we propose instead a stable adaptation law of the form

$$\dot{\tau} = -\frac{gA^k \mathrm{e}^{A\tau}\left(\left[\mathrm{I} - W_k(M_k W_k)^{-1} M_k\right]\dot{z} + \gamma_\tau(z + W_k\alpha)\right)}{gA^k \mathrm{e}^{A\tau}\left[\mathrm{I} - W_k(M_k W_k)^{-1} M_k\right] A(z + W_k\alpha)}, \tag{11}$$

where α is given by (8) and $\gamma_\tau > 0$ is an adaptive gain.

THEOREM 4.4. *Under the adaptation law* (11), $gA^k\mathrm{e}^{A\tau}\big[z+W_k(M_kW_k)^{-1}(\overline{u}e_1-M_kz)\big]\to 0$ *exponentially in* t.

Proof. Let $\xi = z + W_k\alpha$, where α is defined for given τ by (8). Then

$$gA^k\mathrm{e}^{A\tau}\big[z+W_k(M_kW_k)^{-1}(\overline{u}e_1-M_kz)\big] = gA^k\mathrm{e}^{A\tau}\xi,$$

and it suffices to show that $gA^k\mathrm{e}^{A\tau}\xi\to 0$ exponentially in t to prove the theorem. The derivative of ξ with respect to t is given by $\dot\xi = \dot z + W_k\dot\alpha$, and, differentiating (8),

$$\begin{aligned}\dot\alpha &= -(M_kW_k)^{-1}M_k\dot z - (M_kW_k)^{-1}M_kA\big[z+W_k(M_kW_k)^{-1}(\overline{u}e_1-M_kz)\big]\dot\tau\\ &= -(M_kW_k)^{-1}M_k(\dot z + A\xi\dot\tau),\end{aligned}$$

we therefore have $\dot\xi = \big[\mathrm{I}-W_k(M_kW_k)^{-1}M_k\big]\dot z - W_k(M_kW_k)^{-1}M_kA\xi\dot\tau$. Hence

$$\begin{aligned}\frac{d}{dt}(gA^k\mathrm{e}^{A\tau}\xi) &= gA^k\mathrm{e}^{A\tau}\dot\xi + gA^{k+1}\mathrm{e}^{A\tau}\xi\dot\tau\\ &= gA^k\mathrm{e}^{A\tau}\big[\mathrm{I}-W_k(M_kW_k)^{-1}M_k\big]\dot z\\ &\quad + gA^k\mathrm{e}^{A\tau}\big[\mathrm{I}-W_k(M_kW_k)^{-1}M_k\big]A\xi\dot\tau,\end{aligned}$$

and since $\dot\tau$ as defined in (11) can be expressed as

$$\dot\tau = -\frac{gA^k\mathrm{e}^{A\tau}[\mathrm{I}-W_k(M_kW_k)^{-1}M_k]\dot z + \gamma_\tau gA^k\mathrm{e}^{A\tau}\xi}{gA^k\mathrm{e}^{A\tau}[\mathrm{I}-W_k(M_kW_k)^{-1}M_k]A\xi},$$

it follows that $d(gA^k\mathrm{e}^{A\tau}\xi)/dt = -\gamma_\tau gA^k\mathrm{e}^{A\tau}\xi$ for all $t\ge 0$. ∎

Remark 4.5. A finite constraint horizon T_{con} such that, for any $z\in\Omega$, $g\mathrm{e}^{A\tau}z = \overline{u}$ only if $\tau\le T_{\mathrm{con}}$ can be determined using the admissible set approximation approach described in [10]. Consequently, the active constraint set can be determined from a finite number of intersection points $\mathcal{L}_k\cap\mathcal{G}_k$ corresponding to prediction times $\tau\in[0,T_{\mathrm{con}}]$.

Theorem 4.4 is the basis of an algorithm for stable tracking of every intersection point $\mathcal{L}_k\cap\mathcal{G}_k$, $k=1,\dots,n-1$, corresponding to $\tau\in[0,T_{\mathrm{con}}]$ as $z(t)$ varies continuously in Ω. The approach (which is summarized in Algorithm 4.6) involves finding all solutions to (9) for $k=1,\dots,n-1$ on the interval $[0,T_{\mathrm{con}}]$ offline and then adapting these solutions online via (11) in response to the prediction state derivative $\dot z(t)$. The appearance of new solutions and the disappearance of existing solutions are handled via Theorem 4.2 and Remark 4.3.

In particular, let τ_i be a solution to (9) for $z=z(t)$ and $k=k_i$, and define α_i as the corresponding value of α in (8). Denote the set of all solutions to (9) for $k=1,\dots,n-1$ such that $\tau_i\in[0,T_{\mathrm{con}}]$ as $\{(\tau_i)_{i\in\mathcal{I}}\}$. Then Theorem 4.4 shows that (11) is an asymptotically stable estimator for each τ_i, $i\in\mathcal{I}$, such that $\dot\tau_i$ is well defined at time t. Alternatively, if τ_i is discontinuous at time t, then Theorem 4.2 shows that $\mathcal{L}_{k_i}\cap\mathcal{G}_{k_i}\in\mathcal{G}_{k_i+1}$, and, therefore, $gA^{k_i+1}\mathrm{e}^{A\tau_i}(z+W_{k_i}\alpha_i)=0$ is necessarily satisfied; in this case, i is removed from the index set $\mathcal{I}$, indicating that the prediction time τ_i is no longer to be tracked. On the other hand, from Remark 4.3 and the continuity of $\mathcal{L}_{n-1}\cap\mathcal{G}_{n-1}$, a new solution on the interval $(0,T_{\mathrm{con}})$ to (9) can appear at time t only if $\mathcal{L}_{k_i}\cap\mathcal{G}_{k_i}\in\mathcal{L}_{k_i-1}$ (which implies $e_{k_i}^T\alpha_i=0$) for some $i\in\mathcal{I}$ at time t, and, in this case, a new solution is given by $\tau=\tau_i$ for $k=k_i-1$. Accordingly, we introduce new indices $\{i_1,i_2\}$ into the set $\mathcal{I}$ whenever $e_{k_i}^T\alpha_i=0$ for some $i\in\mathcal{I}$ and set $k_{i_1}=k_{i_2}=k_i-1$, $\tau_{i_1}=\tau_i-\epsilon$, $\tau_{i_2}=\tau_i+\epsilon$ for some suitably small ϵ. Furthermore, to ensure that $\tau_i\in[0,T_{\mathrm{con}}]$ for each $i\in\mathcal{I}$, we remove i from $\mathcal{I}$ whenever

$\tau_i = 0$ and $\dot{\tau}_i < 0$ or $\tau_i = T_{\text{con}}$ and $\dot{\tau}_i > 0$. Finally, new intersection points $\mathcal{L}_k \cap \mathcal{G}_k$ corresponding to $\tau = 0$ and $\tau = T_{\text{con}}$ are detected and introduced to the scheme by continuously monitoring the values of the left-hand side of (9) for $\tau = 0$, $\tau = T_{\text{con}}$, and $k = 1, \dots, n-1$.

ALGORITHM 4.6. *Offline: Determine the solutions* $\{(\tau_i)_{i\in\mathcal{I}}\}$ *to* (9) *for* $k = 1, \dots, n-1$ *satisfying* $\tau_i \in [0, T_{\text{con}}]$. *Online at times* $t \geq 0$*:*

1. *For each* $i \in \mathcal{I}$*:*
   (a) *Compute* α_i *using* (8) *and* $\dot{\tau}_i$ *using* (11).
   (b) *If* $gA^{k_i+1}\mathrm{e}^{A\tau_i}(z + W_{k_i}\alpha_i) = 0$*, remove* i *from* $\mathcal{I}$.
   (c) *If* $e_{k_i}^T\alpha_i = 0$*, add* $\{i_1, i_2\}$ *to* $\mathcal{I}$ *and define* $\tau_{i_1} = \tau_i + \epsilon$, $\tau_{i_2} = \tau_i - \epsilon$, $k_{i_1} = k_{i_2} = k_i - 1$.
   (d) *If* $\tau_i = 0$ *and* $\dot{\tau}_i < 0$*, remove* i *from* $\mathcal{I}$.
   (e) *If* $\tau_i = T_{\text{con}}$ *and* $\dot{\tau}_i > 0$*, remove* i *from* $\mathcal{I}$.
2. *For* $\tau = 0$*, and* $k = 1, \dots, n-1$*:*
   Compute α *using* (8). *If* $gA^k(z + W_k\alpha) = 0$ *and* $(\partial\tau/\partial z)\dot{z} > 0$*, add* i *to* $\mathcal{I}$ *and define* $\tau_i = 0$, $k_i = k$.
3. *For* $\tau = T_{\text{con}}$ *and* $k = 1, \dots, n-1$*:*
   Compute α *using* (8). *If* $gA^k\mathrm{e}^{AT_{\text{con}}}(z + W_k\alpha) = 0$ *and* $(\partial\tau/\partial z)\dot{z} < 0$*, add* i *to* $\mathcal{I}$ *and define* $\tau_i = T_{\text{con}}$, $k_i = k$.

Remark 4.7. From property (P4) and Remark 4.5, Algorithm 4.6 enables $z(t) \in \partial\Omega$ to be detected by checking the following simple conditions:

(i) $\alpha_i = 0$ for some $i \in \mathcal{I}$ such that $k_i = 1$.
(ii) $gz(t) = \overline{u}$.

Furthermore, the prediction times associated with active constraints are given by τ_i if (i) is satisfied or by 0 if (ii) is satisfied.

5. Prediction parameter adaptation. From the definition of Ω, the prediction parameters $c^*(t)$ that minimize $J(t)$ over the prediction class (4) are the solution of the following problem:

$$\begin{array}{ll} \underset{c}{\text{minimize}} & J(t) \\ \text{subject to} & \begin{bmatrix} c \\ x(t) \end{bmatrix} \in \Omega. \end{array} \tag{12}$$

Below we describe a method of incrementally optimizing performance by adapting the prediction parameters $c(t)$ online via $\dot{c}(t)$ and specifying the control law as

$$\dot{c}(t) = A_\phi c(t) + \theta(t), \tag{13a}$$

$$u(t) = -kx(t) + \phi(0)^T c(t), \tag{13b}$$

where the direction $\theta(t)$ of adaptation is chosen according to the gradient of $J(t)$ subject to the constraint that $z(t)$ remains in Ω. This approach ensures that the rate of decrease of $J(t)$ is greater than along the prediction system trajectory passing through $z(t)$ while maintaining the feasibility of the receding horizon optimization problem. As a result, (13) exponentially stabilizes the plant from any initial condition $x(0)$ for which a predicted control law of the form (4) is stabilizing (i.e., from any $x(0)$ for which there exists some $c(0)$ such that $z(0) \in \Omega$).

We also show below that this approach guarantees asymptotic convergence of $z(t)$ to a point $z^*(t) = \begin{bmatrix} c^{*T}(t) & x^T(t) \end{bmatrix}^T$, which is optimal with respect to (12) for the given value of $x(t)$, provided the direction $\theta(t)$ in which $c(t)$ is adapted satisfies an optimality criterion. This direction is determined simply by the gradient of $J(t)$

if $z(t)$ lies in the interior of Ω and by a projection of the cost gradient into the subspace tangent to the surface of Ω if $z(t)$ lies in the boundary of Ω. We describe an efficient method of computing a suitable projection based on a knowledge of the normal subspace to $\partial\Omega$ at $z(t)$. Noting that the normal subspace to $\partial\Omega$ at a point $z(t) \in \partial\Omega$ can be determined using Algorithm 4.6 and property (P3), the adaptation law (13a) can be implemented online and, in conjunction with (13b), results in a receding horizon control law with guaranteed stability and asymptotic convergence to a solution of (12).

The predicted performance index evaluated along trajectories of (6) is given by the quadratic form $J(t) = z^T(t)Pz(t)$, where P is defined by

$$PA + A^T P = -\overline{Q}, \;\; \overline{Q} = \begin{bmatrix} 0 & 0 \\ 0 & Q \end{bmatrix} + G^T R G.$$

In the following, we assume that $(\overline{Q}^{1/2}, A)$ is observable so that P is positive definite; given that $(A_m, B_m, Q^{1/2})$ is controllable and observable, this is ensured if $(\Phi(0), A_\phi)$ is observable. From (13) we have

$$\dot{z}(t) = Az(t) + \begin{bmatrix} \theta(t) \\ 0 \end{bmatrix} \tag{14}$$

along closed-loop trajectories, and the derivative of $J(t)$ can therefore be expressed as

$$\dot{J}(t) = -z^T(t)\overline{Q}z(t) + [\nabla_c J(t)]^T \theta(t), \tag{15}$$

where $\nabla_c J(t)$ denotes the gradient of J with respect to c evaluated at $z(t)$. The following theorem uses standard monotonicity arguments (see, e.g., [17]) to show that (13b) asymptotically stabilizes the plant (1) when combined with any adaptation law of the form (13a) with the property that $\nabla_c J^T \theta$ is nonpositive and which ensures that $z(t)$ remains in Ω at all times t.

THEOREM 5.1. *If $\theta(t)$ in* (13a) *satisfies $[\nabla_c J(t)]^T \theta(t) \le 0$ and $z(t) \in \Omega$ for all $t \ge 0$, then $x = 0$ is an asymptotically stable equilibrium of the closed-loop system of* (1) *under* (13b).

Proof. If $\theta(t)$ satisfies $[\nabla_c J(t)]^T \theta(t) \le 0$ and $z(t) \in \Omega$, then from (15), $\dot{J}(t) \le -z^T(t)\overline{Q}z(t)$ for all $t \ge 0$, and it follows that $z = 0$ is a stable equilibrium of the closed-loop dynamics (14). The integral of this bound over $t \ge 0$ gives

$$\int_0^\infty z^T(t')\overline{Q}z(t')\,dt' \le J(0) - \lim_{t\to\infty} J(t),$$

and since $\theta(t)$ is finite for all t by assumption, it follows from Barbalat's lemma [14] that $\lim_{t\to\infty} z^T(t)\overline{Q}z(t) = 0$, which implies that $\lim_{t\to\infty} z(t) = 0$ due to the observability of $(\overline{Q}^{1/2}, A)$. □

We define the unconstrained adaptation law by setting

$$\theta(t) = -\gamma_c \nabla_c J(t) \tag{16}$$

in (13a), where $\gamma_c > 0$ is an adaptation gain. It is easy to show that (16) forces $c(t)$ to converge exponentially to the unconstrained optimal prediction parameters $c = 0$, and provided that γ_c is sufficiently large, (16) will therefore drive $z(t)$ to the admissible

set boundary if the unconstrained optimal control $u(t) = -kx(t)$ is initially infeasible. In order to ensure that $z(t)$ remains in Ω whenever $z(t) \in \partial\Omega$ and that $z(t)$ converges to a constrained optimal point $z^*(t)$, a modified adaptation law for $c(t)$ based on a projection of $\nabla_c J(t)$ into $\partial\Omega$ is needed.

Before describing the constrained adaptation law, we first give conditions for an optimal point $z^*(t)$. In the following, we denote the set of (outward) normal vectors to $\partial\Omega$ at a point $z \in \partial\Omega$ as $\{(n_i)_{i\in\mathcal{I}_a}\}$ and define $n_{i,c} = [\mathrm{I}_{n_c} \quad 0]n_i$ for all $i \in \mathcal{I}_a$. Thus in the case of the single input constraint $u(t) \le \overline{u}$, for example, every $i \in \mathcal{I}_a$ corresponds to a prediction time $\tau_i \ge 0$ for which $g\mathrm{e}^{A\tau_i}z = \overline{u}$, and $n_i = (g\mathrm{e}^{A\tau_i})^T$ for all $i \in \mathcal{I}_a$.

LEMMA 5.2. *$z(t) = z^*(t)$ if and only if there exist multipliers λ_i, $i \in \mathcal{I}_a$, satisfying*

$$\begin{aligned} &\sum_{i\in\mathcal{I}_a} \lambda_i n_{i,c} = -\nabla_c J(t), \\ &\lambda_i \ge 0 \quad \forall i \in \mathcal{I}_a. \end{aligned} \tag{17}$$

Proof. These are the Kuhn–Tucker (KT) conditions for constrained optimality applied to problem (12). The necessity of (17) can be shown by considering the effect on J of a perturbation $c = c^* + \delta c$. If $\nabla_c J$ does not lie in the span of $\{(n_i)_{i\in\mathcal{I}_a}\}$ or if $\lambda_i < 0$ for some $i \in \mathcal{I}_a$, then there exists a direction s such that $n_{i,c}^T s \le 0$ for all $i \in \mathcal{I}_a$ and $\nabla_c J^T s < 0$. A reduction in J can therefore be achieved without violating constraints by choosing $\delta c = \epsilon s$ for sufficiently small $\epsilon > 0$, and, in this case, c^* cannot be a minimizer of (12). The sufficiency of (17) follows from the convexity of J and of Ω. □

From (15) it can be seen that an adaptation law for c should minimize $\nabla_c J^T\theta$ in order to maximize the rate of convergence of $J(t)$. At a point $z(t) \in \partial\Omega$, we require, in addition, that $n_i^T\dot{z}$ be negative for all $i \in \mathcal{I}_a$ in order that $z(t)$ remains in Ω. These criteria are conveniently expressed as a linear program: minimize $\nabla_c J^T\theta$ over θ subject to $n_{i,c}^T\theta \le -n_i^T Az$ for all $i \in \mathcal{I}_a$. However, as a consequence of Lemma 5.2, the solution of this problem is unbounded for any $z \ne z^*$. In order to define a meaningful optimization on which to base the design of the constrained adaptation law, we therefore include the additional constraint $\theta^T S\theta \le r^2$ for some symmetric positive definite S and $r > 0$ and consider the following convex subproblem:

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & \nabla_c J^T\theta \\ \text{subject to} \quad & n_{i,c}^T\theta \le -n_i^T Az \quad \forall i \in \mathcal{I}_a, \\ & \theta^T S\theta \le r^2. \end{aligned} \tag{18}$$

Introducing Lagrange multipliers λ_i, $i \in \mathcal{I}_a$, and λ^0 and defining the Lagrangian function

$$L(\theta, (\lambda_i)_{i\in\mathcal{I}_a}, \lambda^0) = \nabla_c J^T\theta + \sum_{i\in\mathcal{I}_a} \lambda_i(n_{i,c}^T\theta + n_i^T Az) + \frac{1}{2}\lambda^0(\theta^T S\theta - r^2),$$

the solution θ^* of (18) is characterized by the conditions

$$\nabla_\theta L = \nabla_c J + \sum_{i\in\mathcal{I}_a} \lambda_i n_{i,c} + \lambda^0 S\theta = 0, \tag{19a}$$

$$\lambda_i > 0, \quad n_{i,c}^T\theta^* = -n_i^T Az \quad \forall i \in \mathcal{A}^*, \tag{19b}$$

$$\lambda_i = 0, \quad n_{i,c}^T\theta^* \le -n_i^T Az \quad \forall i \in \mathcal{I}_a - \mathcal{A}^*, \tag{19c}$$

$$\lambda^0 \ge 0, \quad \lambda^0(\theta^{*T} S\theta^* - r^2) = 0, \tag{19d}$$

where $\mathcal{A}^* \subseteq \mathcal{I}_a$ is the set of active constraints at the solution. The following theorem shows that the solution of the subproblem (18) ensures asymptotic convergence to a point z^* satisfying the KT conditions of Lemma 5.2, subject to mild conditions on S and r.

THEOREM 5.3. *If S and $r > 0$ are continuous functions of z and $\theta(t) = \theta^*(t)$ in* (13a)*, then $z(t)$ converges asymptotically to $z^*(t)$.*

Proof. The solution θ^* of (18) is also the solution of the QP problem

$$\begin{array}{ll} \underset{\theta}{\text{minimize}} & \nabla_c J^T \theta + \frac{1}{2}\lambda^0 \theta^T S \theta \\ \text{subject to} & n_{i,c}^T \theta \le -n_i^T A z \quad \forall i \in \mathcal{I}_a, \end{array} \tag{20}$$

where λ^0 is defined by the optimality conditions (19). From the definition of the admissible set boundary, we have $n_i^T A z \le 0$ for all $i \in \mathcal{I}_a$, and it follows that $\theta = 0$ satisfies the constraints of (20), which implies that $\nabla_c J^T \theta^* + \frac{1}{2}\lambda^0 \theta^{*T} S \theta^* \le 0$. Noting that $\theta^{*T} S \theta^* = r^2$ whenever $z \ne z^*$ since the constraint $\theta^T S \theta \le r^2$ is then necessarily active, we have

$$\nabla_c J^T \theta \le -\frac{1}{2} r^2 \lambda^0.$$

The integral of (15) with $\theta(t) = \theta^*(t)$ therefore gives

$$\frac{1}{2}\int_0^t r^2 \lambda^0(t')\, dt' \le J(0) - J(t).$$

The right-hand side of this expression is finite, and since λ^0 is nonnegative and continuous (due to the continuous dependence of the constraints and objective of (18) on z), Barbalat's lemma shows that $\lambda^0 \to 0$ as $t \to \infty$. Furthermore, from (19a), we have

$$\sum_{i \in \mathcal{I}_a} \lambda_i n_{i,c} + \lambda^0 S \theta^* = -\nabla_c J, \quad \lambda_i \ge 0 \ \ \forall i \in \mathcal{I}_a,$$

for all $t \ge 0$, and $\lambda^0 \to 0$ therefore implies that the conditions of Lemma 5.2 are satisfied asymptotically. ∎

The solution of (18) is easily computed if $\{(n_i)_{i \in \mathcal{I}_a}\}$ consists of a single normal vector. In practice, this is often the case since the argument of Appendix B shows that the set of points in $\partial\Omega$ at which $\operatorname{span}\{(n_i)_{i \in \mathcal{I}_a}\}$ has dimension greater than 1 has zero measure on $\partial\Omega$. However, the computation of θ^* is nontrivial if $\{(n_i)_{i \in \mathcal{I}_a}\}$ contains several elements and may require a number of iterative steps. To overcome this difficulty, we make use of the freedom in choice of S to derive an efficiently computed suboptimal solution which nevertheless ensures convergence to z^*.

The approach is based on decoupling the Lagrange multipliers for the linear constraints in (18). Define $\lambda \in \mathbb{R}^{n_a}$ as the vector of multipliers λ_i $i \in \mathcal{I}_a$ and $N \in \mathbb{R}^{n \times n_a}$ as the matrix with columns n_i $i \in \mathcal{I}_a$, and let $N_c = [\mathrm{I}_{n_c} \quad 0]N$; then (19a) can be expressed as

$$N_c^T S^{-1} N_c \lambda = -N_c^T S^{-1} \nabla_c J - \lambda^0 N_c^T \theta.$$

Provided the normal vectors $\{(n_{i,c})_{i \in \mathcal{I}_a}\}$ are linearly independent, the factor $N_c^T S^{-1} N_c$ can be diagonalized by defining S^{-1} in terms of a generalized left inverse of N_c. This is conveniently done via a QR factorization of N_c:

$$N_c = Q \begin{bmatrix} R \\ 0 \end{bmatrix} = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R \\ 0 \end{bmatrix} = Q_1 R,$$

where $Q \in \mathbb{R}^{n_c \times n_c}$ is orthogonal, $R \in \mathbb{R}^{n_a \times n_a}$ is upper triangular, and $Q_1 \in \mathbb{R}^{n_c \times n_a}$, $Q_2 \in \mathbb{R}^{n_c \times n_c - n_a}$. Then the definition

$$S^{-1} = Q_1 R^{-T} R^{-1} Q_1^T + Q_2 Q_2^T \Leftrightarrow S = Q_1 R R^T Q_1^T + Q_2 Q_2^T$$

yields $\lambda = -R^{-1} Q_1^T \nabla_c J - \lambda^0 N^T \theta$, and the active set $\mathcal{A}^*$ (defined as the set of indices of constraints which are active at the solution θ^*) is given simply by

$$\mathcal{A}^* = \left\{ i;\ -e_i^T (R^{-1} Q_1^T \nabla_c J - \lambda^0 N^T A z) > 0 \right\}$$

(where e_i denotes the ith column of the identity I_{n_a}).

Computation of λ^0 requires a knowledge of λ at the solution, however, and to avoid this complication we choose a suboptimal active set $\mathcal{A} \supseteq \mathcal{A}^*$ as

$$\mathcal{A} = \{ i;\ -e_i^T R^{-1} Q_1^T \nabla_c J > 0 \}. \tag{21}$$

Setting $\lambda_i = 0$ for all $i \in \mathcal{I}_a - \mathcal{A}$, the corresponding solution for θ is given by

$$\begin{aligned} \theta &= -\frac{1}{\lambda^0} \Pi \nabla_c J - Q_1 R^{-T} E_{\mathcal{A}} E_{\mathcal{A}}^T N^T A z, \\ \lambda^0 &= \frac{\|\nabla_c J\|_\Pi}{[r^2 - \|E_{\mathcal{A}}^T N^T A z\|^2]^{1/2}}, \\ \Pi &= Q_1 R^{-T} (\mathrm{I}_{n_a} - E_{\mathcal{A}} E_{\mathcal{A}}^T) R^{-1} Q_1^T + Q_2 Q_2^T, \end{aligned}$$

where $E_{\mathcal{A}} = [(e_i)_{i \in \mathcal{A}}]$ and $\|\nabla_c J\|_\Pi^2 = \nabla_c J^T \Pi \nabla_c J$.

Remark 5.4. The requirement that the normal vectors $\{(n_{i,c})_{i \in \mathcal{I}_a}\}$ are linearly independent can be enforced by removing index i from $\mathcal{I}_a$ if the absolute value of the ith diagonal element of R falls below a threshold.

Defining $r^2 = \gamma_c^2 + \|E_{\mathcal{A}}^T N^T A z\|^2$ for some adaptation gain γ_c, the corresponding constrained adaptation law is given by (13a) with

$$\theta(t) = -\gamma_c \frac{\Pi \nabla_c J(t)}{\|\nabla_c J(t)\|_\Pi} - Q_1 R^{-T} E_{\mathcal{A}} E_{\mathcal{A}}^T N^T A z(t). \tag{22}$$

This suboptimal choice for $\theta(t)$ is shown below to satisfy the conditions for closed-loop stability of Theorem 5.1 and to ensure asymptotic convergence of $z(t)$ to $z^*(t)$.

LEMMA 5.5. *The constrained adaptation law defined by* (13a) *and* (22) *ensures that $z(t)$ remains in Ω and that $[\nabla_c J(t)]^T \theta(t) \le 0$ for all $t \ge 0$.*

Proof. From the definition of $\mathcal{A}$, we have $e_i^T R^{-1} Q_1^T \nabla_c J < 0$ if $i \in \mathcal{A}$ and $e_i^T R^{-1} Q_1^T \nabla_c J \ge 0$ if $i \in \mathcal{I}_a - \mathcal{A}$. Hence

$$n_{i,c}^T \theta = \begin{cases} -n_i^T A z & \text{if } i \in \mathcal{A}, \\ -\frac{1}{\lambda^0} e_i^T R^{-1} Q_1^T \nabla_c J \le 0 \le -n_i^T A z & \text{if } i \in \mathcal{I}_a - \mathcal{A}, \end{cases}$$

which implies that $n_i^T \dot{z} \le 0$ whenever $z \in \partial\Omega$. Also,

$$\nabla_c J^T \theta = -\gamma_c \|\nabla_c J\|_\Pi - \nabla_c J^T Q_1 R^{-T} E_{\mathcal{A}} E_{\mathcal{A}}^T N^T A z \le -\gamma_c \|\nabla_c J\|_\Pi \tag{23}$$

(where $n_i^T A z \le 0$ for all $i \in \mathcal{I}_a$ has been used). ∎

THEOREM 5.6. *Under the constrained adaptation law of* (13a) *with* (22), *$z(t)$ is asymptotically convergent to $z^*(t)$.*

Proof. The integral of (15) with the bound (23) gives

$$\gamma_c \int_0^t \|\nabla_c J(t')\|_\Pi \, dt' \leq J(0) - J(t),$$

and since $\nabla_c J$ and Π are continuous in z, it follows from Barbalat's lemma that $\Pi \nabla_c J \to 0$ as $t \to \infty$. This implies that $-\nabla_c J \to \sum_{i \in \mathcal{I}_a} \lambda_i n_{i,c}$ and $\lambda^0 \to 0$. Furthermore, $\lambda_i = 0$ if $i \in \mathcal{I}_a - \mathcal{A}$ and $\lambda_i - \lambda^0 N^T A z > 0$ if $i \in \mathcal{A}$, so conditions of Lemma 5.2 are satisfied asymptotically. □

Remark 5.7. As well as providing a feasible gradient descent direction, (22) has the property that $\nabla_c J^T \theta = 0$ only if $\theta = 0$. From (15), this implies that it is always possible to construct an exponentially convergent upper bound for $J(t)$. For example, if $\theta(t_0) \neq 0$, then there must exist $t_1 > t_0$ such that $J(t) \leq z^T(t_0) \mathrm{e}^{A^T(t-t_0)} P \mathrm{e}^{A(t-t_0)} z(t_0)$ for all $t \in [t_0, t_1]$. On the other hand, this bound clearly holds with equality if $\theta(t) = 0$ for $t \in [t_0, t_1]$. It follows that (13) ensures exponential convergence $z(t) \to 0$ along closed-loop trajectories.

The following algorithm summarizes the receding horizon control law, which by Theorem 5.1 and Remark 5.7 is exponentially stable and from Theorem 5.6 is asymptotically optimal with respect to (12).

ALGORITHM 5.8. *At $t = 0$: Perform the offline calculations of Algorithm* 4.6 *and find $c(0)$ such that $z(0) \in \Omega$. At times $t \geq 0$:*

1. *Update $\mathcal{I}$ via Algorithm* 4.6 *and determine $\mathcal{I}_a$ using Remark* 4.7.
2. *Update $c(t)$ using* (13a) *and* (16) *if $\mathcal{I}_a = \emptyset$ or* (22) *otherwise.*
3. *Implement $u(t) = Kx(t) + \Phi(0)^T c(t)$.*

Remark 5.9. The solution (22) for θ resembles part of a single iteration of a primal active set method for QP (see, e.g., [11]). This observation emphasizes the computational advantages of Algorithm 5.8; online computational load is reduced significantly by exploiting prior knowledge of the dependence of the active constraint set on the plant state. The approach proposed in [2] is similar in that the optimal active constraint sets corresponding to different regions of state-space are computed offline so that the online computation reduces to checking in which region the plant state lies. However, Algorithm 5.8 has the further advantage of an explicit characterization of the state-space in terms of active constraints, and this results in lower online computational burden than the approach of [2], which can require checking large numbers of linear inequalities online. This reduction in computational load is gained at the expense of suboptimality since Algorithm 5.8 can only ensure asymptotic convergence to a solution of the receding horizon optimization.

Remark 5.10. The closed-loop cost, $\mathfrak{J}(t)$, is defined as the objective of (3) evaluated along closed-loop trajectories under Algorithm 5.8 as follows:

$$\mathfrak{J}(t) = \int_t^\infty (x^T(t')Qx(t') + u^T(t')Ru(t')) \, dt' = \int_t^\infty z^T(t') \overline{Q} z(t') \, dt' \tag{24}$$

and can be expressed using (15) as

$$\mathfrak{J}(t) = J(t) + \int_t^\infty [\nabla_c J(t')]^T \theta(t') \, dt'.$$

Although Algorithm 5.8 is based on the MPC strategy of minimizing the predicted cost $J(t)$, the policy of minimizing $\nabla_c J^T \theta$ also turns out to be optimal with respect to

the constrained minimization of $\mathfrak{J}(t)$ under certain conditions which we now discuss. Substituting for $\nabla_c J^T\theta$ using (15), the minimization of (18) is equivalent to

$$\min_{\theta}\{\nabla J^T\dot{z} + z^T\overline{Q}z \quad \text{subject to} \quad (14) \text{ and } n_i^T\dot{z} \le 0,\ i \in \mathcal{I}_a\} = \nabla_c J^T\theta^*. \tag{25}$$

This is an HJB equation which shows that ∇J is the gradient of the optimal closed-loop cost for the dynamics of (14) subject to $z \in \Omega$ and that θ^* is the corresponding optimal value of θ [1] whenever $\nabla_c J^T\theta^* = 0$. This condition requires that $z = z^*$ and $n_i^T Az = 0$ for all $i \in \mathcal{A}^*$ (i.e., all active constraints correspond to predicted trajectories that are tangent at future points in time to their constraint boundaries).

6. Simulation examples. This section describes example simulations which compare the performance and computational load of Algorithm 5.8 with QP-based MPC strategies. Given the plant state-space realization (A_m, B_m, C_m), the predicted performance cost is defined by (3), with

$$Q = C_m^T C_m, \qquad R = 1.$$

The matrix A_ϕ is chosen in both examples as

$$A_\phi = \text{diag}\{-\sigma_1, \ldots, -\sigma_{n_c}\}, \quad \sigma_i > 0,$$

with $\sigma_1, \ldots, \sigma_{n_c}$ logarithmically spaced on the interval $[\omega_{3\text{dB}}/\sqrt{n_c}, \sqrt{n_c}\omega_{3\text{dB}}]$, where $\omega_{3\text{dB}}$ is the 3dB bandwidth of the closed-loop system under $u = Kx$; and $\Phi(0) = [\ 1 \quad \ldots \quad 1\ ]$ is employed. The initial point $z(0) \in \Omega$ is determined in each simulation by choosing $c(0)$ as the center of an ellipsoidal bound (derived from a small number of discrete samples of the continuous-time constraints) on the set $\{c;\ [\ c^T \quad x^T(0)\ ]^T \in \Omega\}$.

Example 1. Consider the third order plant

$$A_m = \begin{bmatrix} -0.50356 & 2.4417 & -1.8849 \\ 8 & 0 & 0 \\ 0 & 2 & 0 \end{bmatrix}, \quad B_m = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}, \quad C_m = \begin{bmatrix} -0.79081 \\ 0.10112 \\ 0.94243 \end{bmatrix}^T$$

subject to input constraints

$$-1 \le u \le 1$$

for which the unconstrained LQ-optimal feedback gain is $k = [\ 5.0662 \quad 3.4678 \quad 0.3297\ ]$. The unconstrained closed-loop system under $u = -kx$ has poles at $\{-5.3073, -2.7261 \pm 0.7773i\}$, and we choose a prediction class of dimension $n_c = 3$ and set $A_\phi = \text{diag}\{-1.6207, -2.8072, -4.8622\}$.

The performance of Algorithm 5.8 is compared with a QP-based MPC law described in [5] in Figures 2 and 3 and in Table 1. The QP-based controller uses the same parameterization of predicted inputs as Algorithm 5.8 (i.e., the exponential basis functions of (4)–(5), with $n_c = 3$ and A_ϕ as defined above) but solves the corresponding receding horizon optimization (12) periodically with period $T = 0.1$. In order to formulate (12) as a QP problem, this algorithm approximates the continuous-time constraints (2) by a set of discrete-time constraints which are artificially tightened to ensure the satisfaction of (2) at all prediction times.

The input and output responses of Figure 2 show that Algorithm 5.8 with $\gamma_c = 1$ gives poorer performance than the QP-based algorithm. This is due to the use of a

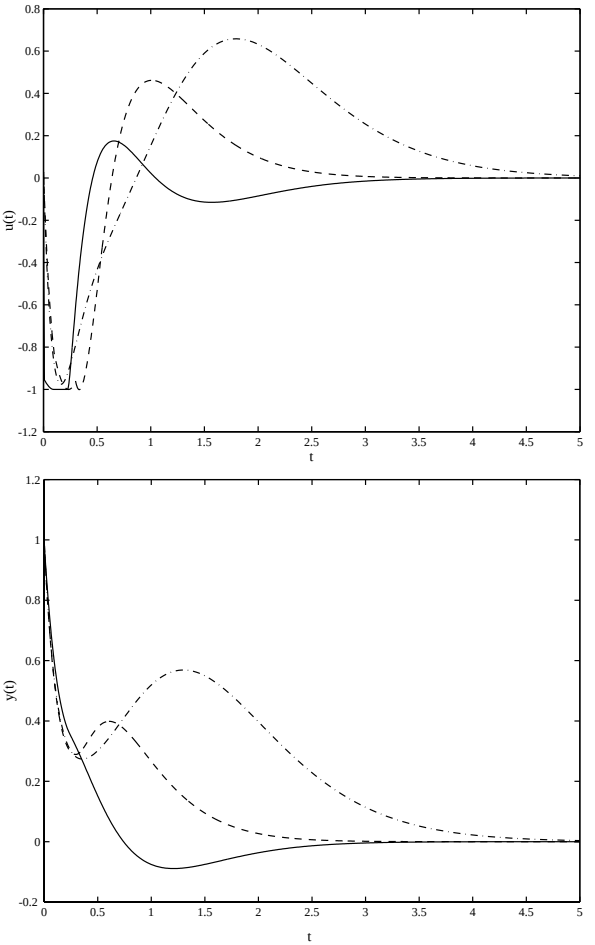


FIG. 2. *Input and output responses for Example* 1. *Solid line:* $\gamma_c = 100$. *Dashed-dotted:* $\gamma_c = 1$. *Dashed: Exponential basis functions and optimization via QP.*

very suboptimal value for $z(0)$, as can be seen from the difference in initial prediction costs $J(0)$ (Figure 3(a)). In fact, $z(t)$ remains in the interior of Ω for all $t \geq 0$ if $\gamma_c = 1$ in this example since $\|\nabla_c J\|_\Pi = \|\nabla_c J\|$ at all times (Figure 3(c)). However, the use of $\gamma_c = 100$ results in faster convergence of $J(t)$ than is achieved by the QP-based algorithm (Figure 3(a)) and, moreover, forces $z(t)$ to converge to a point in the boundary of Ω satisfying the KT conditions for constrained optimality for $0.15 \lesssim t \lesssim 0.3$ (Figure 3(c)), at which point the unconstrained optimal control becomes feasible.

Closed-loop performance is measured in Table 1 by the value of the closed-loop cost

$$\mathfrak{J} = \int_0^\infty [x^T(t)Qx(t) + u^T(t)Ru(t)]\, dt.$$

The computation times refer to average CPU times required by the parameter update law in Algorithm 5.8 and the QP solver (initialized using "warm starts") in the QP-based algorithm, with both controllers implemented in Matlab on a 440MHz Sun Ultra workstation. The computational load of Algorithm 5.8 per iteration is clearly

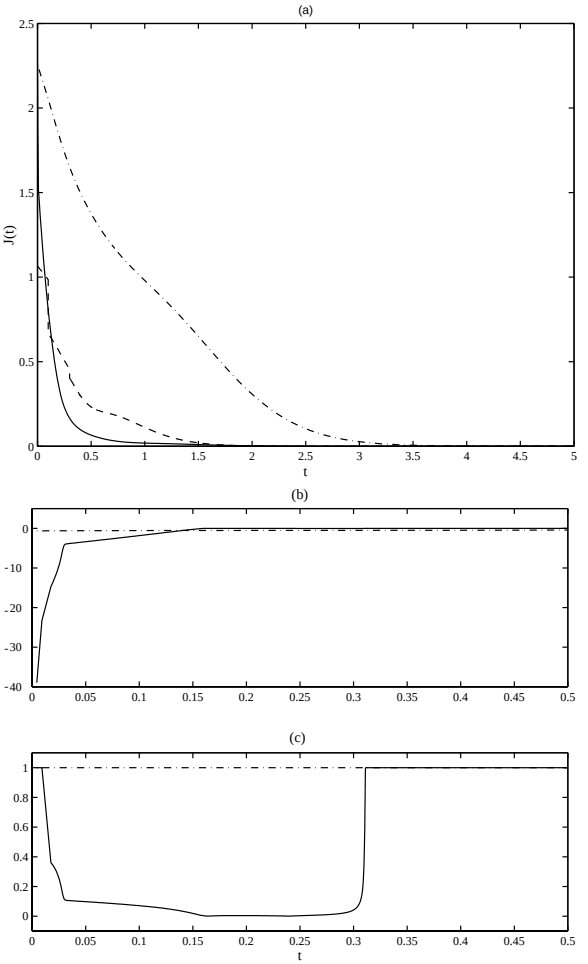


FIG. 3. *Prediction costs* (a)*, and evolution of* $\nabla_c J^T \theta$ (b) *and* $\|\nabla_c J\|_\Pi / \|\nabla_c J\|$ (c) *for Example* 1*. Solid line:* $\gamma_c = 100$*. Dashed-dotted:* $\gamma_c = 1$*. Dashed: Exponential basis functions and optimization via QP.*

TABLE 1
Closed-loop costs and computation times for Example 1*.*

		Cost $\mathfrak{J}$	CPU time (ms)
Alg. 5.8	$\gamma_c = 1$	1.3558	0.92
	$\gamma_c = 10$	0.4496	0.76
	$\gamma_c = 100$	0.4169	1.14
QP – exponential BF		0.6959	240

much lower than that of the QP-based algorithm.

Example 2. Consider the fifth order plant

$$A_m = \begin{bmatrix} -8.06 & 73.10 & 9.82 & 9.48 & 29.59 \\ -122.6 & 40.50 & -41.54 & -40.11 & -125.2 \\ -2.03 & 3.22 & -12.16 & 4.76 & -7.06 \\ 1.46 & -2.32 & -5.34 & -8.18 & 5.08 \\ 4.06 & -6.44 & 4.69 & 4.53 & 5.69 \end{bmatrix}, \quad B_m = \begin{bmatrix} 0.34 \\ -1.43 \\ -0.08 \\ 0.06 \\ 0.16 \end{bmatrix}, \quad C_m = \begin{bmatrix} 0.0028 \\ -0.0008 \\ -113.30 \\ 99.09 \\ -92.92 \end{bmatrix}^T$$

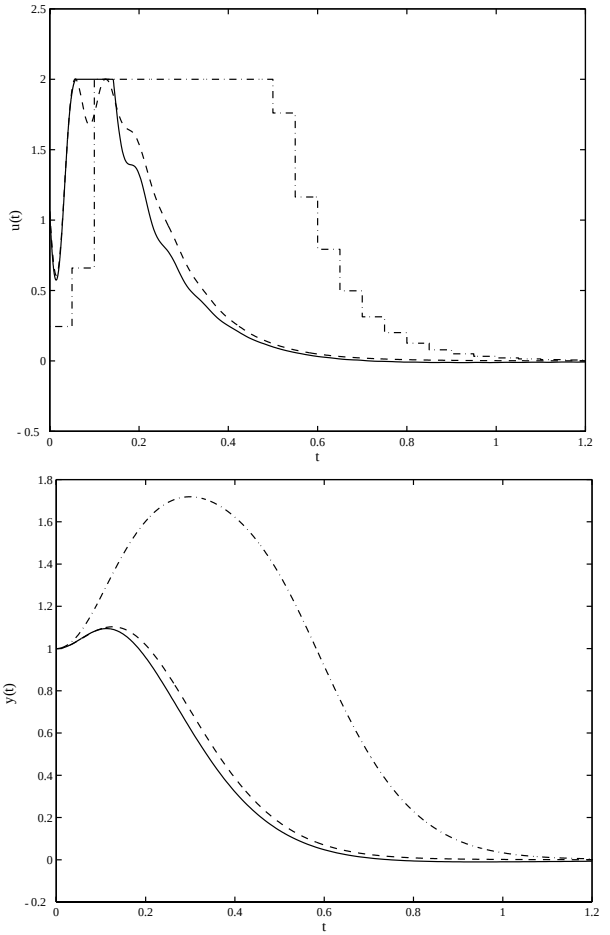


FIG. 4. *Inputs and outputs for Example* 2. *Solid line: Algorithm* 5.8 *with* $\gamma_c = 100$. *Dashed: Exponential basis functions and QP optimization. Dashed-dotted: Discrete-time inputs and QP optimization.*

subject to input constraints

$$-2 \leq u \leq 2.$$

The unconstrained LQ gain is $k = [\ 25.19 \quad -39.93 \quad 29.06 \quad 28.07 \quad 87.58\ ]$, and the corresponding closed-loop poles are located at $\{-8.43, -9.81 \pm 7.03i, -16.57 \pm 86.58i\}$. For a prediction class of dimension $n_c = 4$, we choose $A_\phi = \text{diag}\{-1.89, -3.78, -5.66, -7.55\}$.

Figures 4 and 5 and Table 2 compare the performance of Algorithm 5.8 with the QP-based receding horizon control law described in Example 1 and also with a discrete-time MPC law employing piecewise-constant predicted input trajectories centered on the unconstrained discrete-time LQ-optimal controller. Both QP-based algorithms use "warm starts" determined from the solution to the previous optimization problem. This control problem is considerably more challenging than Example 1

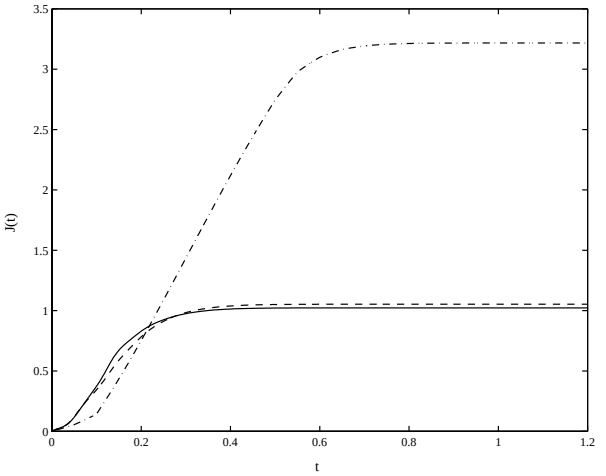


FIG. 5. *The evolution of closed-loop costs* $\mathfrak{J}(t)$ *for Example* 2. *Solid line: Algorithm* 5.8 *with* $\gamma_c = 100$. *Dashed: Exponential basis functions and QP optimization. Dashed-dotted: Discrete-time inputs and QP optimization.*

TABLE 2
Closed-loop costs and computation times for Example 2.

	Cost $\mathfrak{J}$	CPU time (ms)
Alg. 5.8, $n_c = 4$	1.0270	1.37
QP – exponential BF, $n_c = 4$	1.0532	380
QP – piecewise constant BF, $n_c = 10$	3.2175	232

since the plant has an almost uncontrollable unstable mode. For a sample period of $T = 0.05$, the discrete-time algorithm needs at least 10 degrees of freedom to stabilize the plant, and its closed-loop performance is considerably worse than the two other algorithms, which are centered on a continuous-time LQ-optimal control law. The long prediction horizon required in this example makes the QP-based algorithm employing exponential basis functions computationally expensive. On the other hand, Algorithm 5.8 combines good performance with low computational burden.

Appendix A. Conditions on predictions for feasibility given past feasibility. This section determines conditions on the structure of the class of input predictions given by (4) in order that the optimization problem (3) remains feasible at all times $t > 0$ given feasibility at $t = 0$. The receding horizon control law $u(t) = \hat{u}_t(0)$ is implemented for all $t \geq 0$, where $\hat{u}_t$ has the form (4). If the input to the plant on the interval $[t-\delta, t)$ for some $\delta > 0$ is $u(t+\tau) = \hat{u}_{t-\delta}(\tau+\delta)$, $\tau \in [-\delta, 0)$, then feasibility at t is guaranteed if the prediction defined by $\hat{u}_t(\tau) = \hat{u}_{t-\delta}(\tau+\delta)$, $\tau \geq 0$, is realizable by the prediction class of (4) since $\hat{u}_{t-\delta} \in \mathcal{U}$ by assumption. The following theorem determines an equivalent condition on Φ, the matrix of basis functions in the linear expansion of (4).

THEOREM A.1. *If* $u(t+\tau) = \hat{u}_{t-\delta}(\tau+\delta)$ *is implemented for all* $\tau \in [-\delta, 0)$ *for some* $\delta > 0$, *then at time* t *the prediction class* (4) *contains the signal* $\{\hat{u}_{t-\delta}(\tau+\delta),\ \tau \geq 0\}$ *if and only if* $M : \mathbb{R} \to \mathbb{R}^{n_c \times n_c}$ *exists satisfying*

$$\Phi(\tau+\delta) = \Phi(\tau)M(\delta) \quad \forall \tau \geq 0. \tag{26}$$

Proof. In the absence of disturbances, we have $x(t) = \hat{x}_{t-\delta}(\delta)$, so $\hat{u}_{t-\delta}$ can be expressed using (4) as

$$\hat{u}_{t-\delta}(\tau+\delta) = K\hat{x}_t(\tau) + \Phi(\tau+\delta)c(t-\delta) \quad \forall \tau \geq 0.$$

The prediction class of (4), therefore, contains $\{\hat{u}_{t-\delta}(\tau+\delta),\ \tau \geq 0\}$ if and only if $c(t)$ exists satisfying

$$\Phi(\tau)c(t) = \Phi(\tau+\delta)c(t-\delta) \quad \forall \tau \geq 0. \tag{27}$$

If Φ satisfies (26), then (27) holds for any $c(t-\delta) \in \mathbb{R}^{n_c}$ with the choice $c(t) = M(\delta)c(t-\delta)$. To show the necessity of (26), suppose that Φ satisfies

$$\Phi(\tau+\delta) = \Phi(\tau)M(\delta) + \Psi \quad \forall \tau \geq 0, \tag{28}$$

for some $\Psi \in \mathbb{R}^{n_u \times n_c}$ and that (27) holds for some $c(t) \in \mathbb{R}^{n_c}$. Assuming, without loss of generality, that Φ is full rank and solving (27) for $c(t)$ for the case when $n_u < n_c$, we have

$$c(t) = M(\delta)c(t-\delta) + \Phi^T(\tau)\big(\Phi(\tau)\Phi^T(\tau)\big)^{-1}\Psi c(t-\delta) + \Phi_\perp(\tau)\gamma,$$

where $\Phi_\perp(\tau)\gamma$ is a matrix representation of the kernel of $\Phi(\tau)$. On the other hand, if $n_u < n_c$, then $\tilde{\Phi}_\perp^T(\tau)\Psi c(t-\delta) = 0$ must hold in order that (27) has a solution for $c(t)$ (where the columns of $\tilde{\Phi}_\perp(\tau)$ form a basis for the kernel of $\Phi^T(\tau)$), and the solution is then given by

$$c(t) = M(\delta)c(t-\delta) + \big(\Phi^T(\tau)\Phi(\tau)\big)^{-1}\Phi^T(\tau)\Psi c(t-\delta).$$

However, $c(t)$ is independent of τ, and it follows that either $\gamma = 0$ and $\Psi c(t-\delta) = 0$ or $\Phi(\tau)$ is constant. For nonconstant Φ and arbitrary $c(t-\delta)$, (27) therefore holds for some $c(t)$ only if $\Psi = 0$ in (28). ∎

Applying (26) in the limit as $\delta \to 0$, with the additional condition $\Phi \in C^1$ (which is necessary due to the rate constraints of (2)), we have $M(0) = \mathrm{I}$ and $\dot{\Phi}(\tau) = \Phi(\tau)\dot{M}(0)$ for all $\tau \geq 0$. It follows that

$$\Phi(\tau) = \Phi(0)\mathrm{e}^{A_\phi \tau} \quad \forall \tau \geq 0,$$

where $A_\phi = \dot{M}(0)$.

Appendix B. Dimension of $\mathcal{G}_k$. This section determines the dimension and smoothness of the hypersurface defined by

$$\mathcal{G}_k = \{z \in \mathbb{R}^n;\ g\mathrm{e}^{A\tau}z = \overline{u},\ gA^i\mathrm{e}^{A\tau}z = 0,\ i = 1, \dots, k,\ \tau \geq 0\}$$

for $k = 1, \dots, n-1$.

THEOREM B.1. *At every point $z \in \mathcal{G}_k - \mathcal{G}_{k+1}$, $\mathcal{G}_k$ is locally a smooth hypersurface in $\mathbb{R}^n$ of dimension $n-k$ for $k = 1, \dots, n-2$.*

Proof. For any $z \in \mathcal{G}_k$, there exists $\tau \geq 0$ such that $f_i(z,\tau) = 0$ for $i = 0, \dots, k$, where

$$f_i(z,\tau) = \begin{cases} g\mathrm{e}^{A\tau}z - \overline{u}, & i = 0, \\ gA^i\mathrm{e}^{A\tau}z, & i = 1, \dots, k. \end{cases}$$

Let F denote the Jacobian matrix

$$\begin{bmatrix} \partial f_0/\partial z & \partial f_0/\partial \tau \\ \vdots & \vdots \\ \partial f_k/\partial z & \partial f_k/\partial \tau \end{bmatrix};$$

then, for all $(z,\tau) \in \mathcal{G}_k \times \mathbb{R}$ such that $f_i(z,\tau) = 0$, $i = 0, \dots, k$, F is given by

$$F(z,\tau) = \begin{bmatrix} g\mathrm{e}^{A\tau} & 0 \\ \vdots & \vdots \\ gA^{k-1}\mathrm{e}^{A\tau} & 0 \\ gA^k\mathrm{e}^{A\tau} & gA^{k+1}\mathrm{e}^{A\tau}z \end{bmatrix}.$$

The observability of (g, A), therefore, implies that $\operatorname{rank}(F(z,\tau)) = k+1$ for all $(z,\tau) \in \mathcal{G}_k \times \mathbb{R}$ such that $f_i(z,\tau) = 0$, $i = 0, \dots, k$, and $gA^{k+1}\mathrm{e}^{A\tau}z \neq 0$. Since f_i, $i = 0, \dots, k$, are C^∞ functions (i.e., infinitely continuously differentiable with respect to z and τ), it follows from the implicit function theorem that $\mathcal{G}_k - \mathcal{G}_{k+1}$ is a smooth hypersurface in $\mathbb{R}^n$ of dimension $n-k$. Specifically, let $z^T = [z_1^T \ z_2^T]$, where $z_2 \in \mathbb{R}^k$. Then it is clear that the Jacobian matrix

$$\begin{bmatrix} \partial f_0/\partial z_2 & \partial f_0/\partial \tau \\ \vdots & \vdots \\ \partial f_k/\partial z_2 & \partial f_k/\partial \tau \end{bmatrix}$$

is nonsingular for all $(z,\tau) \in \mathcal{G}_k \times \mathbb{R}$ such that $f_i(z,\tau) = 0$, $i = 0, \dots, k$, and $gA^{k+1}\mathrm{e}^{A\tau}z \neq 0$; and z_2, τ, therefore, are C^∞ functions of z_1 by the implicit function theorem. This implies that $\mathcal{G}_k$ is a smooth hypersurface parameterized by the $n-k$ coordinates of z_1 in a neighborhood of every point $z \in \mathcal{G}_k - \mathcal{G}_{k+1}$. ∎

COROLLARY B.2. *$\mathcal{G}_{n-1}$ is a smooth 1-dimensional hypersurface in $\mathbb{R}^n$.*

Proof. For $k = n-1$, the argument of Theorem B.1 shows that $\mathcal{G}_{n-1}$ is smooth and one-dimensional for all $(z,\tau) \in \mathcal{G}_{n-1} \times \mathbb{R}$ such that $f_i(z,\tau) = 0$, $i = 0, \dots, n-1$, and $gA^n\mathrm{e}^{A\tau}z \neq 0$. However, for any $z \in \mathbb{R}^n$, there is no value of $\tau \in \mathbb{R}$ such that $f_i(z,\tau) = 0$, $i = 0, \dots, n-1$, and $gA^n\mathrm{e}^{A\tau}z = 0$, due to the observability of (g, A). It follows that $\mathcal{G}_{n-1}$ is everywhere smooth and of dimension one. ∎

REFERENCES

[1] M. ATHANS AND P. FALB, *Optimal Control: An Introduction to the Theory and Its Applications*, McGraw–Hill, New York, 1966.

[2] A. BEMPORAD, M. MORARI, V. DUA, AND E. PISTIKOPOULOS, *The explicit solution of model predictive control via multiparametric quadratic programming*, in Proceedings of the American Control Conference, Chicago, IL, 2000, pp. 872–876.

[3] M. CANNON AND B. KOUVARITAKIS, *Fast suboptimal predictive control with guaranteed stability*, Systems Control Lett., 35 (1998), pp. 19–29.

[4] M. CANNON AND B. KOUVARITAKIS, *Continuous-time predictive control of constrained nonlinear systems*, in Nonlinear Model Predictive Control: Assessment and Future Directions for Research, Birkhäuser-Verlag, Basel, 2000.

[5] M. CANNON AND B. KOUVARITAKIS, *Infinite horizon predictive control of constrained continuous-time linear systems*, Automatica J. IFAC, 36 (2000), pp. 943–955.

[6] C. CHEN AND L. SHAW, *On receding horizon feedback control*, Automatica J. IFAC, 18 (1982), pp. 349–352.

[7] P. DAVE, F. DOYLE, AND J. PEKNY, *Customization strategies for the solution of linear programming problems arising from large scale model predictive control of a paper machine*, J. Process Control, 9 (1999), pp. 385–396.

[8] P. Dave, D. Willig, G. Kudva, J. Pekny, and F. Doyle, *LP methods in MPC of large-scale systems: Application to paper-machine CD control*, AIChE J., 43 (1997), pp. 1016–1031.
[9] R. Fletcher, *Practical Methods of Optimization*, Wiley-Interscience, New York, 1987.
[10] E. Gilbert and K. Tan, *Linear systems with state and control constraints: The theory and practice of maximal admissible sets*, IEEE Trans. Automat. Control, 36 (1991), pp. 1008–1020.
[11] P. Gill and W. Murray, *Numerically stable methods for quadratic programming*, Math. Programming, 14 (1978), pp. 349–372.
[12] S. Keerthi and E. Gilbert, *Optimal infinite-horizon feedback laws for a general class of constrained discrete-time systems. Stability and moving-horizon approximations*, J. Optim. Theory Appl., 57 (1988), pp. 265–293.
[13] B. Kouvaritakis, M. Cannon, and J. Rossiter, *Removing the need for QP in constrained predictive control*, in Proceedings of the IFAC International Symposium on Advanced Control of Chemical Processes, Pisa, Italy, 2000, pp. 311–316; to appear as *Who needs QP for linear MPC anyway?*, Automatica J. IFAC.
[14] M. Krstić, I. Kannellakopoulos, and P. Kokotović, *Nonlinear and Adaptive Control Design*, Wiley-Interscience, New York, 1995.
[15] W. Kwon and A. Pearson, *A modified quadratic cost problem and feedback stabilization of a linear system*, IEEE Trans. Automat. Control, 22 (1977), pp. 838–842.
[16] Y. Lee and B. Kouvaritakis, *Robust receding horizon predictive control for systems with uncertain dynamics and input saturation*, Automatica J. IFAC, 36 (2000), pp. 1497–1504.
[17] D. Mayne, J. B. Rawlings, C. Rao, and P. Scokaert, *Constrained model predictive control: Stability and optimality*, Automatica J. IFAC, 36 (2000), pp. 789–814.
[18] H. Michalska and D. Mayne, *Receding horizon control of nonlinear systems*, IEEE Trans. Automat. Control, 38 (1993), pp. 1623–1632.
[19] S. Qin and T. Badgwell, *Chemical Process Control* V*: Assessment and New Directions for Research*, AIChE Symposium Series 93, American Institute of Chemical Engineers, 1997, pp. 232–256.
[20] C. Rao, J. Campbell, J. Rawlings, and S. Wright, *Efficient implementation of model predictive control for sheet and film forming processes*, in Proceedings of the American Control Conference, Albuquerque, NM, 1997, pp. 2940–2944.
[21] P. Scokaert, D. Mayne, and J. Rawlings, *Suboptimal model predictive control (feasibility implies stability)*, IEEE Trans. Automat. Control, 44 (1999), pp. 648–654.
[22] J. VanAntwerp and R. Braatz, *Fast model predictive control of sheet and film processes*, IEEE Trans. Control Systems Technology, 8 (2000), pp. 408–417.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 1, pp. 83–88

A STOCHASTIC JURDJEVIC–QUINN THEOREM*

PATRICK FLORCHINGER†

Abstract. The purpose of this paper is to state sufficient conditions for the stabilizability of stochastic differential systems when both the drift and diffusion terms are affine in the control. This result extends to stochastic differential systems the well-known theorem of Jurdjevic–Quinn [*J. Differential Equations*, 28 (1978), pp. 381–389] and incorporates earlier stabilization results proved in [P. Florchinger, *Stochastic Anal. Appl.*, 12 (1994), pp. 473–480] and [R. Chabour and M. Oumoun, *Stochastic Anal. Appl.*, 16 (1998), pp. 43–50].

Key words. stochastic differential system, asymptotic stability in probability, Lyapunov theorem, La Salle's invariance principle

AMS subject classifications. 60H10, 93C10, 93D05, 93D15, 93E15

PII. S0363012900370788

Introduction. The stabilization of deterministic nonlinear control systems has been widely studied in past years by many authors. Among all of the results proved in this area of research, we wish to outline that of Jurdjevic–Quinn [3], giving stabilizing state feedback laws for deterministic systems affine in the control provided the control Lie algebra of the system has full rank. This result has been the starting point for different works on this topic since it appears that many engineering systems are of "Jurdjevic–Quinn type" (see [6], [8], or [7], for example). Note that the result exposed in [3] has been revisited in [7], where a more easily computable rank condition for stabilizability is stated.

A stochastic version of Jurdjevic–Quinn's theorem has been established by Florchinger [2] for stochastic differential systems, the drift of which is affine in the control. In fact, it is proved in [2] that the stabilizer given in [3] for the deterministic part of the system remains valid in the stochastic context provided the system coefficients satisfy a rank condition, which can easily be deduced from that stated in [7]. An extension of this result has been obtained by Chabour and Oumoun in [1], where it is proved that under the stabilizability condition provided in [2] one can compute stabilizing state feedback laws for stochastic differential systems in the form

$$x_t = x_0 + \int_0^t (b(x_s) + uf(x_s))\, ds + \int_0^t \sigma(x_s) dw_s + \int_0^t ug(x_s) d\tilde{w}_s,$$

where $(w_t)_{t\geq 0}$ and $(\tilde{w}_t)_{t\geq 0}$ are two independent Wiener processes.

The technique used in the last two cited papers is based on the stochastic Lyapunov analysis developed by Khasminskii [4] and the stochastic version of La Salle's invariance principle proved by Kushner [5]. However, in the proofs of the main results in [2] and [1], all of the information given by applying Itô's formula, when using the stochastic La Salle theorem, has not been used. The aim of this paper is to take this fact into account to improve the stabilizability conditions stated in [2] and [1] in order to be able to design stabilizers for a wider class of stochastic differential systems than that considered in [2] and [1].

*Received by the editors April 6, 2000; accepted for publication (in revised form) November 1, 2001; published electronically March 27, 2002.
http://www.siam.org/journals/sicon/41-1/37078.html
†23 Allée des Oeillets, F 57160 Moulins les Metz, France (patrick.florchinger@wanadoo.fr).

This paper is divided into four sections and is organized as follows. In section one, we recall some basic facts on the asymptotic stability in probability for the equilibrium solution of a stochastic differential equation. In section two, we introduce the class of control stochastic differential systems we are dealing with in this paper. In section three, we prove the main result of the paper. In section four, we apply the result proved in section three to a working example which cannot be stabilized by using the results proved in [2] and [1].

1. Asymptotic stability in probability. The purpose of this section is to recall some basic facts concerning the Lyapunov analysis for the asymptotic stability in probability of the equilibrium solution of a stochastic differential system that we need in what follows.

For a complete exposition on the subject, we refer to the book of Khasminskii [4].

Let $(\Omega, \mathcal{F}, P)$ be a complete probability space, and denote by $(w_t)_{t\geq 0}$ a standard $\mathbb{R}^m$-valued Wiener process defined on this space.

Consider the stochastic process solution $x_t \in \mathbb{R}^n$ of the stochastic differential equation written in the sense of Itô,

$$x_t = x_0 + \int_0^t f(x_s)ds + \sum_{i=1}^m \int_0^t g_i(x_s)dw_s^i, \tag{1}$$

where the following hold:

1. x_0 is given in $\mathbb{R}^n$.
2. f and g_i, $1 \leq i \leq m$, are functions mapping $\mathbb{R}^n$ into $\mathbb{R}^n$, vanishing in the origin, and such that there exists a nonnegative constant K such that
$$|f(x)|^2 + \sum_{i=1}^m |g_i(x)|^2 \leq K(1+|x|^2) \quad \forall x \in \mathbb{R}^n,$$
$$|f(x)-f(y)| + \sum_{i=1}^m |g_i(x)-g_i(y)| \leq K|x-y| \quad \forall x, y \in \mathbb{R}^n.$$

For any $s \geq 0$ and $x \in \mathbb{R}^n$, denote by $x_t^{s,x}$, $s \leq t$, the solution at time t of the stochastic differential equation (1) starting from the state x at time s.

Then one can introduce the notion of asymptotic stability in probability for the equilibrium solution of the stochastic differential equation (1) as follows.

DEFINITION 1.1. (1) *The equilibrium solution $x_t \equiv 0$ of the stochastic differential equation* (1) *is stable in probability if, for any $s \geq 0$ and $\epsilon > 0$,*

$$\lim_{x\to 0} P\left(\sup_{s\leq t} |x_t^{s,x}| > \epsilon\right) = 0.$$

(2) *The equilibrium solution $x_t \equiv 0$ of the stochastic differential equation* (1) *is asymptotically stable in probability if it is stable in probability and, for any $s \geq 0$ and $x \in \mathbb{R}^n$,*

$$P\left(\lim_{t\to+\infty} |x_t^{s,x}| = 0\right) = 1.$$

Denoting by L the infinitesimal generator of the stochastic process solution of the stochastic differential equation (1), one can prove the following criterion, which gives sufficient conditions in terms of the Lyapunov function for the asymptotic stability in probability of the equilibrium solution of the stochastic differential equation (1).

THEOREM 1.2 (see Khasminskii [4]). *Assume that there exists a Lyapunov function V defined on $\mathbb{R}^n$ (i.e., a proper C^2 function V mapping $\mathbb{R}^n$ into $\mathbb{R}$ such that $V(0) = 0$ and $V(x) > 0$ for any $x \in \mathbb{R}^n$, $x \neq 0$) such that*

$$LV(x) \leq 0 \quad (\textit{respectively},\ LV(x) < 0)$$

for any $x \in \mathbb{R}^n$, $x \neq 0$. Then the equilibrium solution $x_t \equiv 0$ of the stochastic differential equation (1) *is stable (respectively, asymptotically stable) in probability.*

To conclude this section, recall the stochastic version of La Salle's invariance principle, which gives the ω-limit set of a stochastic process stable in probability.

THEOREM 1.3 (see Kushner [5]). *Assume that there exists a Lyapunov function V defined on $\mathbb{R}^n$ such that*

$$LV(x) \leq 0$$

for any $x \in \mathbb{R}^n$. Then the stochastic process solution x_t of the stochastic differential equation (1) *tends to the largest invariant set whose support is contained in the locus $LV(x_t) = 0$ for any $t \geq 0$ with probability* 1.

2. Problem setting. The purpose of this section is to introduce the class of control stochastic differential systems we are dealing with in the rest of the paper.

Consider the stochastic process solution $x_t \in \mathbb{R}^n$ of the multi-input stochastic differential system written in the sense of Itô,

$$\begin{aligned} x_t = x_0 &+ \int_0^t \left(f_0(x_s) + \sum_{k=1}^{p} u^k f_k(x_s) \right) ds \\ &+ \sum_{i=1}^{m} \int_0^t \left(g_{i,0}(x_s) + \sum_{k=1}^{p} u^k g_{i,k}(x_s) \right) dw_s^i, \end{aligned} \tag{2}$$

where the following hold:

1. x_0 is given in $\mathbb{R}^n$.
2. u^k, $1 \leq k \leq p$, are real-valued measurable control laws.
3. f_k, $0 \leq k \leq p$, and $g_{i,k}$, $1 \leq i \leq m$, $0 \leq k \leq p$, are smooth Lipschitz functions mapping $\mathbb{R}^n$ into $\mathbb{R}^n$, vanishing in the origin, and such that there exists a nonnegative constant K such that

$$\sum_{k=0}^{p} \left(|f_k(x)|^2 + \sum_{i=1}^{m} |g_{i,k}(x)|^2 \right) \leq K(1 + |x|^2) \qquad \forall x \in \mathbb{R}^n.$$

The aim of this paper is to design a state feedback law u such that the equilibrium solution of the closed-loop system deduced from the stochastic differential system (2) is asymptotically stable in probability. Note that this problem has already been solved in [2] when the diffusion term in (2) does not depend on the control, i.e., when $g_{i,k} \equiv 0$ for every $i \in \{1, \ldots, m\}$ and $k \in \{1, \ldots, p\}$.

3. The main result. Before stating the main result of the paper, which extends to the stochastic differential system (2) the well-known theorem of Jurdjevic–Quinn established in [3], we introduce the following notation.

Denote by L_0 the infinitesimal generator of the stochastic process solution of the unforced stochastic differential system deduced from (2); that is, L_0 is the second

order differential operator defined for any function ϕ in $C^2(\mathbb{R}^n;\mathbb{R})$ by

$$L_0\phi(x) = \sum_{i=1}^n f_0^i(x)\frac{\partial\phi(x)}{\partial x_i} + \frac{1}{2}\sum_{k,r=1}^n\sum_{j=1}^m g_{j,0}^k(x)g_{j,0}^r(x)\frac{\partial^2\phi(x)}{\partial x_k\partial x_r}.$$

For any $i \in \{1,\dots,p\}$, denote by L_i the second order differential operator defined for any function ϕ in $C^2(\mathbb{R}^n;\mathbb{R})$ by

$$L_i\phi(x) = \sum_{k=1}^n f_i^k(x)\frac{\partial\phi(x)}{\partial x_k} + \sum_{k,r=1}^n\sum_{j=1}^m g_{j,0}^k(x)g_{j,i}^r(x)\frac{\partial^2\phi(x)}{\partial x_k\partial x_r},$$

and, for any $i,j \in \{1,\dots,p\}$, denote by L_{ij} the second order differential operator defined for any function ϕ in $C^2(\mathbb{R}^n;\mathbb{R})$ by

$$L_{ij}\phi(x) = \frac{1}{2}\sum_{k,r=1}^n\sum_{\nu=1}^m g_{\nu,i}^k(x)g_{\nu,j}^r(x)\frac{\partial^2\phi(x)}{\partial x_k\partial x_r}.$$

Moreover, for any $i \in \{1,\dots,m\}$, denote by G_i the first order differential operator defined for any function ϕ in $C^1(\mathbb{R}^n;\mathbb{R})$ by

$$G_i\phi(x) = \sum_{k=1}^n g_{i,0}^k(x)\frac{\partial\phi(x)}{\partial x_k}.$$

Then the following result, which gives sufficient conditions for asymptotic stabilizability in probability of the stochastic differential system (2), holds.

THEOREM 3.1. *Assume that there exists a smooth Lyapunov function V defined on $\mathbb{R}^n$ such that the following hold:*

(1) *$L_0V(x) \le 0$ for every $x \in \mathbb{R}^n$.*

(2) *The set*

$$\mathcal{K} = \Bigg\{ x \in \mathbb{R}^n \;/\; G_{i_0}^{\alpha_0}L_0^{\beta_0}\cdots G_{i_k}^{\alpha_k}L_0^{\beta_k}L_jV(x) = 0$$
$$\text{and} \quad G_{i_0}^{\alpha_0}L_0^{\beta_0}\cdots G_{i_k}^{\alpha_k}L_0^{\beta_k+1}V(x) = 0$$
$$\forall j \in \{1,\dots,p\}, \forall k \in \mathbb{N}, \forall i_0,\dots,i_k \in \{1,\dots,m\},$$
$$\forall \alpha_0,\beta_0,\dots,\alpha_k,\beta_k \in \{0,\dots,k\},$$
$$\text{such that } \sum_{i=0}^k(\alpha_i+\beta_i) = k \Bigg\}$$

is reduced to $\{0\}$.

Then the control law u, defined on $\mathbb{R}^n$ by

$$u^j(x) = -\frac{L_jV(x)}{\beta(x)} \quad 1 \le j \le p, \tag{3}$$

where $\beta(x) = 1 + (\sup_{1\le i,j\le p} L_{ij}V(x))^2$, renders the stochastic differential system (2) asymptotically stable in probability.

REMARK 3.2. (1) *Hypothesis (1) implies, according to the stochastic Lyapunov theorem (Theorem 1.2), that the equilibrium solution of the unforced stochastic differential system deduced from (2) is stable in probability.*

(2) *In the definition of the control law u given in* (3)*, one can choose, instead of the one given above, any positive function β mapping $\mathbb{R}^n$ into $\mathbb{R}$ such that*

$$L_{ij}V(x) < \beta(x)$$

for any $x \in \mathbb{R}^n$ and $i, j \in \{1, \ldots, p\}$.

(3) *The existence of a unique solution for the closed-loop system deduced from* (2) *is ensured by the application of Theorem* 4.1 *from Khasminskii* [4].

Proof of Theorem 3.1. Denoting by $\mathcal{L}$ the infinitesimal generator of the stochastic process solution of the closed-loop system deduced from (2) with the state feedback law u given by (3), one gets, for every $x \in \mathbb{R}^n$,

$$\mathcal{L}V(x) = L_0V(x) - \frac{1}{\beta(x)}\sum_{i=1}^{p}(L_iV(x))^2 + \frac{1}{\beta(x)^2}\sum_{i,j=1}^{p} L_iV(x)L_jV(x)L_{ij}V(x). \tag{4}$$

Then, taking hypothesis (1) and the definition of $\beta(x)$ into account, one has

$$\mathcal{L}V(x) \leq 0$$

for every $x \in \mathbb{R}^n$, and, according to the stochastic Lyapunov theorem (Theorem 1.2), the equilibrium solution $x_t \equiv 0$ of the closed-loop system deduced from (2) is stable in probability.

Furthermore, the stochastic version of La Salle's invariance theorem (Theorem 1.3) asserts that the stochastic process solution of the closed-loop system tends to the largest invariant set whose support is contained in the locus $\mathcal{L}V(x_t) \equiv 0$ for every $t \geq 0$ with probability 1.

However, one can deduce easily from (4) that $\mathcal{L}V(x_t) \equiv 0$ for every $t \geq 0$ if and only if $L_iV(x_t) \equiv 0$, $0 \leq i \leq p$, for every $t \geq 0$.

Moreover, applying Itô's formula to the stochastic processes $L_iV(x_t)$, $0 \leq i \leq p$, yields that if $L_iV(x_t) \equiv 0$, $0 \leq i \leq p$, for every $t \geq 0$, one has $L_0^2V(x_t) \equiv 0$, $G_iL_0V(x_t) \equiv 0$, $1 \leq i \leq m$, $L_0L_iV(x_t) \equiv 0$, $1 \leq i \leq p$, and $G_i(L_jV)(x_t) \equiv 0$, $1 \leq i \leq m$, $1 \leq j \leq p$, for every $t \geq 0$.

Therefore, by inductive applications of Itô's formula, one can prove that if $\mathcal{L}V(x_t) \equiv 0$ for every $t \geq 0$, one has $x_t \in \mathcal{K}$ for every $t \geq 0$ and, consequently, according to hypothesis (2), $x_t \equiv 0$ for every $t \geq 0$.

Hence, according to Theorem 1.3, the stochastic process solution of the closed-loop system deduced from (2) tends to 0 with probability 1 and thus is asymptotically stable in probability. □

Remark 3.3. *As in* [2]*, by not fully using all of the information obtained after applying Itô's formula to the stochastic processes $L_iV(x_t) \equiv 0$, $0 \leq i \leq p$, one does not take into account the fact that $G_jL_iV(x_t) \equiv 0$, $1 \leq j \leq m$, $0 \leq i \leq p$, for every $t \geq 0$. As a consequence, the stabilizability condition obtained in* [2] *is simpler than that stated in hypothesis* (2) *but does not allow us to compute stabilizing state feedback laws in many applications. In particular, if the Lyapunov function V is such that $L_0V(x) \equiv 0$ for every $x \in \mathbb{R}^n$, then the result proved in* [2] *does not permit us to make a conclusion about the asymptotic stabilizability in probability of the system, whereas the result of the above theorem still applies.*

4. A working example. Let x_0 be given in $\mathbb{R}^2$, and denote by $x_t \in \mathbb{R}^2$ the solution of the stochastic differential system

$$dx_t = \begin{pmatrix} -\frac{1}{2}x_{1,t} \\ -\frac{1}{2}x_{2,t} \end{pmatrix} dt + u \begin{pmatrix} x_{2,t} \\ 0 \end{pmatrix} dt + \begin{pmatrix} x_{2,t} \\ x_{1,t} \end{pmatrix} dw_t, \tag{5}$$

where $(w_t)_{t\geq 0}$ is a standard real-valued Wiener process and u is a real-valued measurable control law.

Note that the equilibrium solution of the stochastic differential system (5) is stable in probability but not asymptotically stable in probability.

Then, taking the Lyapunov function V defined on $\mathbb{R}^2$ by

$$V(x) = \frac{1}{2}\left(x_1^2 + x_2^2\right),$$

one has

$$L_0V(x) = 0, \qquad L_1V(x) = x_1x_2, \qquad \text{and} \qquad L_0L_1V(x) = 0$$

for every $x \in \mathbb{R}^2$, and the stabilizability conditions stated in Theorem 3.2 in [2] are not satisfied.

However, for any $x \in \mathbb{R}^2$, one has

$$G_1L_1V(x) = x_1^2 + x_2^2,$$

and hence it is obvious that the set $\mathcal{K}$ defined in hypothesis (2) of Theorem 3.1 is reduced to $\{0\}$.

Therefore, by application of the result of Theorem 3.1, one gets that the state feedback law u defined on $\mathbb{R}^2$ by

$$u(x) = -x_1x_2$$

renders the stochastic differential system (5) asymptotically stable in probability.

REFERENCES

[1] R. Chabour and M. Oumoun, *A Jurdjevic–Quinn theorem for stochastic nonlinear systems*, Stochastic Anal. Appl., 16 (1998), pp. 43–50.

[2] P. Florchinger, *A stochastic version of Jurdjevic–Quinn theorem*, Stochastic Anal. Appl., 12 (1994), pp. 473–480.

[3] V. Jurdjevic and J. P. Quinn, *Controllability and stability*, J. Differential Equations, 28 (1978), pp. 381–389.

[4] R. Z. Khasminskii, *Stochastic Stability of Differential Equations*, Sijthoff and Noordhoff, Alphen aan den Rijn, Germantown, MD, 1980.

[5] H. J. Kushner, *Stochastic stability*, in Stability of Stochastic Dynamical Systems, R. Curtain, ed., Lecture Notes in Math. 294, Springer-Verlag, Berlin, Heidelberg, New York, 1972, pp. 97–124.

[6] K. K. Lee and A. Arapostathis, *Remarks on smooth feedback stabilization of nonlinear systems*, Systems Control Lett., 10 (1988), pp. 41–44.

[7] R. Outbib and G. Sallet, *Stabilizability of the angular velocity of a rigid body revisited*, Systems Control Lett., 18 (1992), pp. 93–98.

[8] J. Tsinias, *Sufficient Lyapunov–like conditions for stabilization*, Math. Control Signals Systems, 2 (1989), pp. 343–357.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 1, pp. 89–112

STABILITY OF PLANAR SWITCHED SYSTEMS: THE LINEAR SINGLE INPUT CASE*

UGO BOSCAIN†

Abstract. We study the stability of the origin for the dynamical system $\dot{x}(t) = u(t)Ax(t) + (1 - u(t))Bx(t)$, where A and B are two 2×2 real matrices with eigenvalues having strictly negative real part, $x \in \mathbf{R}^2$, and $u(.) : [0, \infty[\to [0, 1]$ is a completely random measurable function. More precisely, we find a (coordinates invariant) necessary and sufficient condition on A and B for the origin to be asymptotically stable for each function $u(.)$.

The result is obtained without looking for a common Lyapunov function but studying the locus in which the two vector fields Ax and Bx are collinear. There are only three relevant parameters: the first depends only on the eigenvalues of A, the second depends only on the eigenvalues of B, and the third contains the interrelation among the two systems, and it is the cross ratio of the four eigenvectors of A and B in the projective line $\mathbf{C}P^1$. In the space of these parameters, the shape and the convexity of the region in which there is stability are studied.

This bidimensional problem assumes particular interest since linear systems of higher dimensions can be reduced to our situation.

Key words. stability, planar, random switching function, switched systems

AMS subject classifications. 93D20, 37N35

PII. S0363012900382837

1. Introduction.

By a switched system we mean a family of continuous-time dynamical systems and a rule that determines at any time which dynamical system is responsible for the time evolution. More precisely, let $\{f_u : \ u \in U\}$ be a (finite or infinite) set of sufficiently regular vector fields on a manifold M, and consider the family of dynamical systems:

$$\dot{x} = f_u(x), \quad x \in M. \tag{1}$$

The rule is given assigning the so-called switching function $u(.) : [0, \infty[\to U$. Here we consider the situation in which the switching function cannot be predicted a priori; it is given from outside and represents some phenomena (e.g., a disturbance) that it is not possible to control or include in the dynamical system model.

In the following, we use the notation $u \in U$ to label a fixed individual system and $u(.)$ to indicate the switching function.

Suppose now that all of the f_u have a given property for every $u \in U$. A typical problem is to study under which conditions this property holds for the system (1) for arbitrary switching functions. For a discussion of various issues related to switched systems, we refer the reader to [8].

In [1, 7] the case of switched linear systems was considered:

$$\dot{x} = A_u x, \quad x \in \mathbf{R}^n, \quad A_u \in \mathbf{R}^{n \times n}, \quad u \in U, \tag{2}$$

and the problem of the asymptotic stability of the origin for arbitrary switching functions was investigated. Clearly we need the asymptotic stability of each single

*Received by the editors December 22, 2000; accepted for publication (in revised form) October 17, 2001; published electronically April 2, 2002. This work was supported by a TMR fellowship (Non Linear Control Network), contract FMRX-CT97-0137 (DG 12-BDCN) (CNRS CON00P140DR04).
http://www.siam.org/journals/sicon/41-1/38283.html

†Université de Bourgogne, Département de Mathématiques, Analyse Appliquée et Optimisation, 9, Avenue Alain Savary B.P., 47870-21078 Dijon, France (uboscain@u-bourgogne.fr).

subsystem $\dot{x} = A_u x$, $u \in U$, in order to have the asymptotic stability of (2) for each switching function (i.e., the eigenvalues of each matrix A_u must have strictly negative real part). This will be assumed to be the case throughout the paper.

Notice the important point that in the case of linear systems, the asymptotic stability for arbitrary switching functions is equivalent to the more often quoted property of global exponential stability, uniform with respect to switching (GUES); see, for example, [2] and references therein.

In [1, 7], it is shown that the structure of the Lie algebra generated by the matrices A_u,

$$\mathbf{g} = \{A_u : \ u \in U\}_{L.A.},$$

is crucial for the stability of the system (2) (i.e., the interrelation among the systems). The main result of [7] is the following theorem.

THEOREM 1.1 (Hespanha, Morse, Liberzon). *If* $\mathbf{g}$ *is a solvable Lie algebra, then the switched system* (2) *is asymptotically stable for each switching function* $u(.) : [0, \infty[\to U$.

In [1] a generalization was given. Let $\mathbf{g} = \mathbf{r} \ \exists \!\!\!+ \ \mathbf{s}$ be the Levi decomposition of $\mathbf{g}$ in its radical (i.e., the maximal solvable ideal of $\mathbf{g}$) and a semisimple subalgebra, where the symbol $\exists \!\!\!+$ indicates the semidirect sum.

THEOREM 1.2 (Agrachev, Liberzon). *If* $\mathbf{s}$ *is a compact Lie algebra, then the system* (2) *is asymptotically stable for every switching function* $u(.) : [0, \infty[\to U$.

Theorem 1.2 contains Theorem 1.1 as a special case. Anyway, the converse of Theorem 1.2 is not true in general: if $\mathbf{s}$ is noncompact, the system can be stable or unstable. This case was also investigated. In particular, if $\mathbf{s}$ is noncompact, then it contains as a subalgebra $sl(2, \mathbf{R})$. Due to that, in the case in which $\mathbf{g}$ has dimension at most 4 as Lie algebra, the authors were able to reduce the problem of the asymptotic stability of the system (2) to the problem of the asymptotic stability of an auxiliary bidimensional system. We refer the reader to [1] for details. For this reason, the bidimensional problem assumes particular interest, and in this paper we give the complete description of that case for a single input system.

More precisely, we study the stability of the origin for the switched system

$$\dot{x}(t) = u(t)Ax(t) + (1 - u(t))Bx(t), \tag{3}$$

where A and B are two 2×2 real matrices with eigenvalues having strictly negative real part, $x \in \mathbf{R}^2$, and $u(.) : [0, \infty[\to [0, 1]$ is an arbitrary measurable switching function.

It is well known that asymptotic stability for linear switching systems is equivalent to the existence of a common Lyapunov function. In [11] necessary and sufficient conditions were obtained for linear bidimensional systems to share a common *quadratic* Lyapunov function, but there are linear bidimensional systems for which this function may fail to be quadratic (see [6]) so that the problem of finding necessary and sufficient conditions on A and B for the asymptotic stability of the system (3) was open in general.

In this paper, we give the solution to this problem. Our result is obtained with a direct method without looking for a common Lyapunov function but analyzing the locus in which the two vector fields are collinear, to build the "worst trajectory," similarly to what people do in optimal synthesis problems on the plane (see [4, 5, 9, 10]). We also use the concept of feedback. The idea of building the worst trajectory was used also in [6] for analyzing an example.

Three cases are analyzed separately. In the first case, both matrices have complex eigenvalues (in the following **(CC)** case). In the second case, one of the two matrices has real and the other has complex eigenvalues (in the following **(RC)** case). In the third case, both the matrices have real eigenvalues (in the following **(RR)** case).

There are only three relevant parameters: one depends on the eigenvalues of A, one on the eigenvalues of B (we call them, respectively, ρ_A and ρ_B), and the last contains the interrelation among the two systems, and it is the cross ratio of the four eigenvectors of A and B in the projective line $\mathbf{C}P^1$.

The result can be obtained quite easily except in one case in which the integration of the vector fields has to be done. In this case, the computations are not difficult but long, and they are collected in Appendices A and B. In the **(CC)** and **(RR)** cases, we are even able to write the final result in a relatively compact way (see formulas (5) and (7)).

Fixing the value of the cross ratio, we study the region $\mathcal{R}$ in which the system is asymptotically stable for arbitrary switching functions in the space of the parameters ρ_A and ρ_B. In the **(CC)** and **(RR)** cases it is constituted by one or two open unbounded convex regions, while in the **(RC)** case it is an open unbounded region but not always convex.

In section 2 we give the basic definitions, we study the properties of the parameters describing the problem, and we state the stability theorem giving the main ideas of the proof. In section 3 we prove the stability theorem separately for the three cases **(CC)**, **(RC)**, **(RR)**, and we give some examples. In section 4 we study the shape and the convexity of the region $\mathcal{R}$ for fixed values of the cross ratio. In section 5 we make some final remarks.

2. Basic definitions and statement of the main results. Let A and B be two diagonalizable 2×2 real matrices with eigenvalues having strictly negative real part. Consider the following property:

($\mathcal{P}$) The dynamical system in $\mathbf{R}^2$: $\dot{x}(t) = u(t)Ax(t) + (1-u(t))Bx(t)$ is asymptotically stable at the origin for each measurable function $u(.) : [0, \infty[\to [0, 1]$.

In this section we state the necessary and sufficient conditions on A and B under which ($\mathcal{P}$) holds. Moreover, we state under which conditions we have at least stability (not asymptotic) for each function $u(.)$.

Set $M(u) := uA + (1-u)B$, $u \in [0,1]$. In the class of constant functions the asymptotic stability of the origin of the system (3) occurs iff the matrix $M(u)$ has eigenvalues with strictly negative real part for each $u \in [0,1]$. So this is a necessary condition. On the other hand, it is known that if $[A, B] = 0$, then the system (3) is asymptotically stable for each function $u(.)$. So in the following we will always assume the following conditions:

H1. Let λ_1, λ_2 (resp., λ_3, λ_4) be the eigenvalues of A (resp., B). Then $\mathrm{Re}(\lambda_1)$, $\mathrm{Re}(\lambda_2)$, $\mathrm{Re}(\lambda_3)$, $\mathrm{Re}(\lambda_4) < 0$.

H2. $[A, B] \neq 0$. (That implies that neither A nor B are proportional to the identity.)

For simplicity we will also assume the following.

H3. A and B are diagonalizable. (Notice that if **H2** and **H3** hold, then $\lambda_1 \neq \lambda_2$, $\lambda_3 \neq \lambda_4$.)

H4. Let $\mathbf{V}_1, \mathbf{V}_2 \in \mathbf{C}P^1$ (resp., $\mathbf{V}_3, \mathbf{V}_4 \in \mathbf{C}P^1$) be the eigenvectors of A (resp., B). From **H2** and **H3** we know that they are uniquely defined, and $\mathbf{V}_1 \neq \mathbf{V}_2$ and $\mathbf{V}_3 \neq \mathbf{V}_4$. We assume $\mathbf{V}_i \neq \mathbf{V}_j$ for $i \in \{1,2\}$, $j \in \{3,4\}$.

The degenerate cases, in which **H1** and **H2** hold and **H3** or **H4** or both do not, are the following:

- A or B is not diagonalizable. This case (in which $(\mathcal{P})$ can be true or false) can be treated with techniques entirely similar to the ones of this paper.
- A or B is diagonalizable, but one eigenvector of A coincides with one eigenvector of B. In this case, using arguments similar to the ones of the next section, it is possible to conclude that $(\mathcal{P})$ is true.

Remark 1. One can easily prove that (under the hypotheses **H2** and **H3**), **H4** can be violated only in the **(RR)** case (see also subsection 3.3). Moreover, hypotheses **H2**, **H3**, and **H4** imply that $\mathbf{V}_i \neq \mathbf{V}_j$ for $i, j \in \{1, 2, 3, 4\}$, $i \neq j$. This fact permits us to define the cross ratio without additional hypotheses (see the definition of cross ratio below).

Theorem 2.3 gives necessary and sufficient conditions for the stability of the system (3) in terms of three (coordinates invariant) parameters defined in Definition 2.1. The first (ρ_A) depends on the eigenvalues of A, the second (ρ_B) depends on the eigenvalues of B, and the third ($\mathcal{K}$) depends on $\mathrm{Tr}(AB)$, which is a standard scalar product in the space of 2×2 matrices. Proposition 2.2 gives some properties of these parameters. Finally, Proposition 2.4 shows the geometrical meaning of $\mathcal{K}$. It is in one-to-one correspondence with the cross ratio of the four points in the projective line $\mathbf{C}P^1$ that corresponds to the four eigenvectors of A and B. This parameter contains the interrelation among the two systems.

Definition 2.1. *Let A and B be two 2×2 real matrices, and suppose that* **H1**, **H2**, **H3**, *and* **H4** *hold. Moreover, choose the labels* $(_1)$ *and* $(_2)$ *(resp.,* $(_3)$ *and* $(_4)$*) in such a way that* $|\lambda_2| > |\lambda_1|$ *(resp.,* $|\lambda_3| > |\lambda_4|$*) if they are real or* $\mathrm{Im}(\lambda_2) < 0$ *(resp.,* $\mathrm{Im}(\lambda_4) < 0$*) if they are complex. Define*

$$\rho_A := -i\frac{\lambda_1 + \lambda_2}{\lambda_1 - \lambda_2}; \qquad \rho_B := -i\frac{\lambda_3 + \lambda_4}{\lambda_3 - \lambda_4}; \qquad \mathcal{K} := 2\frac{\mathrm{Tr}(AB) - \frac{1}{2}\mathrm{Tr}(A)\mathrm{Tr}(B)}{(\lambda_1 - \lambda_2)(\lambda_3 - \lambda_4)}.$$

Moreover, define the following function of $\rho_A, \rho_B, \mathcal{K}$:

$$\mathcal{D} := \mathcal{K}^2 + 2\rho_A\rho_B\mathcal{K} - (1 + \rho_A^2 + \rho_B^2). \tag{4}$$

Notice that $\rho_A \in \mathbf{R}$, $\rho_A > 0$, iff A has complex eigenvalues and $\rho_A \in i\mathbf{R}$, $\rho_A/i > 1$, iff A has real eigenvalues. The same holds for B. Moreover, $\mathcal{D} \in \mathbf{R}$. The parameter $\mathcal{K}$ contains important information about the matrices A and B. They are stated in the following proposition, which can be easily proved using the systems of coordinates of the next section (see also [3]).

Proposition 2.2. *Let A and B be as in Definition* 2.1. *We have the following:*

- *if A and B have both complex eigenvalues, then $\mathcal{K} \in \mathbf{R}$ and $|\mathcal{K}| > 1$;*
- *if A and B have both real eigenvalues, then $\mathcal{K} \in \mathbf{R} \setminus \{\pm 1\}$;*
- *A and B have one complex and the other real eigenvalues iff $\mathcal{K} \in i\mathbf{R}$.*

Theorem 2.3. *Let A and B be two real matrices such that* **H1**, **H2**, **H3**, *and* **H4** *hold, and define $\rho_A, \rho_B, \mathcal{K}, \mathcal{D}$ as in Definition* 2.1. *We have the following stability conditions:*

Case (CC) *If A and B have both complex eigenvalues, then:*

Case (CC.1) *if $\mathcal{D} < 0$, then $(\mathcal{P})$ is true;*

Case (CC.2) *if $\mathcal{D} > 0$, then:*

Case (CC.2.1) *if $\mathcal{K} < -1$, then $(\mathcal{P})$ is false;*

Case (CC.2.2) *if $\mathcal{K} > 1$, then $(\mathcal{P})$ is true iff the following condition holds:*

$$\rho_{CC} := \exp\left[-\rho_A \arctan\left(\frac{-\rho_A\mathcal{K}+\rho_B}{\sqrt{\mathcal{D}}}\right) - \rho_B \arctan\left(\frac{\rho_A-\rho_B\mathcal{K}}{\sqrt{\mathcal{D}}}\right) - \frac{\pi}{2}(\rho_A+\rho_B)\right] \times \sqrt{\frac{(\rho_A\rho_B+\mathcal{K})+\sqrt{\mathcal{D}}}{(\rho_A\rho_B+\mathcal{K})-\sqrt{\mathcal{D}}}} < 1. \tag{5}$$

Case (CC.3) *If $\mathcal{D} = 0$, then $(\mathcal{P})$ is true or false according, respectively, to the fact that $\mathcal{K} > 1$ or $\mathcal{K} < -1$.*

Case (RC) *If A and B have one complex and the other real eigenvalues, define $\chi := \rho_A\mathcal{K} - \rho_B$, where ρ_A and ρ_B are chosen in such a way that $\rho_A \in i\mathbf{R}$, $\rho_B \in \mathbf{R}$. Then:*

Case (RC.1) *if $\mathcal{D} > 0$, then $(\mathcal{P})$ is true;*

Case (RC.2) *if $\mathcal{D} < 0$, then $\chi \neq 0$, and we have:*

Case (RC.2.1) *if $\chi > 0$, then $(\mathcal{P})$ is false. Moreover, in this case $\mathcal{K}/i < 0$;*

Case (RC.2.2) *if $\chi < 0$, then:*

Case (RC2.2.A) *if $\mathcal{K}/i \leq 0$, then $(\mathcal{P})$ is true;*

Case (RC2.2.B) *if $\mathcal{K}/i > 0$, then $(\mathcal{P})$ is true iff the following condition holds:*

$$\rho_{RC} := e^{-\rho_B(\xi^+-\xi^-)}\sqrt{\frac{\cos^2\xi^+ + E^2\sin^2\xi^+}{\cos^2\xi^- + E^2\sin^2\xi^-}} \times \sqrt{\left(\frac{m^+}{m^-}\right)^{\frac{1}{2}(-\rho_A/i+1)}\cos^2\theta^+ + \left(\frac{m^+}{m^-}\right)^{\frac{1}{2}(-\rho_A/i-1)}\sin^2\theta^+} < 1, \tag{6}$$

where:
$$E := \mathcal{K}/i + \sqrt{-\mathcal{K}^2+1},$$
$$m^\pm := \frac{-\chi \pm \sqrt{-\mathcal{D}}}{(-\rho_A/i-1)\mathcal{K}/i},$$
$$\theta^+ := \arctan m^+,$$
$$\xi^\pm := \arctan\left(\frac{m^\pm - 1}{E(m^\pm+1)}\right), \quad \xi^+ \in]\xi^-, \xi^- + \pi[.$$

Case (RC.3) *If $\mathcal{D} = 0$, then $(\mathcal{P})$ is true or false according, respectively, to the fact that $\chi < 0$ or $\chi > 0$.*

Case (RR) *If A and B have both real eigenvalues, then:*

Case (RR.1) *if $\mathcal{D} < 0$, then $(\mathcal{P})$ is true. Moreover, we have $|\mathcal{K}| > 1$;*

Case (RR.2) *if $\mathcal{D} > 0$, then $\mathcal{K} \neq -\rho_A\rho_B$ (notice that $-\rho_A\rho_B > 1$) and:*

Case (RR.2.1) *if $\mathcal{K} > -\rho_A\rho_B$, then (P) is false*

Case (RR.2.2) *if $\mathcal{K} < -\rho_A\rho_B$, then:*

Case (RR.2.2.A) *if $\mathcal{K} > -1$, then (P) is true;*

Case (RR.2.2.B) *if $\mathcal{K} < -1$, then (P) is true iff the following condition holds:*

$$\rho_{RR} := -f^{sym}(\rho_A,\rho_B,\mathcal{K})f^{asym}(\rho_A,\rho_B,\mathcal{K}) \times f^{asym}(\rho_B,\rho_A,\mathcal{K}) < 1, \tag{7}$$

where:

$$f^{sym}(\rho_A,\rho_B,\mathcal{K}) := \frac{1+\rho_A/i+\rho_B/i+\mathcal{K}-\sqrt{\mathcal{D}}}{1+\rho_A/i+\rho_B/i+\mathcal{K}+\sqrt{\mathcal{D}}};$$

$$f^{aym}(\rho_A,\rho_B,\mathcal{K}) := \left(\frac{\rho_B/i-\mathcal{K}\rho_A/i-\sqrt{\mathcal{D}}}{\rho_B/i-\mathcal{K}\rho_A/i+\sqrt{\mathcal{D}}}\right)^{\frac{1}{2}(\rho_A/i-1)}.$$

Case (RR.3) *If $\mathcal{D}=0$, then $(\mathcal{P})$ is true or false according, respectively, to the fact that $\mathcal{K}<-\rho_A\rho_B$ or $\mathcal{K}>-\rho_A\rho_B$.*

Finally, if $(\mathcal{P})$ is false, then in case **(CC.2.2)** *with $\rho_{CC}=1$, case* **(RC.2.2.B)** *with $\rho_{RC}=1$, case* **(RR.2.2.B)** *with $\rho_{RR}=1$, case* **(CC.3)** *with $\mathcal{K}<-1$, case* **(RC.3)** *with $\chi>0$, and case* **(RR.3)** *with $\mathcal{K}>-\rho_A\rho_B$, for every $C>0$, there exists $C'\le C$ such that if $|\gamma(0)|<C'$, then $|\gamma(t)|<C$ for every $t\in[0,\infty[$ (i.e., we have stability of the origin). In the other cases, there exists a trajectory $\gamma(t)$ such that $\lim_{t\to\infty}|\gamma(t)|=\infty$.*

Notice that the expressions (5) and (7) are invariant if we exchange ρ_A with ρ_B. The last statement says when we have at least stability (not asymptotic) for every switching function.

Let us study the geometric meaning of $\mathcal{K}$. Let $\mathbf{V}_1,\mathbf{V}_2,\mathbf{V}_3,\mathbf{V}_4$ belong to the complex projective line $\mathbf{C}P^1$. Suppose $\mathbf{V}_1\neq\mathbf{V}_2\neq\mathbf{V}_3$, and let (v_1,v_1'), (v_2,v_2'), (v_3,v_3'), (v_4,v_4') be the corresponding homogeneous coordinates. The *cross ratio* $\beta(\mathbf{V}_1,\mathbf{V}_2,\mathbf{V}_3,\mathbf{V}_4)$ is defined in the following way. Make a Moebius transformation such that $\mathbf{V}_1,\mathbf{V}_2$ become the fundamental points (i.e., of homogeneous coordinates, respectively, $(0,1)$ and $(1,0)$) and $\mathbf{V}_3$ the unity point (i.e., of homogeneous coordinates $(1,1)$), and let $(\bar v_4,\bar v_4')$ be the new homogeneous coordinates of $\mathbf{V}_4$. By definition we have

$$\beta(\mathbf{V}_1,\mathbf{V}_2,\mathbf{V}_3,\mathbf{V}_4) := \bar v_4'/\bar v_4 = \frac{\begin{vmatrix} v_1 & v_4 \\ v_1' & v_4' \end{vmatrix}\begin{vmatrix} v_2 & v_3 \\ v_2' & v_3' \end{vmatrix}}{\begin{vmatrix} v_2 & v_4 \\ v_2' & v_4' \end{vmatrix}\begin{vmatrix} v_1 & v_3 \\ v_1' & v_3' \end{vmatrix}}. \tag{8}$$

Now the four eigenvectors of A and B are exactly four directions in $\mathbf{C}^2$; i.e., they can be regarded as four points of $\mathbf{C}P^1$, and under the conditions **H2**, **H3**, **H4**, it makes sense to compute their cross ratio (cf. Remark 1).

One can immediately obtain (suggestion: use the systems of coordinates of the next section) the following proposition.

PROPOSITION 2.4. *Let A and B be two 2×2 real matrices such that* **H1**, **H2**, **H3**, *and* **H4** *hold, and let $\mathbf{V}_1,\mathbf{V}_2,\mathbf{V}_3,\mathbf{V}_4$ be the four points in the space $\mathbf{C}P^1$ corresponding, respectively, to the four eigenvectors of A and B chosen in such a way that they correspond, respectively, to $\lambda_1,\lambda_2,\lambda_3,\lambda_4$ (see Definition* 2.1*). Let $\beta(\mathbf{V}_1,\mathbf{V}_2,\mathbf{V}_3,\mathbf{V}_4)$ be their cross ratio and $\mathcal{K}$ the quantity defined in Definition* 2.1*. Then $\beta(\mathbf{V}_1,\mathbf{V}_2,\mathbf{V}_3,\mathbf{V}_4)$ and $\mathcal{K}$ are in the one-to-one relation from $\mathbf{C}\cup\{\infty\}$ to $\mathbf{C}\cup\{\infty\}$:*

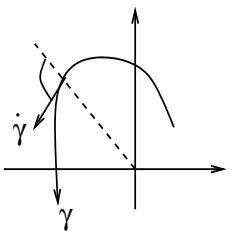
$$\mathcal{K}=\frac{\beta(\mathbf{V}_1,\mathbf{V}_2,\mathbf{V}_3,\mathbf{V}_4)+1}{\beta(\mathbf{V}_1,\mathbf{V}_2,\mathbf{V}_3,\mathbf{V}_4)-1},\qquad \beta(\mathbf{V}_1,\mathbf{V}_2,\mathbf{V}_3,\mathbf{V}_4)=\frac{\mathcal{K}+1}{\mathcal{K}-1}.$$

Notice that $\mathcal{K} \neq \infty$ so that $\beta \neq 1$. From Proposition 2.4 and Definition 2.1 we have the following expression for the cross ratio of the eigenvectors of A and B:

$$\beta = \frac{\mathrm{Tr}(AB) - (\lambda_1\lambda_4 + \lambda_2\lambda_3)}{\mathrm{Tr}(AB) - (\lambda_1\lambda_3 + \lambda_2\lambda_4)}.$$

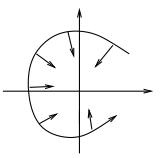
Theorem 2.3 is proved in the next section. Here we describe the main idea of the proof.

We build the "worst trajectory," i.e., the trajectory that at each point has the velocity forming the angle, with the (exiting) radial direction, having the smallest absolute value, without taking care of the module of the velocity.

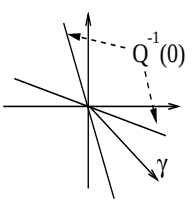


The main idea is that the system (3) is asymptotically stable iff this trajectory tends to the origin. The worst trajectory is constructed in the following way. We study the locus $Q^{-1}(0)$ (the notation is clarified later) in which the two vector fields Ax and Bx are collinear. We have several cases:

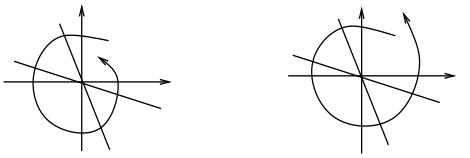
- If $Q^{-1}(0)$ contains only the origin, then, in the **(CC)** and **(RC)** cases, one vector field always points on the same side of the other, and the worst trajectory is a trajectory of the vector field Ax or Bx. In this case, the system is asymptotically stable (cases **(CC.1)** and **(RC.1)** of Theorem 2.3).

  The situation is similar in case **(RR.1).** (The worst trajectory tends to the origin.)
- If $Q^{-1}(0)$ does not contain only the origin, then it is a couple of straight lines passing from the origin (see the next section). If at each point of $Q^{-1}(0)$ the two vector fields have opposite versus, then there exists a trajectory going to infinity corresponding to a constant switching function (see the following figure).

  This corresponds to cases **(CC.2.1)**, **(RC.2.1)**, and **(RR.2.1)** of Theorem 2.3, and it is the situation in which there exists $u \in [0, 1]$ such that the matrix $M(u)$ admits an eigenvalue with positive real part. If at each point of Q the two vector fields have the same versus, then the system is asymptotically stable iff the worst trajectory turns around the origin and after one turn the distance from the origin is increasing.



This corresponds to cases **(CC.2.2)**, **(RC.2.2)**, and **(RR.2.2)** of Theorem 2.3.

- Finally, **(CC.3)**, **(RC.3)**, and **(RR.3)** are the degenerate cases in which the two straight lines coincide.

More details are given later.

3. Proof of the stability theorem. In the following, we prove Theorem 2.3 separately for the three cases in which A and B have both complex, one complex and the other real, and both real eigenvalues.

3.1. The case in which A and B have both complex eigenvalues. Let $-\delta_A \pm i\omega_A$ $(\delta_A, \omega_A > 0)$ be the eigenvalues of A and $-\delta_B \pm i\omega_B$ $(\delta_B, \omega_B > 0)$ be the eigenvalues of B. We have $\rho_A = \delta_A/\omega_A$, $\rho_B = \delta_B/\omega_B$. Choose a system of coordinates in which

$$A = \begin{pmatrix} -\delta_A & -\omega_A/E \\ \omega_A E & -\delta_A \end{pmatrix}, \qquad B = \begin{pmatrix} -\delta_B & -\omega_B \\ \omega_B & -\delta_B \end{pmatrix},$$

where $E \in \mathbf{R}\setminus\{0\}$. This corresponds to put B in the normal form in which its integral curves are "circular spirals" and then, using the invariance of B under rotation, to rotate the coordinates in such a way that the integral curves of A are elliptical spirals with axes corresponding to the x_1 and x_2 directions (see, for example, Figure 3.1). We have

$$[A, B] = \omega_A\omega_B(E - 1/E)\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

so we assume $E \neq \pm 1$; otherwise, $[A, B] = 0$. In this case we have $\mathcal{K} = \frac{1}{2}(E + \frac{1}{E})$, and without loss of generality we may assume $|E| > 1$.

The locus in which Ax and Bx are collinear is $Q^{-1}(0)$, where

$$\begin{aligned} Q = \det(Ax, Bx) &= x_1^2(-\delta_A\omega_B + \delta_B\omega_A E) \\ &+ x_1x_2(\omega_A\omega_B(E - 1/E)) + x_2^2(-\delta_A\omega_B + \delta_B\omega_A/E) \end{aligned}$$

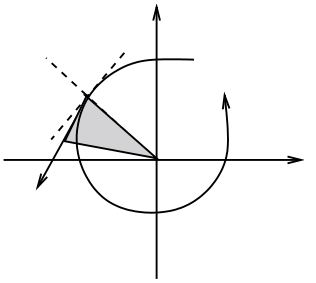
and $x = (x_1, x_2)$. Now let D_{CC} be the discriminant of the quadratic form Q. We have

$$\begin{aligned} D_{CC} &= \omega_A^2\omega_B^2(E - 1/E)^2 - 4(-\delta_A\omega_B + \delta_B\omega_A E)(-\delta_A\omega_B + \delta_B\omega_A/E) \\ &= 4\omega_A^2\omega_B^2\mathcal{D}, \end{aligned} \tag{9}$$

where $\mathcal{D}$ is defined in Definition 2.1.

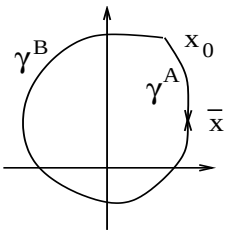
Case 1. If $\mathcal{D} < 0$, then the quadratic form Q has strictly defined sign and $Q^{-1}(0) = \{0\}$. In this case, one vector field always points on the same side of the other. Making a suitable change of coordinates and possibly exchanging the labels $(_A)$ and $(_B)$, we can realize the situation in which Ax always points on the left of Bx for every $x \in \mathbf{R}^2 \setminus \{0\}$. We have two cases.

- Suppose first that $E > 1$. In this case, Ax always points in the grey region of the following picture.



Fix an arbitrary measurable switching function $u(.) : [0, \infty[\to [0, 1]$, and let $(x_1(t), x_2(t))$ (resp., $(\rho(t), \theta(t))$) be the Cartesian (resp., polar) coordinates of the solution of $\dot{x}(t) = u(t)Ax(t) + (1-u(t))Bx(t)$, $x(0) = x_0 \in \mathbf{R}^2 \setminus \{0\}$. In this case, we have $\dot{\rho}(t) < 0$ for almost every $t \in [0, +\infty[$ and $(\mathcal{P})$ is true.

- Suppose now that $E < -1$. Fix $x_0 \in \mathbf{R}^2 \setminus \{0\}$, and let γ be a trajectory of the switched system (3) such that $\gamma(0) = x_0$. Let $\gamma^A : [0, t_A] \to \mathbf{R}^2$ (resp., $\gamma^B : [0, t_B] \to \mathbf{R}^2$) be a trajectory of the vector field Ax (resp., Bx) such that $\gamma^A(0) = x_0$ (resp., $\gamma^B(0) = x_0$), and define t_A and t_B in such a way that $\gamma^A(t_A) = \gamma^B(t_B) =: \bar{x}$ is the first intersection point of γ^A and γ^B after x_0.



Let Ω be the simply connected closed set whose border is

$$\partial\Omega = \mathrm{Supp}(\gamma^A|_{[0,t_A]} \cup \gamma^B|_{[0,t_B]}).$$

For every $x \in \partial\Omega$ we have the following. Define $V_u = uAx + (1-u)Bx$. For each $u \in]0, 1[$, V_u points inside Ω. Moreover, if $x \notin \{x_0, \bar{x}\}$, V_1 (resp., V_0) points inside Ω or it is tangent to $\partial\Omega$. Fix $\bar{t} > \max\{t_A, t_B\}$. We clearly have $\bar{x} := \gamma(\bar{t}) \in \mathrm{int}(\Omega)$. Using homothety invariance of the system (3), we may easily conclude that $\lim_{t\to\infty} \gamma(t) = 0$ and $(\mathcal{P})$ is true. This proves case **(CC.1)** of Theorem 2.3 (see Example 1 below).

Case 2. If $\mathcal{D} > 0$, then Q has no definite sign and $Q^{-1}(0)$ is a couple of noncoinciding straight lines passing from the origin and forming the following angles with the x_1 axis:

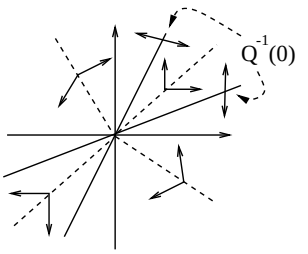
$$\theta^{\pm} = \arctan(m^{\pm}), \text{ where} \tag{10}$$

$$\begin{aligned} m^{\pm} &= \frac{-\omega_A\omega_B(E - 1/E) \pm \sqrt{D_{CC}}}{2(-\delta_A\omega_B + \delta_B\omega_A/E)} \\ &= \frac{-(E - 1/E) \pm 2\sqrt{\mathcal{D}}}{2(-\rho_A + \rho_B/E)} \text{ if } -\rho_A + \rho_B/E \neq 0, \end{aligned} \tag{11}$$

$$\begin{aligned} m^- = \infty, \quad m^+ &= \frac{\delta_A\omega_B - \delta_B\omega_A E}{\omega_A\omega_B(E - 1/E)} \\ &= \frac{\rho_A - \rho_B E}{E - 1/E} \text{ if } -\rho_A + \rho_B/E = 0, \end{aligned} \tag{12}$$

where we assume that $\theta^- \in [-\frac{\pi}{2}, \frac{\pi}{2}[$ and $\theta^+ \in]\theta^-, \theta^- + \pi[$. Notice that if $E < -1$, then $4(-\delta_A\omega_B + \delta_B\omega_A E)(-\delta_A\omega_B + \delta_B\omega_A/E) > 0$, which implies that $\theta^\pm \in [-\frac{\pi}{2}, 0[$.

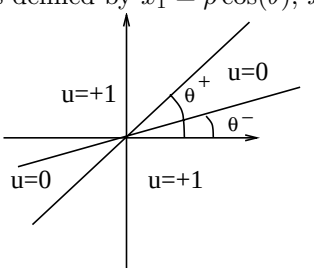
Case 2.1. If $E < -1$ ($\mathcal{K} < -1$), then at each point of $Q^{-1}(0) \setminus \{0\}$ the two vector fields have opposite versus. Consider the four connected components of $\mathbf{R}^2 \setminus Q^{-1}(0)$. In this case, for each point x_0 belonging to two of these regions (see the figure below), it is possible to find $u_0 \in [0, 1]$ such that $u_0 Ax_0 + (1 - u_0)Bx_0$ has the exiting radial direction. So the system is not stable for arbitrary switching functions. This situation corresponds to the case in which there exists $u \in [0, 1]$ such that $M(u) := uA + (1 - u)B$ admits an eigenvalue with positive real part; i.e., there exist trajectories γ corresponding to constant switching functions such that $\lim_{t\to\infty} |\gamma(t)| = \infty$. Case **(CC.2.1)** of Theorem 2.3 is proved (see Example 4 below).



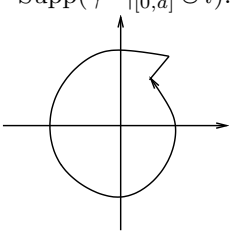
Case 2.2. If $E > 1$ ($\mathcal{K} > 1$), then at each point of $Q^{-1}(0) \setminus \{0\}$ the two vector fields have the same versus (counterclockwise). Fix $x_0 \in \mathbf{R}^2 \setminus \{0\}$, and let $\gamma^M : [0, \infty[\to \mathbf{R}^2$, $\gamma^M(0) = x_0$ be the trajectory corresponding to the feedback

$$u^M(x) = \begin{cases} 0 \text{ if } \theta \in [\theta^-, \theta^+[\text{ or } \theta \in [\theta^- + \pi, \theta^+ + \pi[, \\ +1 \text{ if } \theta \in [\theta^+, \theta^- + \pi[\text{ or } \theta \in [\theta^+ + \pi, \theta^- + 2\pi[, \end{cases} \tag{13}$$

where $\theta \in [\theta^-, \theta^- + 2\pi[$ is defined by $x_1 = \rho\cos(\theta)$, $x_2 = \rho\sin(\theta)$.



Let $(\rho^M(t), \theta^M(t))$ be the polar coordinates of γ^M and a the time defined by $\theta^M(a) = \theta^M(0) + 2\pi$. If $\rho^M(a) < \rho^M(0)$, then let l be the segment joining the points $(\rho^M(0), \theta^M(0))$ with $(\rho^M(a), \theta^M(a))$ and Ω the simply connected region whose border is $\partial\Omega := \mathrm{Supp}(\gamma^M|_{[0,a]} \cup l)$.



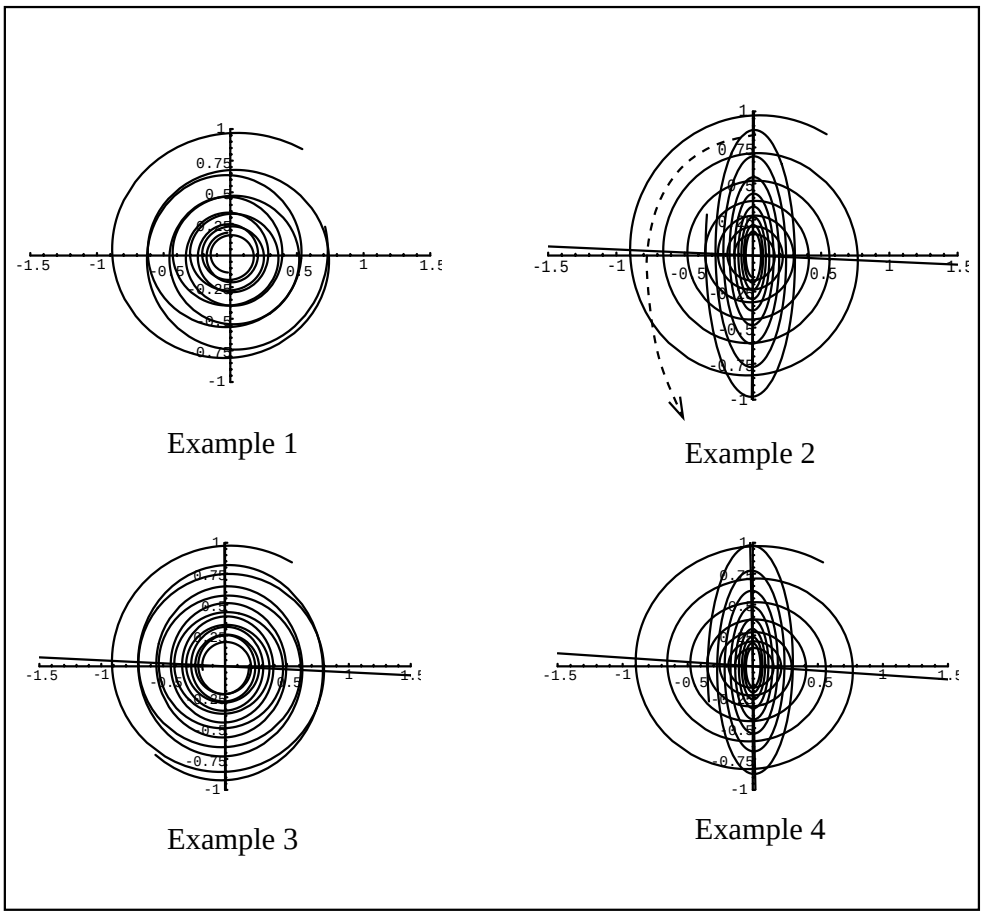


FIG. 3.1. *Examples in the* **(CC)** *case.*

For every $x \in \partial\Omega$, we have the following. Define V_u as in Case 1, $E < -1$. For each $u \in]0,1[$, V_u points inside Ω. Moreover, if $x \notin \{\gamma^M(0), \gamma^M(a)\}$, V_1 (resp., V_0) points inside Ω or is tangent to $\partial\Omega$. Similarly to Case 1 ($E < 1$), we can conclude that $(\mathcal{P})$ is true (see Example 3 below). On the other hand if $\rho^M(a) \geq \rho^M(0)$, then $\gamma^M(t)$ does not tend to the origin and $(\mathcal{P})$ is false (see Example 2 below). The condition $\rho^M(a) < \rho^M(0)$ is satisfied iff condition (5) holds. Formula (5) is obtained in Appendix A . The condition $\rho^M(a) = \rho^M(0)$ (i.e., $\boldsymbol{\rho_{CC} = 1}$) is the case in which we have at least stability (not asymptotic) for every switching function. This concludes the proof of case **(CC.2.2)**.

Case 3. If $\mathcal{D} = 0$, then the two straight lines coincide. If $E > 1$, it is easy to understand that we are in the same situation as that of Case 1. If $E < -1$, then to every $x \in Q$ there exists $u \in [0,1]$ such that $uAx + (1-u)Bx = 0$. In this case, $(\mathcal{P})$ is false, but we have at least stability (not asymptotic). This proves case **(CC.3)** of Theorem 2.3.

Examples. In the following, we give some examples of the various situations in the **(CC)** case. We refer to Figure 3.1.

Example 1. $\rho_A = 0.05$, $\rho_B = 0.06$, $\mathcal{K} = -1.005$. In this case, $\mathcal{D} \sim -0.002$,

and $(\mathcal{P})$ is true. In Figure 3.1, two integral curves of the vector fields Ax and Bx are shown. A similar situation but with the two trajectories rotating with the same versus can be obtained with the same values of ρ_A and ρ_B but with $\mathcal{K} = +1.00001$. In this case, $\mathcal{D} \sim -0.00008$ and $(\mathcal{P})$ is true (see case **(CC.1)**).

Example 2. $\rho_A = 0.0375,\ \rho_B = 0.05,\ \mathcal{K} = 1.67$. In this case, $\mathcal{D} \sim 1.79,\ \rho_{CC} \sim 2.62$ and $(\mathcal{P})$ is false. In Figure 3.1, two integral curves of the vector fields Ax and Bx, $\mathcal{D}$ that are the two straight lines (one almost coincides with the x_2 axis) and a trajectory γ such that $\lim_{t\to\infty} |\gamma(t)| = \infty$ (cf. case **(CC.2.2)**) are shown.

Example 3. $\rho_A = 0.0375,\ \rho_B = 0.0425,\ \mathcal{K} = 1.00455$. In this case, $\mathcal{D} \sim 0.0091,\ \rho_{CC} \sim 0.96$, and $(\mathcal{P})$ is true (cf. case **(CC.1)**).

Example 4. Suppose $\rho_A = 0.0375,\ \rho_B = 0.05,\ \mathcal{K} = -1.67$. In this case, $\mathcal{D} \sim 1.77$ and $(\mathcal{P})$ is false (cf. case **(CC.2.1)**).

3.2. The case in which A and B have one complex and the other real eigenvalues. Suppose that A has real eigenvalues $\lambda_1,\ \lambda_2$ $(\lambda_1, \lambda_2 < 0,\ |\lambda_2| > |\lambda_1|)$ and B complex eigenvalues $\lambda_3 = -\delta + i\omega,\ \lambda_4 = -\delta - i\omega$ $(\delta, \omega > 0)$. We have $\rho_A = -i(\lambda_1 + \lambda_2)/(\lambda_1 - \lambda_2)$ and $\rho_B = \delta/\omega$. We recall that $\rho_A/i > 1,\ \rho_B > 0$. Define

$$R(\varphi) := \begin{pmatrix} \cos(\varphi) & \sin(\varphi) \\ -\sin(\varphi) & \cos(\varphi) \end{pmatrix} \in SO(2), \tag{14}$$

and choose a system of coordinates in which

$$A = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \tag{15}$$

$$\begin{aligned} B &= \begin{pmatrix} a & b \\ c & d \end{pmatrix} := R^{-1}(\varphi) \begin{pmatrix} -\delta & -\omega/E \\ \omega E & -\delta \end{pmatrix} R(\varphi) \\ &= \begin{pmatrix} -\delta - \omega(E - 1/E)\sin(\varphi)\cos(\varphi) & -\omega(E\sin^2(\varphi) + 1/E\cos^2(\varphi)) \\ \omega(E\cos^2(\varphi) + 1/E\sin^2(\varphi)) & -\delta + \omega(E - 1/E)\sin(\varphi)\cos(\varphi) \end{pmatrix}. \end{aligned} \tag{16}$$

We have $\mathcal{K} = i(E - 1/E)\cos(\varphi)\sin(\varphi) \in i\mathbf{R}$, and without loss of generality we may assume that $\varphi \in [0, \pi/2[$, $|E| \geq 1$. Notice that in this case

$$[A, B] = (\lambda_1 - \lambda_2) \begin{pmatrix} 0 & b \\ -c & 0 \end{pmatrix} \neq 0 \text{ for each } \mathcal{K} \in i\mathbf{R}.$$

Similarly to the previous subsection, the locus in which Ax and Bx are collinear is $Q^{-1}(0)$, where

$$Q = \det(Ax, Bx) = x_1^2(\lambda_1 c) + x_1 x_2 \bar{\chi} + x_2^2(-\lambda_2 b),$$

and by definition $\bar{\chi} := \lambda_1 d - \lambda_2 a = (\lambda_1 + \lambda_2)\omega\mathcal{K}/i - (\lambda_1 - \lambda_2)\delta = (\lambda_1 - \lambda_2)\omega\chi$, where $\chi := \rho_A\mathcal{K} - \rho_B$ (see Theorem 2.3). In this case, the discriminant of the quadratic form Q is

$$D_{RC} = \bar{\chi}^2 + 4\lambda_1\lambda_2 bc = \bar{\chi}^2 - 4\lambda_1\lambda_2\omega^2(-\mathcal{K}^2 + 1) = -\omega^2(\lambda_1 - \lambda_2)^2\mathcal{D}. \tag{17}$$

Notice that $\chi = 0$ implies $\bar{\chi} = 0$, which implies $D_{RC} < 0$, i.e., $\mathcal{D} > 0$. Moreover, $\chi > 0$ implies $\mathcal{K}/i < 0$, which implies $E < -1$. Similarly to the previous subsection, we have the following cases.

Case 1. If $D_{RC} < 0$ $(\mathcal{D} > 0)$, then $(\mathcal{P})$ is true (see Example 1 below).

Case 2. If $D_{RC} > 0$ ($\mathcal{D} < 0$), then $Q^{-1}(0)$ is a couple of noncoinciding straight lines passing from the origin and forming the following angles with the x_1 axis:

$$\theta^{\pm} = \arctan(m^{\pm}), \tag{18}$$
$$m^{\pm} := \frac{-\bar{\chi} \pm \sqrt{D_{RC}}}{2(-\lambda_2 b)} = \frac{-\chi \pm \sqrt{-\mathcal{D}}}{2\frac{\lambda_2}{\lambda_1-\lambda_2}(E\sin^2(\varphi) + 1/E\cos^2(\varphi))}.$$

From (17) it follows that $D_{RC} < \bar{\chi}^2$ (i.e., $-\mathcal{D} < \chi^2$) so that in this case we have $\chi \neq 0$ and we may assume

$$\begin{cases} \theta^-, \theta^+ \in]0, \pi/2[\text{ if } \chi \text{ and } E \text{ have the same sign,} \\ \theta^-, \theta^+ \in]-\pi/2, 0[\text{ if } \chi \text{ and } E \text{ have opposite sign.} \end{cases}$$

Case 2.1 If $\chi > 0$, then $\mathcal{K}/i < 0$, which implies $E < -1$, and we have $\theta^-\theta^+ \in]-\pi/2, 0[$. In this case, at each point of $Q^{-1}(0) \setminus \{0\}$ the two vector fields have opposite versus. The same argument of Case 2.1 of section 3.1 shows that $(\mathcal{P})$ is false (see Example 4 below).

Case 2.2 If $\chi < 0$, then in both cases where $E \geq 1$, $E \leq -1$ at each point of $Q^{-1}(0) \setminus \{0\}$ the two vector fields have the same versus.

Case 2.2.A If $E \leq -1$ (which implies $\mathcal{K}/i \leq 0$), then $(\mathcal{P})$ is true because of the following argument.

From $\chi = \rho_A\mathcal{K} - \rho_B < 0$ we have

$$-\frac{\mathcal{K}/i}{\rho_B} < \frac{1}{\rho_A/i} < 1. \tag{19}$$

Now let γ be an integral of the vector field Bx and $(\rho(t), \theta(t))$ its polar coordinates. We have

$$\gamma(t) = R(\varphi) \begin{pmatrix} \rho_0 e^{-\delta t}\cos(\omega t + \varphi_0) \\ \rho_0 E e^{-\delta t}\sin(\omega t + \varphi_0), \end{pmatrix}$$

and $\rho(t) = \rho_0 e^{-\delta t}\sqrt{\cos^2(\omega t + \varphi_0) + E^2\sin^2(\omega t + \varphi_0)}$. Now we prove that the condition (19) implies $\dot{\rho}(t) \leq 0$ for every $t \in Dom(\gamma)$, which clearly implies that $(\mathcal{P})$ is true. We have

$$\dot{\rho}(t) = \rho_0 e^{-\delta t}\left(\frac{(E^2-1)\omega\sin(\omega t + \varphi_0)\cos(\omega t + \varphi_0)}{\sqrt{\cos^2(\omega t + \varphi_0) + E^2\sin^2(\omega t + \varphi_0)}} - \delta\sqrt{\cos^2(\omega t + \varphi_0) + E^2\sin^2(\omega t + \varphi_0)}\right).$$

Therefore, $\dot{\rho}(t) < 0$ iff

$$\frac{(E^2-1)\omega\sin(\omega t + \varphi_0)\cos(\omega t + \varphi_0)}{\sqrt{\cos^2(\omega t + \varphi_0) + E^2\sin^2(\omega t + \varphi_0)}} - \delta\sqrt{\cos^2(\omega t + \varphi_0) + E^2\sin^2(\omega t + \varphi_0)} < 0$$

or, equivalently, iff

$$\cos^2(\omega t + \varphi_0) + E^2\sin^2(\omega t + \varphi_0) - (E^2-1)\frac{\omega}{\delta}\sin(\omega t + \varphi_0)\cos(\omega t + \varphi_0) > 0. \tag{20}$$

Now if $(E^2-1)\omega/\delta \leq -2E$ (that choosing a system of coordinates in which $\varphi = \pi/4$ is equivalent to $-\mathcal{K}/(i\rho_B) \leq 1$; see Appendix B), then the condition (20) is satisfied. Hence from (19) we can conclude that $\dot{\rho}(t) < 0$ for each $t \in Dom(\gamma)$ and $(\mathcal{P})$ is true (see Example 5 below).

Case 2.2.B If $E \geq 1$ (which implies $\mathcal{K}/i \geq 0$), then $(\mathcal{P})$ is true iff condition (6) is satisfied (see Appendix B). Notice that in the case where $\mathcal{K} = 0$ we clearly have that $\rho_{RC} < 1$ and $(\mathcal{P})$ is true (see Examples 2 and 3 below). The case in which $\rho_{RC} = 1$ is the case in which we have at least stability (but not asymptotic) for every switching function.

Case 3. If $\mathcal{D} = 0$, then the two straight lines coincide. If $\chi < 0$, it is easy to understand that we are in the same situation as that of Case 1. If $\chi > 0$, then to every $x \in Q^{-1}(0)$ there exists $u \in [0,1]$ such that $uAx+(1-u)Bx = 0$. In this case, $(\mathcal{P})$ is false, but we have at least stability (not asymptotic). This proves case **(RC.3)** of Theorem 2.3.

This concludes the proof of cases **(RC)**.

Examples. In the following, we give some examples of the various situations in the **(RC)** case. We refer to Figure 3.2.

Example 1. $\rho_A/i = 1.11$, $\rho_B = 0.045$, $\mathcal{K}/i = 0.095$. In this case, $\chi \sim -0.15$, $\mathcal{D} \sim 0.2$, and $(\mathcal{P})$ is true (cf. case **(RC.1)**).

Example 2. $\rho_A/i = 1.11$, $\rho_B = 0.02$, $\mathcal{K}/i = 1.33$. In this case, $\chi \sim -1.49$, $\mathcal{D} \sim -1.62$, $\rho_{RC} \sim 1.4$, and $(\mathcal{P})$ is false (cf. case **(RC.2.2.B)**).

Example 3. $\rho_A/i = 1.11$, $\rho_B = 0.03$, $\mathcal{K}/i = 0.75$. In this case, $\chi \sim -0.85$, $\mathcal{D} \sim -0.37$, $\rho_{RC} \sim 0.98$, and $(\mathcal{P})$ is true (cf. case **(RC.2.2.B)**).

Example 4. $\rho_A/i = 1.11$, $\rho_B = 0.045$, $\mathcal{K}/i = -2.4$. In this case, $\chi \sim 2.6$, $\mathcal{D} \sim -5.3$, and $(\mathcal{P})$ is false (cf. case **(RC.2.1)**).

Example 5. $\rho_A/i = 1.14$, $\rho_B = 1.67$, $\mathcal{K}/i = -0.42$. In this case, $\chi \sim -1.19$, $\mathcal{D} \sim -1.06$, and $(\mathcal{P})$ is true (cf. case **(RC.2.2.A)**).

3.3. The case in which *A* and *B* have both real eigenvalues. Let λ_1, λ_2 $(\lambda_1, \lambda_2 < 0,\ |\lambda_2| > |\lambda_1|)$ be the eigenvalues of A and λ_3, λ_4 $(\lambda_3, \lambda_4 < 0,\ |\lambda_4| > |\lambda_3|)$ be the eigenvalues of B. Choose a system of coordinates such that

$$A = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \tag{21}$$

$$\begin{aligned} B = \begin{pmatrix} a & b \\ c & d \end{pmatrix} &:= R^{-1}(\pi/4) \begin{pmatrix} \lambda_3 & \alpha(\lambda_4-\lambda_3) \\ 0 & \lambda_4 \end{pmatrix} R(\pi/4) \\ &= \frac{1}{2} \begin{pmatrix} (\lambda_3+\lambda_4)-\alpha(\lambda_4-\lambda_3) & (\lambda_3-\lambda_4)+\alpha(\lambda_4-\lambda_3) \\ (\lambda_3-\lambda_4)-\alpha(\lambda_4-\lambda_3) & (\lambda_3+\lambda_4)+\alpha(\lambda_4-\lambda_3) \end{pmatrix}, \end{aligned} \tag{22}$$

where $R(\varphi)$ is defined as in formula (14) and $\alpha \in \mathbf{R} \setminus \{\pm 1\}$. In this system of coordinates the eigenvectors of A are proportional to $\mathbf{V}_1 = (1,0)$, $\mathbf{V}_2 = (0,1)$ and the eigenvectors of B to $\mathbf{V}_3 = (1,1)$, $\mathbf{V}_4 = ((\alpha-1)/(\alpha+1), 1)$. The geometric meaning of α is the following. $\mathrm{Arctan}(\alpha)$ is the angle between the vector $(-1,1)$ and $\mathbf{V}_4$, measured clockwise. We have $\mathcal{K} = \alpha$. Notice that

$$[A,B] = -\frac{1}{2}(\lambda_1-\lambda_2)(\lambda_3-\lambda_4) \begin{pmatrix} 0 & (\alpha-1) \\ (\alpha+1) & 0 \end{pmatrix},$$

so $[A,B] \neq 0$ for every value of α. The case $\alpha = \pm 1$ is excluded (otherwise $\mathbf{V}_4$ is parallel to $\mathbf{V}_2$ or to $\mathbf{V}_1$, respectively, and **(H4)** fails).

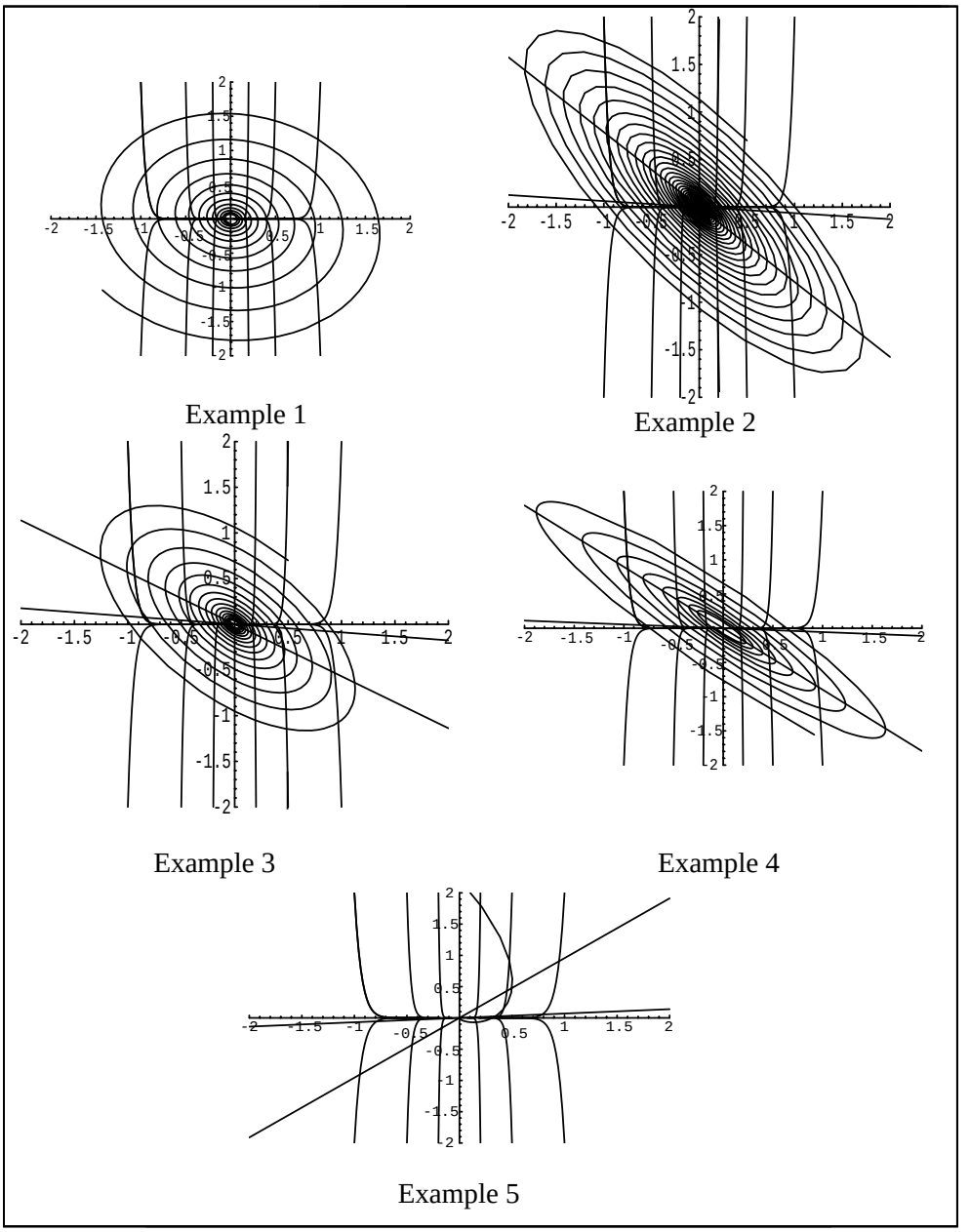


FIG. 3.2. *Examples in the* **(RC)** *case.*

The locus in which Ax and Bx are collinear is $Q^{-1}(0)$, where

$$Q = \det(Ax, Bx) = x_1^2(\lambda_1 c) + x_1 x_2 \bar{\chi} + x_2^2(-\lambda_2 b),$$

and by definition $\bar{\chi} := \lambda_1 d - \lambda_2 a = \frac{1}{2}((\lambda_1 - \lambda_2)(\lambda_3 + \lambda_4) - \mathcal{K}(\lambda_1 + \lambda_2)(\lambda_3 - \lambda_4)) = -\frac{1}{2}i(\lambda_1 - \lambda_2)(\lambda_3 - \lambda_4)\chi$, where $\chi := \rho_A \mathcal{K} - \rho_B \in i\mathbf{R}$. In this case, the discriminant

of the quadratic form Q is

$$D_{RR} = \bar{\chi}^2 + 4\lambda_1\lambda_2 bc = \bar{\chi}^2 + \lambda_1\lambda_2(\lambda_3-\lambda_4)^2(-\mathcal{K}^2+1) = \frac{1}{4}(\lambda_1-\lambda_2)^2(\lambda_3-\lambda_4)^2\mathcal{D}. \tag{23}$$

Notice that if $\mathcal{K} < 1$, then $\mathcal{D} > 0$. The following lemma states that in the case where $|\mathcal{K}| < 1$ $(\mathcal{P})$ is true.

LEMMA 3.1. *Let A, B be two 2×2 real matrices satisfying* **H1**, **H2**, **H3**, *and* **H4** *and such that their eigenvalues are real. Fix an arbitrary measurable switching function $u(.) : [0,\infty[\to[0,1]$, and let $(x_1(t), x_2(t))$ (resp., $(\rho(t),\theta(t))$) be the Cartesian (resp., polar) coordinates of the solution of $\dot{x}(t) = u(t)Ax(t) + (1-u(t))Bx(t)$, $x(0) = x_0 \in \mathbf{R}^2\setminus\{0\}$. If $\mathcal{K}\in]-1,1[$, we have that $\dot{\rho}(t) < 0$ for almost every $t\in[0,+\infty[$.*

Proof. In this case, it is possible to choose a system of coordinates such that

$$A = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix},$$
$$B = R^{-1}(\varphi)\begin{pmatrix} \lambda_3 & 0 \\ 0 & \lambda_4 \end{pmatrix} R(\varphi) = \begin{pmatrix} \cos^2(\varphi)\lambda_3 + \sin^2(\varphi)\lambda_4 & (\lambda_3-\lambda_4)\sin(\varphi)\cos(\varphi) \\ (\lambda_3-\lambda_4)\sin(\varphi)\cos(\varphi)(\lambda_3-\lambda_4) & \sin^2(\varphi)\lambda_3 + \cos^2(\varphi)\lambda_4 \end{pmatrix},$$

where we assume $\varphi\in]0,\pi/2[$. Notice that $\varphi = 0$ is excluded (otherwise $[A,B]=0$). We have

$$\begin{aligned}\dot{\rho}(t) &= \dot{x}_1(t)\cos\theta(t) + \dot{x}_2(t)\sin\theta(t)\\ &= \rho(t)\big(u(t)(\lambda_1\cos^2\theta(t) + \lambda_2\sin^2\theta(t))\\ &\quad + (1-u(t))(\lambda_3\cos^2(\theta(t)-\varphi) + \lambda_4\sin^2(\theta(t)-\varphi)\big).\end{aligned}$$

This means that $\rho(t)$ has the expression

$$\rho(t) = \rho(0)\exp\left(\int_0^t (u(t)f_1(t) + (1-u(t))f_2(t))\,dt\right),$$

where f_1 and f_2 are analytic functions satisfying $f_1 < \lambda_1$, $f_2 < \lambda_3$. ∎

If $|\mathcal{K}| > 1$ we have the following cases.

Case 1. If $\mathcal{D} < 0$, then $(\mathcal{P})$ is true.

Case 2. If $\mathcal{D} > 0$, then $Q^{-1}(0)$ is a couple of noncoinciding straight lines passing from the origin and forming the following angles with the x_1 axis:

$$\theta^{\pm} = \arctan(m^{\pm}),\quad m^{\pm} := \frac{-\bar{\chi}\pm\sqrt{D_{RR}}}{2(-\lambda_2 b)} = \frac{-\chi/i\pm\sqrt{\mathcal{D}}}{(\rho_A/i+1)(1-\mathcal{K})}. \tag{24}$$

From (23) it follows that $D_{RR} < \bar{\chi}^2$ so that in this case we have $\chi\neq 0$ and we may assume

$$\begin{cases} \theta^-,\theta^+\in]0,\pi/2[\text{ if } \chi/i \text{ and } \mathcal{K} \text{ have the same sign,}\\ \theta^-,\theta^+\in]-\pi/2,0[\text{ if } \chi/i \text{ and } \mathcal{K} \text{ have opposite sign.}\end{cases}$$

We have the following lemma.

LEMMA 3.2. *Let $\mathcal{D} > 0$; then*

(a) $\mathcal{K} > -\rho_A\rho_B$;
(b) *at each point of* $Q^{-1}(0)\setminus\{0\}$, *Ax and Bx have the same (resp., opposite) sign iff* $\mathcal{K} < -\rho_A\rho_B$ *(resp.,* $\mathcal{K} > -\rho_A\rho_B$*).*

Proof. (a) can be checked directly. Let us prove (b). Define $\Lambda^{\pm} := \mathcal{K} \pm \sqrt{\mathcal{D}} - \rho_A\rho_B$. By direct computation it follows that

- $\Lambda^{\pm} > 0$ iff $\mathcal{K} > -\rho_A\rho_B$;
- $(\frac{Ax}{\|Ax\|}) = (\frac{Bx}{\|Bx\|})$ for every $x = (h, m^{\pm}h),\ h \in \mathbf{R}\setminus\{0\},$ iff $\Lambda^{\pm} < 0$;
- $(\frac{Ax}{\|Ax\|}) = -(\frac{Bx}{\|Bx\|})$ for every $x = (h, m^{\pm}h),\ h \in \mathbf{R}\setminus\{0\},$ iff $\Lambda^{\pm} > 0$.

This concludes the proof. □

From Lemma (3.2) we have the following cases (notice that $-\rho_A\rho_B > 1$).

Case 2.1 If $\mathcal{K} > -\rho_A\rho_B$, then $(\mathcal{P})$ is false.

Case 2.2 If $\mathcal{K} < -\rho_A\rho_B$, then:

Case 2.2.A If $\mathcal{K} > 1$, one can easily check that the worst trajectory cannot rotate around the origin and $(\mathcal{P})$ is true.

Case 2.2.B If $\mathcal{K} < -1$, then the worst trajectory rotates around the origin and $(\mathcal{P})$ is true iff condition (7) is satisfied. Condition (7) can be obtained with arguments entirely similar to the ones of Appendices A and B. The case in which $\rho_{RC} = 1$ is the case in which we have at least stability (not asymptotic) for every switching function.

Case 3. If $\mathcal{D} = 0$, then the two straight lines coincide. Similarly to the **(CC)** and **(RC)** cases, if $\mathcal{K} < -\rho_A\rho_B$, then $(\mathcal{P})$ is true. Vice versa, if $\mathcal{K} > -\rho_A\rho_B$, then $(\mathcal{P})$ is false but we have stability (not asymptotic).

4. Asymptotic stability in the space of parameters. Fix a value of the cross ratio, and let $\mathcal{R}$ (resp., $\bar{\mathcal{R}}$) be the region in the (ρ_A, ρ_B) plane in which the system is asymptotically stable (resp., asymptotically stable or only stable) for arbitrary switching functions. In this section, we study the shape and the convexity of $\mathcal{R}$ and $\bar{\mathcal{R}}$.

4.1. The complex-complex case. In Figure 4.1 we show $\mathcal{R}$ for a fixed value of $\mathcal{K}$ in the case in which both A and B have complex eigenvalues.

In the case $\mathcal{K} < -1$, $\mathcal{R}$ is determined by the condition $\mathcal{D} < 0$. The set of values of ρ_A and ρ_B such that $\mathcal{D} = 0$ is the two curved lines of equations $\rho_B = \rho_A\mathcal{K} \pm \sqrt{(\rho_A^2+1)(\mathcal{K}^2-1)}$ of Figure 4.1 (case $\mathcal{K} < -1$). The points of intersection with the two axes are

$$(\rho_A, \rho_B) = (\sqrt{\mathcal{K}^2-1}, 0), \tag{25}$$

$$(\rho_A, \rho_B) = (0, \sqrt{\mathcal{K}^2-1}). \tag{26}$$

In this case, $\mathcal{R}$ is constituted by two connected open convex unbounded regions, while to get $\bar{\mathcal{R}}$ we have to add the points in which $\mathcal{D} = 0$.

In the case where $\mathcal{K} > 1$, $\mathcal{R}$ is determined by the condition $\rho_{CC} < 1$. In Figure 4.1 (case $\mathcal{K} > 1$) the locus $\mathcal{D} = 0$ is drawn with dotted lines, while the locus $\rho_{CC} = 1$ is drawn with a solid line. The points in which the two loci intersect each other and intersect the two axes are given again by formulas (25) and (26). In this case, to study the convexity of $\mathcal{R}$, we have to check if, expressing the locus $\rho_{CC} = 0$ as $\rho_B = f_{\mathcal{K}}(\rho_A)$, we find a convex function. In the following, the label $(_{\mathcal{K}})$ is a parameter, and it will be dropped. Let us indicate the derivative of ρ_{CC} with respect to the first and second

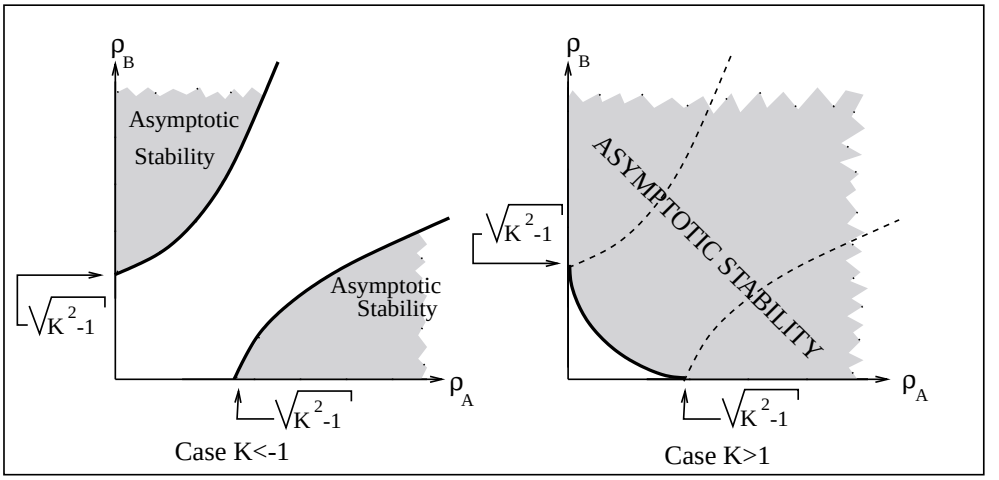


FIG. 4.1. *$\mathcal{R}$ (the grey region) for a fixed value of $\mathcal{K}$, in the complex–complex case.*

variable as $(\rho_{CC})_1$ and $(\rho_{CC})_2$. We have that

$$f'(\rho_A) = F\left(\rho_A, f(\rho_A)\right), \text{ where } F(\rho_A,\rho_B) := -\frac{(\rho_{CC})_1}{(\rho_{CC})_2} = -\frac{\arctan(\frac{\rho_B-\rho_A\mathcal{K}}{\sqrt{\mathcal{D}}})+\frac{\pi}{2}}{\arctan(\frac{\rho_A-\rho_B\mathcal{K}}{\sqrt{\mathcal{D}}})+\frac{\pi}{2}},$$

$$f''(\rho_A) = G\left(\rho_A, f(\rho_A)\right), \text{ where}$$

$$\begin{aligned}
G(\rho_A,\rho_B) &= \frac{\partial F(\rho_A,\rho_B)}{\partial\rho_A} + \frac{\partial F(\rho_A,\rho_B)}{\partial\rho_B}F(\rho_A,\rho_B)\\
&= \frac{2}{(1+\rho_A^2)\,(1+\rho_B^2)\,\sqrt{\mathcal{D}}\left(\pi+2\,\arctan(\frac{\rho_A-\rho_B\mathcal{K}}{\sqrt{\mathcal{D}}})\right)^3}\\
&\times \Bigg(\left[\rho_A^3\,\rho_B+\rho_A\,\rho_B^3+2(1+\rho_A^2+\rho_B^2+\rho_A\rho_B+\rho_A^2\rho_B^2)+\mathcal{K}(2+\rho_A^2+\rho_B^2)\right]\pi^2\\
&\quad+4\left(1+\rho_B^2\right)\left(1+\rho_A^2+\rho_A\,\rho_B+\mathcal{K}\right)\pi\,\arctan\left(\frac{\rho_A-\rho_B\,\mathcal{K}}{\sqrt{\mathcal{D}}}\right)\\
&\quad+4\left(1+\rho_A^2\right)\left(1+\rho_B^2+\rho_B\,\rho_A+\mathcal{K}\right)\pi\,\arctan\left(\frac{\rho_B-\rho_A\,\mathcal{K}}{\sqrt{\mathcal{D}}}\right)\\
&\quad+4\left(1+\rho_A^2\right)\left(\rho_A\,\rho_B+\mathcal{K}\right)\arctan\left(\frac{\rho_B-\rho_A\,\mathcal{K}}{\sqrt{\mathcal{D}}}\right)^2\\
&\quad+4\left(1+\rho_B^2\right)\left(\rho_A\,\rho_B+\mathcal{K}\right)\arctan\left(\frac{\rho_A-\rho_B\,\mathcal{K}}{\sqrt{\mathcal{D}}}\right)^2\\
&\quad+8\left(1+\rho_A^2\right)\left(1+\rho_B^2\right)\arctan\left(\frac{\rho_B-\rho_A\,\mathcal{K}}{\sqrt{\mathcal{D}}}\right)\arctan\left(\frac{\rho_A-\rho_B\,\mathcal{K}}{\sqrt{\mathcal{D}}}\right)\Bigg).
\end{aligned}$$

Now the only terms that can be negative are the ones in the third and fourth rows, but it is easy to check numerically that the sum of these two terms with the one in the second row is always bigger than zero. The convexity follows. In this case, $\mathcal{R}$ is a convex open unbounded region, while $\bar{\mathcal{R}}$ is a convex not-open unbounded region (we

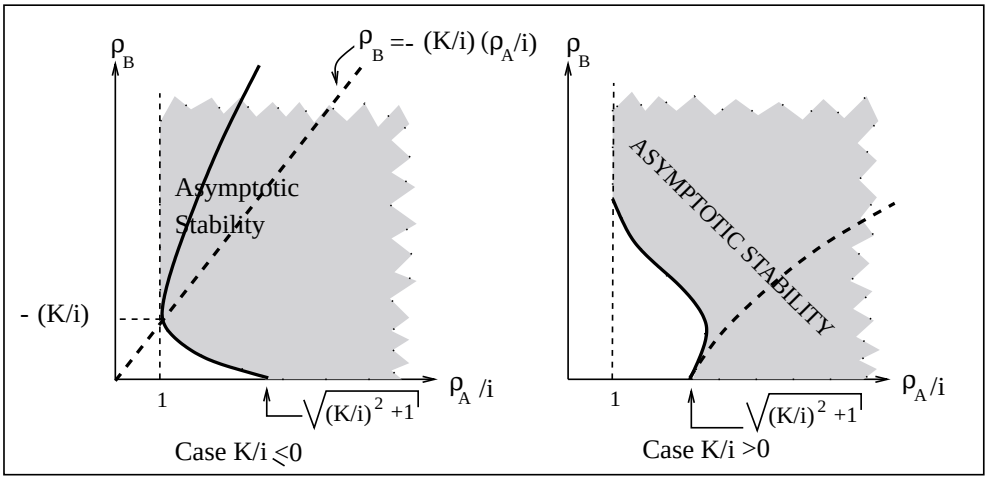


FIG. 4.2. $\mathcal{R}$ *(the grey region) for a fixed value of* $\mathcal{K}$*, in the* **(RC)** *case.*

have to add the points such that $\rho_{CC} = 1$).

4.2. The real-complex case. In the case in which A and B have one complex and the other real eigenvalues, $\mathcal{R}$ is drawn in Figure 4.2. We recall that $\rho_A/i > 1$, $\rho_B > 0$, $\mathcal{K}/i \in \mathbf{R}$.

In the case where $\chi > 0$ (which implies $\mathcal{K}/i < 0$ and $\rho_B < (-\mathcal{K}/i)(\rho_A/i)$), $\mathcal{R}$ is determined by the condition $\mathcal{D} > 0$. The locus $\mathcal{D} = 0$ is the set of points such that $\rho_B = -(\rho_A/i)(\mathcal{K}/i) \pm \sqrt{(-(\rho_A/i)^2 + 1)(-(\mathcal{K}/i)^2 - 1)}$. The intersection point with the ρ_A axis is

$$(\rho_A/i, \rho_B) = (\sqrt{(\mathcal{K}/i)^2 + 1}, 0), \tag{27}$$

and the intersection with the $\rho_A/i = 1$ set is

$$(\rho_A/i, \rho_B) = (1, -(\mathcal{K}/i)).$$

In the case when $\chi < 0$ and $\mathcal{K}/i \leq 0$, we have asymptotic stability. We conclude that in the case when $\mathcal{K}/i \leq 0$, $\mathcal{R}$ is a convex open unbounded region (see Figure 4.2 (case $\mathcal{K}/i \leq 0$)), while to get $\bar{\mathcal{R}}$, we have to add the points in which $\mathcal{D} = 0$.

In the case when $\chi < 0$ and $\mathcal{K}/i > 0$, $\mathcal{R}$ is determined by the condition $\rho_{RC} < 1$. In Figure 4.2 (case $\mathcal{K}/i > 0$), the locus $\mathcal{D} = 0$ is drawn with a dotted line, while the locus $\rho_{CC} = 1$ is drawn with a solid line. The points in which the two loci intersect each other are given by formula (27). In this case, $\mathcal{R}$ is a nonconvex open unbounded region. Again the points in which we have at least stability are the points in which we have asymptotic stability plus the points such that $\rho_{RC} = 1$.

4.3. The real-real case. In the case in which A and B have both real eigenvalues, $\mathcal{R}$ is drawn in Figure 4.3. We recall that $\rho_A/i, \rho_B/i > 1$, $\mathcal{K} \in \mathbf{R} \setminus \{\pm 1\}$. If $\mathcal{K} < -1$, $\mathcal{R}$ is determined by $\rho_{RR} > 0$, while, if $\mathcal{K} > 1$, $\mathcal{R}$ is determined by $\mathcal{D} > 0$. Similarly to the **(CC)** case, we can conclude that $\mathcal{R}$ is a convex open unbounded region, while $\bar{\mathcal{R}}$ is a convex not-open unbounded region. (We have to add the points such that $\rho_{CC} = 1$ and $\mathcal{D} = 0$.)

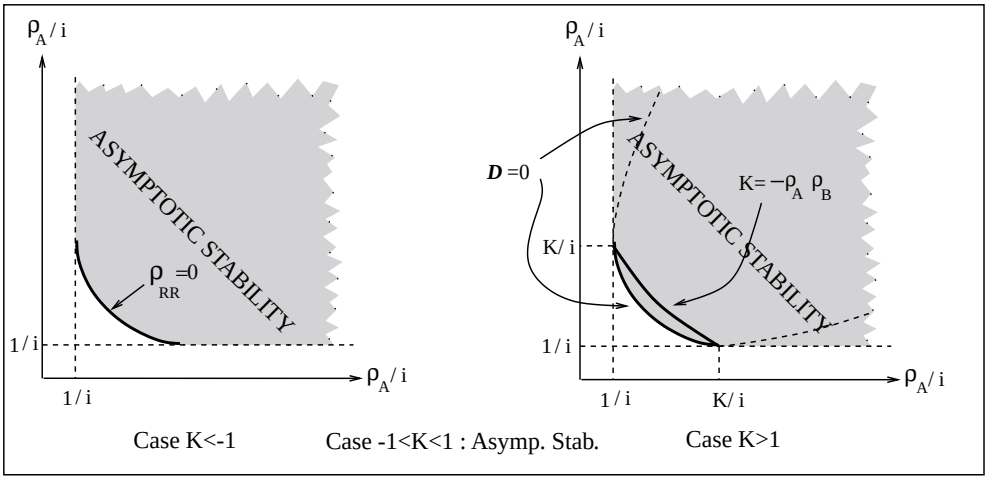


FIG. 4.3. $\mathcal{R}$ *(the grey region) for a fixed value of* $\mathcal{K}$*, in the* **(RR)** *case.*

5. Final remarks. Using the results of [1, 7] and by Theorem 2.3, we have a complete algorithm to study the asymptotic stability of a switched linear system in any dimension at least in the case in which

$$A_u = uA^1x + (1-u)A^2x, \qquad u \in [0,1], \quad A^1, A^2 \in \mathbf{R}^{n\times n},$$

where A and B are diagonalizable and $\dim\{A_1, A_2\}_{L.A.} \leq 4$. The case in which A or B is not diagonalizable can be treated with similar techniques.

Generalization can be done for more complex sets U. One is the following m-input system:

$$\dot{x} = \sum_{i=1}^{m} u_i A^i x, \quad \sum_{i=1}^{m} u_i = 1, \quad u_i \geq 0 \ \ (i = 1, \ldots, m), \quad x \in \mathbf{R}^n, \quad A^1, \ldots, A^m \in \mathbf{R}^{n\times n}.$$

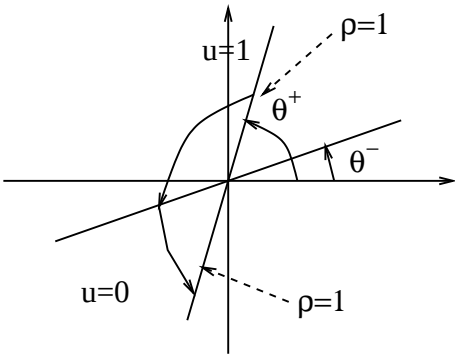
With exactly the same techniques used in this paper, one can find a coordinates invariant necessary and sufficient condition for the stabilizability of a control system of the kind (2), where all the matrices have eigenvalues with strictly *positive* real part. This problem was also studied in [12] but not in terms of a minimum number of coordinate-free parameters. We refer to [12] for details.

Some results can be obtained for the nonlinear version of the problem treated in this paper,

$$\dot{x} = uF(x) + (1-u)G(x), \tag{28}$$

where $x \in \mathbf{R}^2$, $F(.)$, $G(.)$ are $\mathcal{C}^\infty$ generic functions from $\mathbf{R}^2$ to $\mathbf{R}^2$ such that $F(0) = 0$, $G(0) = 0$, and the two dynamical systems $\dot{x} = F(x)$, $\dot{x} = G(x)$ are globally asymptotically stable at the origin. We are interested in studying under which conditions on $F(.)$ and $G(.)$ the origin of the system (28) is globally asymptotically stable for every measurable function $u(.) : [0, \infty[\to [0, 1]$.

Appendix A: Proof of formula (5). We refer to the following figure.



Let $\rho(t)$, $\theta(t)$ (resp., $x(t), y(t)$) be the polar coordinates (resp., Cartesian) of $\gamma^M(t)$, where we fix the initial condition by setting $\rho(0) = 1$, $\theta(0) = \theta^+$. We have to check if at the time a such that $\theta(a) = \theta(0) + 2\pi$ we have $\rho(a) < 1$. Due to the symmetries of the system, this happens iff at the time $\bar{t}$ such that $\theta(\bar{t}) = \theta^+ + \pi$ we have $\rho_{RC} := \rho(\bar{t}) < 1$. Notice that $\bar{t} = a/2$. The trajectory $\gamma^M(t)$ corresponds to the constant switching function $u = +1$ up to the time t' in which $\theta(t') = \theta^- + \pi$. This time is defined by the equations

$$
\begin{aligned}
&x(t') = \rho_0 e^{-\delta_A t'} \cos(\omega_A t' + \theta_E^+),\\
&y(t') = \rho_0 E e^{-\delta_A t'} \sin(\omega_A t' + \theta_E^+),\\
&\theta_E^+ = \arctan\left(\frac{m^+}{E}\right) \in \begin{cases} [-\pi/2, \pi/2[\text{ if } \theta^+ \in [-\pi/2, \pi/2[,\\ ]\pi/2, 3\pi/4[\text{ if } \theta^+ \in]\pi/2, 3\pi/4[, \end{cases}\\
&\rho_0 = (\cos^2(\theta_E^+) + E^2 \sin^2(\theta_E^+))^{-1/2},\\
&y(t') = m^- x(t').
\end{aligned}
$$

It follows that $\tan(\omega_A t' + \theta_E^+) = m^-/E$. If we set $\theta_E^- = \arctan(m^-/E) \in]\theta_E^+, \theta_E^+ + \pi[$, we have $t' = (\theta_E^- - \theta_E^+)/\omega_A$.

After time t', $\gamma^M(t)$ corresponds to the constant switching function $u = 0$ up to the first time $\bar{t}$ in which $\theta(\bar{t}) = \theta^+ + \pi$. This time is defined by the equations

$$
\begin{aligned}
&x(\bar{t}) = \rho(t') e^{-\delta_B(\bar{t} - t')} \cos(\omega_B(\bar{t} - t') + \theta^- + \pi),\\
&y(\bar{t}) = \rho(t') e^{-\delta_B(\bar{t} - t')} \sin(\omega_B(\bar{t} - t') + \theta^- + \pi),\\
&\rho(t') = \rho_0 e^{-\frac{\delta_A}{\omega_A}(\theta_E^- - \theta_E^+)} \sqrt{\cos^2(\theta_E^-) + E^2 \sin^2(\theta_E^-)},\\
&y(\bar{t}) = m^+ x(\bar{t}).
\end{aligned}
$$

It follows that $\tan(\omega_B(\bar{t} - t') + \theta^- + \pi) = \tan(\omega_B(\bar{t} - t') + \theta^-) = m^+ = \tan(\theta^+)$, and we have $\bar{t} = (\theta^+ - \theta^-)/\omega_B + t'$. Finally,

$$
\bar{\rho} = \rho(\bar{t}) = \rho(t') e^{-\frac{\delta_B}{\omega_B}(\bar{t} - t')} = e^{-\frac{\delta_A}{\omega_A}(\theta_E^- - \theta_E^+) - \frac{\delta_B}{\omega_B}(\theta^+ - \theta^-)} \sqrt{\frac{\cos^2\theta_E^- + E^2 \sin^2(\theta_E^-)}{\cos^2\theta_E^+ + E^2 \sin^2(\theta_E^+)}}.
$$

This formula is not in a good form because it is not explicitly invariant for the exchange of δ_A, ω_A with δ_B, ω_B and because the quantity E does not appear only in the form $E + 1/E$. Recalling the definition of $\rho_A, \rho_B, \mathcal{K}$ (see Definition 2.1) and using the equality

$$
\arctan a - \arctan b = \arctan\left(\frac{ab + 1}{b - a} + \pi/2\right),
$$

which holds for $a > b$, it is possible to obtain the relations

$$-\frac{\delta_A}{\omega_A}(\theta_E^- - \theta_E^+) = -\rho_A\left(\arctan\left(\frac{-\rho_A\mathcal{K}+\rho_B}{\sqrt{\mathcal{D}}}\right)+\pi/2\right),$$
$$-\frac{\delta_B}{\omega_B}(\theta^+ - \theta^-) = -\rho_B\left(\arctan\left(\frac{\rho_A-\rho_B\mathcal{K}}{\sqrt{\mathcal{D}}}\right)+\pi/2\right).$$

Moreover, with elementary computation we can show that

$$\sqrt{\frac{\cos^2\theta_E^- + E^2\sin^2(\theta_E^-)}{\cos^2\theta_E^+ + E^2\sin^2(\theta_E^+)}} = \sqrt{\frac{\rho_A\rho_B+\sqrt{\mathcal{D}}}{\rho_A\rho_B-\sqrt{\mathcal{D}}}}.$$

Formula (5) is obtained.

Appendix B: Proof of formula (6). To obtain a result that explicitly does not depend on the choice of the system of coordinates, we need to write the formulas of section 3.2 in a more invariant way. Set

$$\psi = \sqrt{\frac{E\cos^2\varphi + 1/E\sin^2\varphi}{E\sin^2\varphi + 1/E\cos^2\varphi}},$$

and make the coordinates transformation

$$x \to \Psi(\psi)x, \quad \text{where } \Psi(\psi) := \begin{pmatrix} 1 & 0 \\ 0 & \psi \end{pmatrix}.$$

In this case ($E \geq 1$), the new coordinates A, B, and $\theta^\pm$ have the expressions

$$A = \Psi^{-1}(\psi)\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}\Psi(\psi) = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix},$$
$$B = \Psi^{-1}(\psi)\begin{pmatrix} a & b \\ c & d \end{pmatrix}\Psi(\psi) = \begin{pmatrix} -\delta-\omega\mathcal{K}/i & -\omega\sqrt{-\mathcal{K}^2+1} \\ +\omega\sqrt{-\mathcal{K}^2+1} & -\delta+\omega\mathcal{K}/i \end{pmatrix},$$
$$\theta^\pm = \arctan m^\pm, \quad m^\pm = \frac{-\chi\pm\sqrt{-\mathcal{D}}}{2\frac{\lambda_2}{\lambda_1-\lambda_2}\sqrt{-\mathcal{K}^2+1}} = \frac{-\chi\pm\sqrt{-\mathcal{D}}}{(-\rho_A/i-1)\sqrt{-\mathcal{K}^2+1}}.$$

Equivalently, we can use the expressions (15), (16), (18) for $A, B, \theta^\pm$ with $E \geq 1$ and $\varphi = \pi/4$.

$$A = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix},$$
$$B = R^{-1}(\pi/4)\begin{pmatrix} -\delta & -\omega/E \\ \omega E & -\delta \end{pmatrix}R(\pi/4),$$
$$\theta^\pm = \arctan m^\pm, \quad m^\pm := \frac{-\chi\pm\sqrt{-\mathcal{D}}}{\frac{\lambda_2}{\lambda_1-\lambda_2}(E+1/E)} = 2\frac{-\chi\pm\sqrt{-\mathcal{D}}}{(-\rho_A/i-1)(E+1/E)}. \tag{29}$$

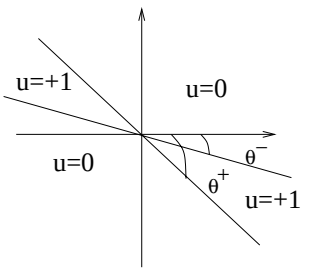
The relation between $\mathcal{K}$ and E is

$$\mathcal{K} = i\frac{1}{2}(E-1/E), \quad E = \mathcal{K}/i + \sqrt{-\mathcal{K}^2+1}.$$

Moreover, we are considering the case $\chi < 0$ so that $\theta^+, \theta^- \in]-\pi/2, 0[$. From (29) it follows that $\theta^+ < \theta^-$.

In this case, $\gamma^M(.)$ corresponds to the feedback (see the following figure):

$$u(x) = \begin{cases} 1 \text{ if } \theta \in]\theta^+, \theta^-[\text{ or } \theta \in]\theta^+ + \pi, \theta^- + \pi[, \\ 0 \text{ if } \theta \in]\theta^-, \theta^+ + \pi[\text{ or } \theta \in]\theta^- + \pi, \theta^+ + 2\pi[. \end{cases}$$



Make the following coordinates transformation: $x \to \bar{x} = R(\pi/4)x$. We have

$$\begin{aligned} A \to \bar{A} &= R(\pi/4) \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} R^{-1}(\pi/4), \\ B \to \bar{B} &= \begin{pmatrix} -\delta & -\omega/E \\ \omega E & -\delta \end{pmatrix}, \\ \theta^\pm \to \bar{\theta}^\pm &= \theta^\pm - \pi/4 = \arctan \bar{m}^\pm \in [3/4\pi, \pi/4[, \quad \bar{m}^\pm := \frac{m^\pm - 1}{m^\pm + 1}. \end{aligned}$$

Similarly to Appendix A, we compute γ^M in polar coordinates with the initial condition $\rho(0) = 1,\ \theta(0) = \bar{\theta}^-$. Let t' be the first time such that $\theta(t') = \bar{\theta}^+ + \pi$. We have

$$\begin{aligned} t' &= (\xi^+ - \xi^-)/\omega, \\ \rho(t') &= e^{-\frac{\delta}{\omega}(\xi^+ - \xi^-)} \sqrt{\frac{\cos^2 \xi^+ + E^2 \sin^2 \xi^+}{\cos^2 \xi^- + E^2 \sin^2 \xi^-}}, \end{aligned}$$

where $\xi^\pm := \arctan(\bar{m}^\pm/E), \quad \xi^+ \in]\xi^-, \xi^- + \pi[$.

Now we come back to the old coordinates ($\bar{x} \to x = R^{-1}(\pi/4)\bar{x}$), and we integrate Bx up to the first time $\bar{t}$ such that $\theta(\bar{t}) = \theta^- + \pi$. We have

$$\begin{aligned} x(\bar{t}) &= \rho(t') \cos(\theta^+ + \pi) e^{\lambda_1(\bar{t} - t')}, \\ y(\bar{t}) &= \rho(t') \sin(\theta^+ + \pi) e^{\lambda_2(\bar{t} - t')}, \\ y(\bar{t}) &= m^- x(\bar{t}). \end{aligned}$$

It follows that

$$m^+ e^{((\lambda_2 - \lambda_1)(\bar{t} - t'))} = n^- \Longrightarrow \bar{t} - t' = \frac{1}{\lambda_2 - \lambda_1} \ln\left(\frac{m^-}{m^+}\right).$$

Finally,

$$\begin{aligned} \rho_{RC} := \rho(\bar{t}) &= \rho(t') \sqrt{\cos^2 \theta^+ e^{\frac{\lambda_1}{\lambda_2 - \lambda_1} \ln(m^-/m^+)} + \sin^2 \theta^+ e^{\frac{\lambda_2}{\lambda_2 - \lambda_1} \ln(m^-/m^+)}} \\ &= e^{-\frac{\delta}{\omega}(\xi^+ - \xi^-)} \sqrt{\frac{\cos^2 \xi^+ + E^2 \sin^2 \xi^+}{\cos^2 \xi^- + E^2 \sin^2 \xi^-}} \end{aligned}$$

$$\times\sqrt{\cos^2\theta^+\left(\frac{m^-}{m^+}\right)^{\frac{\lambda_1}{\lambda_2-\lambda_1}}+\sin^2\theta^+\left(\frac{m^-}{m^+}\right)^{\frac{\lambda_2}{\lambda_2-\lambda_1}}}$$

$$= e^{-\rho_B(\xi^+-\xi^-)}\sqrt{\frac{\cos^2\xi^+ + E^2\sin^2\xi^+}{\cos^2\xi^- + E^2\sin^2\xi^-}}$$

$$\times\sqrt{\cos^2\theta^+\left(\frac{m^+}{m^-}\right)^{\frac{1}{2}(-\rho_A/i+1)}+\sin^2\theta^+\left(\frac{m^+}{m^-}\right)^{\frac{1}{2}(-\rho_A/i-1)}},$$

which is formula (6). This formula is complicated but acceptable because there are no further symmetries.

Acknowledgment. The author is grateful to Andrei Agrachev for suggesting the problem and for helpful discussions that contributed to finding the right invariant.

REFERENCES

[1] A. A. Agrachev and D. Liberzon, *Lie-algebraic stability criteria for switched systems*, SIAM J. Control Optim., 40 (2001), pp. 253–269.

[2] D. Angeli, *A note on stability of arbitrarily switched homogeneous systems*, Systems Control Lett., to appear.

[3] U. Boscain, *Extremal Synthesis and Morse Property for Minimum Time*, Ph.D. Thesis, SISSA Trieste, Italy, 2000.

[4] U. Boscain and B. Piccoli, *Extremal syntheses for generic planar systems*, J. Dynam. Control Systems, 7 (2001), pp. 209–258.

[5] U. Boscain and B. Piccoli, *Morse properties for the minimum time function on 2-D manifolds*, J. Dynam. Control Systems, 7 (2001), pp. 385–423.

[6] W. P. Dayawansa and C. F. Martin, *A converse Lyapunov theorem for a class of dynamical systems which undergo switching*, IEEE Trans. Automat. Control, 44 (1999), pp. 751–760.

[7] D. Liberzon, J. P. Hespanha, and A. S. Morse, *Stability of switched systems: A Lie-algebraic condition*, Systems Control Lett., 37 (1999), pp. 117–122.

[8] D. Liberzon and A. S. Morse, *Basic problems in stability and design of switched systems*, IEEE Control Systems Magazine, 19 (1999), pp. 59–70.

[9] B. Piccoli, *Classification of generic singularities for the planar time-optimal synthesis*, SIAM J. Control Optim., 34 (1996), pp. 1914–1946.

[10] H. J. Sussmann, *The structure of time-optimal trajectories for single-input systems in the plane: The general real-analytic case*, SIAM J. Control Optim., 25 (1987), pp. 868–904.

[11] R. Shorten and K. Narendra, *Necessary and sufficient conditions for the existence of a common quadratic Lyapunov function for two stable second order linear time–invariant systems*, in Proceedings of the American Control Conference, San Diego, 1999, pp. 1410–1414.

[12] X. Xu and P. J. Antsaklis, *Stabilization of second order LTI switched systems*, in Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, AZ, 1999, pp. 1339–1344.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 1, pp. 113–140

DISCRETE-TIME AND SAMPLED-DATA LOW-GAIN CONTROL OF INFINITE-DIMENSIONAL LINEAR SYSTEMS IN THE PRESENCE OF INPUT HYSTERESIS*

H. LOGEMANN† AND A. D. MAWBY†‡

Abstract. We introduce a general class of causal dynamic discrete-time nonlinearities which have certain monotonicity and Lipschitz continuity properties. In particular, the discretizations of a large class of continuous-time hysteresis operators obtained by applying the standard sampling and hold operations belong to this class. It is shown that closing the loop around a power-stable, linear, infinite-dimensional, discrete-time, single-input, single-output system, subject to an input nonlinearity from the class under consideration and compensated by a discrete-time integral controller, guarantees asymptotic tracking of constant reference signals, provided that (a) the positive integrator gain is sufficiently small and (b) the reference value is feasible in a very natural sense. We apply this result in the development of sampled-data low-gain integral control for exponentially stable, regular, linear, infinite-dimensional, continuous-time systems subject to input hysteresis.

Key words. continuous-time regular infinite-dimensional systems, discrete-time infinite-dimensional systems, hysteresis nonlinearities, integral control, robust tracking, sampled-data control

AMS subject classifications. 47H30, 47J40, 93C10, 93C20, 93C25, 93C55, 93C57, 93D10

PII. S0363012901385770

1. Introduction.

The present paper extends the line of work on low-gain integral control of infinite-dimensional linear systems subject to input nonlinearities initiated by the recent papers [9], [11], [14], [15]. Underpinning these contributions are generalizations of the well-known principle (see, for example, [4], [17], [18], and [22]) that closing the loop around a stable, linear, continuous-time, single-input, single-output plant, with transfer function $\mathbf{G}^c(s)$ compensated by a pure integral controller k/s, will result in a stable closed-loop system that achieves asymptotic tracking of arbitrary constant reference signals, provided that $|k|$ is sufficiently small and $\mathbf{G}^c(0)k > 0$.[1] In particular, Logemann and Mawby [11] have shown that the above principle remains true for exponentially stable, regular, linear, infinite-dimensional, continuous-time, single-input, single-output systems subject to input hysteresis belonging to a certain class $\mathcal{C}(\lambda)$ of hysteresis operators, provided the reference value r is feasible in a natural sense. The class $\mathcal{C}(\lambda)$ consists of hysteresis operators which, among other conditions, satisfy a certain Lipschitz condition with Lipschitz constant $\lambda > 0$. We emphasize that $\mathcal{C}(\lambda)$ encompasses a large number of hysteresis nonlinearities important in applications such as relay, backlash, elastic-plastic, and Prandtl hysteresis.

In this paper, we provide discrete-time and sampled-data analogues of the continuous-time results in [11]. More precisely, the contribution of the present work is twofold.

*Received by the editors March 2, 2001; accepted for publication (in revised form) October 8, 2001; published electronically April 2, 2002. This work was supported in part by the UK EPSRC Council (grant GR/L78086).

http://www.siam.org/journals/sicon/41-1/38577.html

†Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK (hl@maths.bath.ac.uk).

‡Current address: Systems Engineering and Assessment, P.O. Box 800, Bath BA3 6TB, UK (adm@sea.co.uk).

[1] Therefore, under the above assumptions on the plant, the problem of tracking constant reference signals reduces to that of tuning the gain parameter k. This so-called tuning regulator theory [4] has been successfully applied in process control (see [3], [19]).

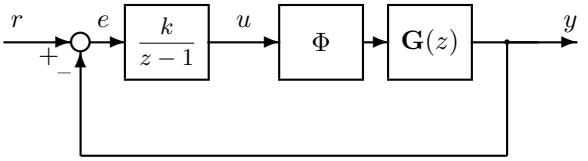


FIG. 1. *Low-gain control with input nonlinearity.*

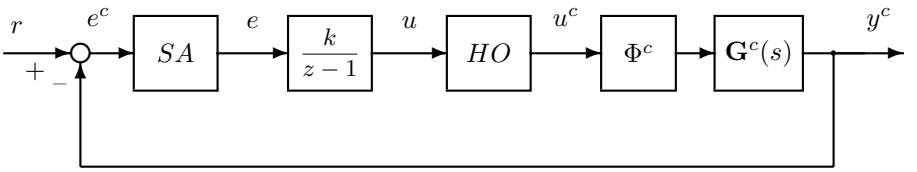


FIG. 2. *Sampled-data low-gain control.*

(i) We introduce a class $\mathcal{D}(\lambda)$ of causal dynamic discrete-time nonlinearities which have a number of properties, one of them being that a functional Lipschitz condition with Lipschitz constant $\lambda > 0$ is satisfied. The class $\mathcal{D}(\lambda)$ contains a large number of discrete-time hysteresis operators, in particular discretizations of continuous-time hysteresis operators in $\mathcal{C}(\lambda)$ obtained by sampling/hold. We derive a discrete-time version of the continuous-time tuning regulator result in [11] by showing that for a power-stable, linear, infinite-dimensional, discrete-time, single-input, single-output plant with transfer function $\mathbf{G}(z)$, subject to a dynamic input nonlinearity $\Phi \in \mathcal{D}(\lambda)$, the output $y(n)$ of the closed-loop system, shown in Figure 1, converges to the reference value r as $n \to \infty$, provided that $\mathbf{G}(1) > 0$, r is feasible in some natural sense, and $k \in (0, K/\lambda)$, where K is the supremum of the set of all numbers $k > 0$ such that

$$1 + k \operatorname{Re} \frac{\mathbf{G}(z)}{z-1} \geq 0 \quad \forall\, |z| > 1\,.$$

(ii) We apply the discrete-time theory in the development of a sampled-data counterpart to the continuous-time low-gain control result in [11]—see Figure 2—where HO denotes a standard hold operation and SA denotes a sampling operation. In the case of unbounded observation, the latter involves an averaging operation. Specifically, we show that for an exponentially stable, regular, linear, infinite-dimensional, continuous-time, single-input, single-output plant with transfer function $\mathbf{G}^c(s)$, subject to a continuous-time dynamic input nonlinearity $\Phi^c \in \mathcal{C}(\lambda)$, the output $y^c(t)$ of the closed-loop system, shown in Figure 2, converges to the reference value r as $t \to \infty$, provided that $\mathbf{G}^c(0) > 0$, r is feasible in some natural sense, and $k > 0$ is sufficiently small. The class of regular, linear, infinite-dimensional, continuous-time systems, introduced by Weiss [25], [26], [27], [28], is rather general. It includes most distributed parameter systems and all time-delay systems (retarded and neutral) which are of interest in applications. With respect to (i), while the structure of the discrete-time analysis parallels that of the continuous-time analysis in [11], there are several points where these analyses differ in an essential manner. With reference to (ii), the sampled-data results constitute the main contribution of the paper. In the derivation of these results, the discrete-time theory plays a central role.

We mention that there exists a substantial literature on the mathematical theory of hysteresis phenomena; see, for example, Brokate [1], Brokate and Sprekels [2], and Krasnosel'skiĭ and Pokrovskiĭ [8]. Of particular importance in a systems and control context is the pioneering work [8]. Our treatment of continuous-time hysteresis operators in section 4 has been strongly influenced by chapter 2 in [2].

The paper is organized as follows. In section 2, we introduce the class $\mathcal{D}(\lambda)$ of dynamic discrete-time nonlinearities, while section 3 contains the low-gain tuning regulator result for discrete-time systems subject to input nonlinearities belonging to $\mathcal{D}(\lambda)$. In section 4, we introduce the class $\mathcal{C}(\lambda)$ of dynamic continuous-time nonlinearities and establish several important properties enjoyed by nonlinearities in this class. At the end of this section, we discretize continuous-time hysteresis operators and show that the resultant discrete-time operators are contained in the class $\mathcal{D}(\lambda)$. In section 5, the low-gain tracking problem for exponentially stable, regular, linear, infinite-dimensional, continuous-time, single-input, single-output systems with input nonlinearity in $\mathcal{C}(\lambda)$ is solved using sampled-data integral control. In section 6, we illustrate our results in the context of a simple linear diffusion process subject to input hysteresis: relay as well as backlash nonlinearities are considered. Finally, several technical details relating to the infinite-dimensional discrete-time positive-real lemma have been relegated to the appendix.

Notation. We define

$$\mathbb{R}_+ := \{x \in \mathbb{R} \,|\, x \geq 0\}, \quad \mathbb{Z}_+ = \{x \in \mathbb{Z} \,|\, x \geq 0\}.$$

For sets M and N, we denote the set of all functions $f : M \to N$ by $F(M, N)$. If $I \subset \mathbb{R}$ is a compact interval, then $AC(I, \mathbb{R})$ denotes the space of absolutely continuous real-valued functions defined on I; $AC(\mathbb{R}_+, \mathbb{R})$ denotes the space of real-valued functions defined on $\mathbb{R}_+$ which are absolutely continuous on any compact interval $I \subset \mathbb{R}_+$, i.e., a function $f \in F(\mathbb{R}_+, \mathbb{R})$ is in $AC(\mathbb{R}_+, \mathbb{R})$ if and only if there exists a function $g \in L^1_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R})$ such that

$$f(t) = f(0) + \int_0^t g(\tau)\, d\tau \quad \forall\, t \geq 0.$$

We say that a function $f \in F(\mathbb{R}_+, \mathbb{R})$ is *piecewise monotone* if there exists a sequence $0 = t_0 < t_1 < t_2 < \cdots$ such that $\lim_{i\to\infty} t_i = \infty$ and f is monotone on each of the intervals (t_i, t_{i+1}). A function $f \in F(\mathbb{R}_+, \mathbb{R})$ is called *piecewise continuous* if there exists a sequence $0 = t_0 < t_1 < t_2 < \cdots$ such that $\lim_{i\to\infty} t_i = \infty$, f is continuous on each of the intervals (t_i, t_{i+1}), and the right and left limits of f exist and are finite at each t_i. We denote the space of all piecewise continuous functions $f : \mathbb{R}_+ \to \mathbb{R}$ by $PC(\mathbb{R}_+, \mathbb{R})$. As usual, for $f \in PC(\mathbb{R}_+, \mathbb{R})$, we define

$$f(t+) := \lim_{\tau \downarrow t} f(\tau) \text{ (for } t \geq 0) \quad \text{and} \quad f(t-) := \lim_{\tau \uparrow t} f(\tau) \text{ (for } t > 0).$$

Let $\mathbb{T} = \mathbb{R}_+, \mathbb{Z}_+$; a function $f \in F(\mathbb{T}, \mathbb{R})$ is called *ultimately constant* if there exists $T \in \mathbb{T}$ such that f is constant on $[T, \infty) \cap \mathbb{T}$. $L(X, Y)$ denotes the space of bounded linear operators from a Banach space X to a Banach space Y, and we set $L(X) := L(X, X)$. The Laplace transform is denoted by $\mathfrak{L}$.

2. A class of discrete-time nonlinear operators. For each $n \in \mathbb{Z}_+$, we define a projection operator $\mathbf{Q}_n : F(\mathbb{Z}_+, \mathbb{R}) \to F(\mathbb{Z}_+, \mathbb{R})$ by

$$(\mathbf{Q}_n u)(m) = \begin{cases} u(m) & \text{for } m \in [0, n] \cap \mathbb{Z}_+, \\ u(n) & \text{for } m \in \mathbb{Z}_+ \setminus [0, n]. \end{cases}$$

Recall that an operator $\Phi : F(\mathbb{Z}_+, \mathbb{R}) \to F(\mathbb{Z}_+, \mathbb{R})$ is called *causal* if for all $u, v \in F(\mathbb{Z}_+, \mathbb{R})$ and all $n \in \mathbb{Z}_+$ with $u(m) = v(m)$ for all $m \in [0, n] \cap \mathbb{Z}_+$ it follows that $(\Phi(u))(m) = (\Phi(v))(m)$ for all $m \in [0, n] \cap \mathbb{Z}_+$.

Let $u \in F(\mathbb{Z}_+, \mathbb{R})$. The function u is called *ultimately nondecreasing* if there exists $m \in \mathbb{Z}_+$ such that u is nondecreasing on $\mathbb{Z}_+ \setminus [0, m]$.

The *numerical value set* $\operatorname{NVS} \Phi$ of an operator $\Phi : F(\mathbb{Z}_+, \mathbb{R}) \to F(\mathbb{Z}_+, \mathbb{R})$ is defined by

$$\operatorname{NVS} \Phi := \{(\Phi(u))(n) \,|\, u \in F(\mathbb{Z}_+, \mathbb{R}),\, n \in \mathbb{Z}_+\}.$$

We introduce the following five assumptions on the operator $\Phi : F(\mathbb{Z}_+, \mathbb{R}) \to F(\mathbb{Z}_+, \mathbb{R})$:

(D1) Φ is causal;

(D2) for all $u \in F(\mathbb{Z}_+, \mathbb{R})$ and all $n \in \mathbb{Z}_+$

$$(\Phi(\mathbf{Q}_n u))(k) = (\Phi(\mathbf{Q}_n u))(n) \quad \forall\, k \in \mathbb{Z}_+ \setminus [0, n];$$

(D3) there exists $\lambda > 0$ such that for all $u \in F(\mathbb{Z}_+, \mathbb{R})$ and all $n \in \mathbb{Z}_+$

$$u(n) \neq u(n+1) \Longrightarrow \frac{(\Phi(u))(n+1) - (\Phi(u))(n)}{u(n+1) - u(n)} \in [0, \lambda];$$

(D4) if $u \in F(\mathbb{Z}_+, \mathbb{R})$ is ultimately nondecreasing and $\lim_{n\to\infty} u(n) = \infty$, then $(\Phi(u))(n)$ and $(\Phi(-u))(n)$ converge to $\sup \operatorname{NVS} \Phi$ and $\inf \operatorname{NVS} \Phi$, respectively, as $n \to \infty$;

(D5) if, for $u \in F(\mathbb{Z}_+, \mathbb{R})$, $L := \lim_{n\to\infty} (\Phi(u))(n)$ exists with $L \in \operatorname{int}(\operatorname{clos}(\operatorname{NVS} \Phi))$, then u is bounded.

REMARK 2.1. (1) *We note that if* (D1) *and* (D2) *hold, then* (D3) *is implied by the monotonicity condition*

$$[(\Phi(u))(n+1) - (\Phi(u))(n)][u(n+1) - u(n)] \geq 0 \quad \forall\, u \in F(\mathbb{Z}_+, \mathbb{R}),\ \ \forall\, n \in \mathbb{Z}_+,$$

together with the Lipschitz continuity condition

$$\sup_{n\in\mathbb{Z}_+} |(\Phi(u))(n) - (\Phi(v))(n)| \leq \lambda \sup_{n\in\mathbb{Z}_+} |u(n) - v(n)| \quad \forall\, u, v \in F(\mathbb{Z}_+, \mathbb{R}).$$

(2) *Assumption* (D2) *says that if the input u of the nonlinearity Φ is constant on $\mathbb{Z}_+ \setminus [0, n-1]$, then the output $\Phi(u)$ is constant and equal to $(\Phi(u))(n)$ on $\mathbb{Z}_+ \setminus [0, n-1]$.*

(3) *If* (D1)–(D3) *hold, then*

$$|(\Phi(u))(n+1) - (\Phi(u))(n)| \leq \lambda\, |u(n+1) - u(n)| \quad \forall\, u \in F(\mathbb{Z}_+, \mathbb{R}),\ \ \forall\, n \in \mathbb{Z}_+.$$

Thus if (D4) *also holds,* $\operatorname{clos}(\operatorname{NVS} \Phi)$ *is an interval. However, it can be shown that* $\operatorname{NVS} \Phi$ *is not necessarily an interval; see Mawby* [21] *for a counterexample.*

(4) *If* (D1)–(D3) *hold, then for all $u \in F(\mathbb{Z}_+, \mathbb{R})$ there exists $d : \mathbb{Z}_+ \to [0, \lambda]$ such that $(\Phi(u))(n+1) - (\Phi(u))(n) = d(n)(u(n+1) - u(n))$ for all $n \in \mathbb{Z}_+$.*

If $\Phi : F(\mathbb{Z}_+, \mathbb{R}) \to F(\mathbb{Z}_+, \mathbb{R})$ satisfies (D3), then any number $l > 0$ such that (D3) holds for $\lambda = l$ is called a *Lipschitz constant* of Φ. We are now in a position to define the class of nonlinear operators we will be considering in the context of the discrete-time integral control problem in section 3.

DEFINITION 2.2. *Let $\lambda > 0$. The set of all operators $\Phi : F(\mathbb{Z}_+, \mathbb{R}) \to F(\mathbb{Z}_+, \mathbb{R})$ satisfying* (D1)–(D5) *and having Lipschitz constant λ is denoted by $\mathcal{D}(\lambda)$.*

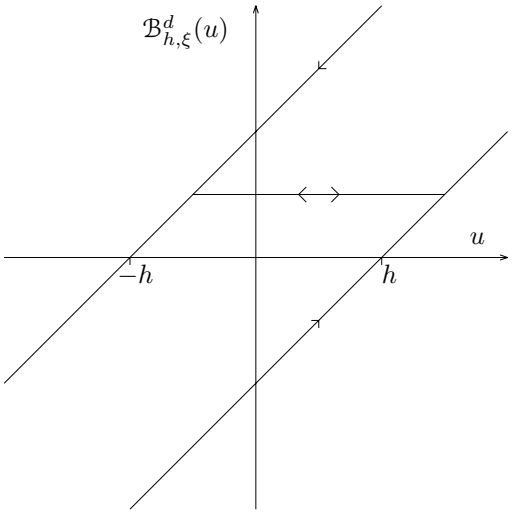


FIG. 3. *Backlash hysteresis.*

We now consider two examples of nonlinearities which satisfy (D1)–(D5).

Static nonlinearities. For a function $f : \mathbb{R} \to \mathbb{R}$, define the corresponding static nonlinearity by

$$\mathcal{S}_f : F(\mathbb{Z}_+, \mathbb{R}) \to F(\mathbb{Z}_+, \mathbb{R}), \quad u \mapsto f \circ u.$$

It is clear that if $f : \mathbb{R} \to \mathbb{R}$ is nondecreasing and globally Lipschitz with Lipschitz constant $\lambda > 0$, then $\mathcal{S}_f \in \mathcal{D}(\lambda)$.

Backlash hysteresis. Let $h \in \mathbb{R}_+$ be arbitrary. Define the function $b_h : \mathbb{R}^2 \to \mathbb{R}$ by

$$b_h(v, w) = \max\{v - h, \min\{v + h, w\}\}. \tag{2.1}$$

We note that

$$b_h(v, w) \in [v - h, v + h] \quad \forall\, v, w \in \mathbb{R}, \tag{2.2}$$

$$b_h(v, w) = w \quad \forall\, (v, w) \in \{(z_1, z_2) \,|\, z_1 \in \mathbb{R}, z_2 \in [z_1 - h, z_1 + h]\}, \tag{2.3}$$

$$(b_h(v_1, w) - b_h(v_2, w))(v_1 - v_2) \geq 0 \quad \forall\, v_1, v_2, w \in \mathbb{R}. \tag{2.4}$$

For each $\xi \in \mathbb{R}$, we introduce the discrete-time backlash operator $\mathcal{B}^d_{h,\,\xi} : F(\mathbb{Z}_+, \mathbb{R}) \to F(\mathbb{Z}_+, \mathbb{R})$ by defining recursively

$$(\mathcal{B}^d_{h,\,\xi}(u))(n) = \begin{cases} b_h(u(0), \xi) & \text{for } n = 0, \\ b_h(u(n), (\mathcal{B}^d_{h,\,\xi}(u))(n-1)) & \text{for } n \in \mathbb{Z}_+ \setminus \{0\}. \end{cases}$$

We remark that ξ plays the role of an "initial state." The discrete-time backlash operator $\mathcal{B}^d_{h,\,\xi}$ is illustrated in Figure 3.

We show that $\mathcal{B}^d_{h,\xi} \in \mathcal{D}(1)$. It is immediately clear from the definition that $\mathcal{B}^d_{h,\xi}$ satisfies (D1). Using (2.2) and (2.3), we see that (D2) holds. Combining (2.1)–(2.4) leads to

(2.5)
$$[(\mathcal{B}^d_{h,\xi}(u))(n+1) - (\mathcal{B}^d_{h,\xi}(u))(n)][u(n+1) - u(n)] \geq 0 \quad \forall\, u \in F(\mathbb{Z}_+, \mathbb{R}), \quad \forall\, n \in \mathbb{Z}_+.$$

It is not difficult to show (see [2, p. 42]) that

$$|b_h(v_1, w_1) - b_h(v_2, w_2)| \leq \max(|v_1 - v_2|, |w_1 - w_2|) \quad \forall\, v_1, v_2, w_1, w_2 \in \mathbb{R}.$$

Thus

$$|(\mathcal{B}^d_{h,\xi}(u))(n+1) - (\mathcal{B}^d_{h,\xi}(u))(n)| \leq |u(n+1) - u(n)| \quad \forall\, u \in F(\mathbb{Z}_+, \mathbb{R}), \quad \forall\, n \in \mathbb{Z}_+,$$

which, combined with (2.5), implies that (D3) holds for $\lambda = 1$. Note that $\operatorname{NVS} \mathcal{B}^d_{h,\xi} = \mathbb{R}$. By (2.2), for all $u \in F(\mathbb{Z}_+, \mathbb{R})$ and all $n \in \mathbb{Z}_+$, $(\mathcal{B}^d_{h,\xi}(u))(n) \in [u(n) - h, u(n) + h]$, showing that (D4) holds. Finally, it is clear that

$$v \in [b_h(v, w) - h, b_h(v, w) + h] \quad \forall\, v, w \in \mathbb{R},$$

and so

$$u(n) \in [(\mathcal{B}^d_{h,\xi}(u))(n) - h, (\mathcal{B}^d_{h,\xi}(u))(n) + h] \quad \forall\, u \in F(\mathbb{Z}_+, \mathbb{R}), \quad \forall\, n \in \mathbb{Z}_+,$$

showing that (D5) is satisfied. We have shown that (D1)–(D5) hold for $\mathcal{B}^d_{h,\xi}$ (with $\lambda = 1$), and hence $\mathcal{B}^d_{h,\xi} \in \mathcal{D}(1)$. We direct the reader to section 4 for a discussion of the backlash operator in a continuous-time setting.

3. Discrete-time integral control. Consider a single-input, single-output, discrete-time system

$$x(n+1) = Ax(n) + Bu(n), \quad x(0) = x_0 \in X, \tag{3.1a}$$
$$y(n) = Cx(n) + Du(n), \tag{3.1b}$$

evolving on a real Hilbert space X. Here $A \in L(X)$, $B \in L(\mathbb{R}, X)$, $C \in L(X, \mathbb{R})$, and $D \in \mathbb{R}$. A system of the form (3.1) is called *power-stable* if A is power-stable, i.e., there exist $M \geq 1$ and $\theta \in (0, 1)$ such that

$$\|A^n\| \leq M\theta^n \quad \forall\, n \in \mathbb{Z}_+.$$

The transfer function $\mathbf{G}$ of (3.1) is given by

$$\mathbf{G}(z) = C(zI - A)^{-1}B + D.$$

Suppose that system (3.1) is subject to a causal input nonlinearity $\Phi : F(\mathbb{Z}_+, \mathbb{R}) \to F(\mathbb{Z}_+, \mathbb{R})$, yielding the nonlinear system

$$x(n+1) = Ax(n) + B(\Phi(u))(n), \quad x(0) = x_0 \in X, \tag{3.2a}$$
$$y(n) = Cx(n) + D(\Phi(u))(n). \tag{3.2b}$$

Denoting the reference value by r, the control law

$$u(n+1) = u(n) + k(r - y(n)),$$

where k is a real parameter, then leads to the following nonlinear system of difference equations:

$$x(n+1) = Ax(n) + B(\Phi(u))(n), \quad x(0) = x_0 \in X, \tag{3.3a}$$

$$u(n+1) = u(n) + k(r - Cx(n) - D(\Phi(u))(n)), \quad u(0) = u_0 \in \mathbb{R}. \tag{3.3b}$$

If $\mathbf{G}$ is holomorphic and bounded on $\{z \in \mathbb{C} \,|\, |z| > \alpha\}$ for some $\alpha < 1$ (which is the case if (3.1) is power-stable) and $\mathbf{G}(1) > 0$, then it can be shown that

$$1 + k \operatorname{Re} \frac{\mathbf{G}(z)}{z-1} \geq 0 \quad \forall\, |z| > 1, \tag{3.4}$$

for all sufficiently small $k > 0$; see [16]. We define

$$K := \sup\{k > 0 \,|\, (3.4) \text{ holds}\}. \tag{3.5}$$

We can now state the main result of this section.

THEOREM 3.1. *Let $\lambda > 0$. Assume that $\Phi \in \mathcal{D}(\lambda)$, (3.1) is power-stable, $\mathbf{G}(1) > 0$, $k \in (0, K/\lambda)$, and $r \in \mathbb{R}$ is such that $\tilde{r} := r/\mathbf{G}(1) \in \operatorname{clos}(\operatorname{NVS}\Phi)$. Then for all $(x_0, u_0) \in X \times \mathbb{R}$, the solution (x, u) of (3.3) satisfies the following:*

(1) $\lim_{n\to\infty}(\Phi(u))(n) = \tilde{r}$;

(2) $\lim_{n\to\infty} x(n) = (I - A)^{-1}B\tilde{r}$;

(3) $\lim_{n\to\infty} y(n) = r$, *where* $y(n) = Cx(n) + D(\Phi(u))(n)$;

(4) *if* $\tilde{r} \in \operatorname{int}(\operatorname{clos}(\operatorname{NVS}\Phi))$, *then u is bounded.*

Proof. Denote the solution of (3.3) by (x, u), and introduce new variables by defining

$$z(n) := x(n) - (I - A)^{-1}B(\Phi(u))(n), \quad v(n) := (\Phi(u))(n) - \tilde{r} \quad \forall\, n \in \mathbb{Z}_+.$$

By Remark 2.1, part (4), there exists $d : \mathbb{Z}_+ \to [0, \lambda]$ such that $(\Phi(u))(n+1) - (\Phi(u))(n) = d(n)(u(n+1) - u(n))$ for all $n \in \mathbb{Z}_+$. Using the identity $A(I - A)^{-1} = (I - A)^{-1} - I$, a straightforward calculation yields

$$z(n+1) = Az(n) - (I - A)^{-1}Bw(n), \quad z(0) = z_0, \tag{3.6a}$$

$$v(n+1) = v(n) + w(n), \quad v(0) = v_0, \tag{3.6b}$$

where

$$w(n) = -kd(n)(Cz(n) + \mathbf{G}(1)v(n)),$$

and

$$z_0 := x_0 - (I - A)^{-1}B(\Phi(u))(0), \quad v_0 := (\Phi(u))(0) - \tilde{r}.$$

Choose $c \in (k\lambda, K)$, and define

$$\mathbf{H}(z) = -C(zI - A)^{-1}(I - A)^{-1}B + J,$$

where $J := 1/c - \mathbf{G}(1)/2$. Then

$$\mathbf{H}(z) = \frac{1}{z-1}(\mathbf{G}(z) - \mathbf{G}(1)) + J.$$

Since $c < K$, there exists $\varepsilon > 0$ such that

$$\frac{1}{c} + \operatorname{Re} \frac{\mathbf{G}(z)}{z-1} \geq \varepsilon \quad \forall\, |z| > 1,$$

and hence, using the identity

$$\operatorname{Re}\left(\frac{1}{e^{i\theta}-1}\right) = -\frac{1}{2} \quad \forall\, \theta \in (0, 2\pi),$$

we may conclude that

$$\operatorname{Re} \mathbf{H}(e^{i\theta}) \geq \varepsilon \quad \forall\, \theta \in [0, 2\pi).$$

An application of the discrete-time positive-real lemma (see the appendix) shows that there exist $P \in L(X)$, $P = P^* \geq 0$, $L \in L(\mathbb{R}, X)$, and $W \in \mathbb{R}$ such that

$$A^*PA - P = -LL^*, \tag{3.7a}$$

$$A^*P(I-A)^{-1}B = LW - C^*, \tag{3.7b}$$

$$W^2 = 2J - B^*(I-A^*)^{-1}P(I-A)^{-1}B. \tag{3.7c}$$

For $n \in \mathbb{Z}_+$, define

$$V(n) = \langle z(n), Pz(n) \rangle + \mathbf{G}(1)v(n)^2.$$

Using (3.6) and (3.7), we obtain for all $n \in \mathbb{Z}_+$

$$\begin{aligned}
V(n+1) - V(n) &= \langle z(n+1), Pz(n+1) \rangle - \langle z(n), Pz(n) \rangle + \mathbf{G}(1)(v(n+1)^2 - v(n)^2) \\
&= -(L^*z(n))^2 - 2(L^*z(n))Ww(n) + 2(Cz(n))w(n) \\
&\qquad + w(n)(2J - W^2)w(n) + \mathbf{G}(1)(w(n)^2 + 2w(n)v(n)) \\
&= -(L^*z(n))^2 - (Ww(n))^2 - 2(L^*z(n))Ww(n) \\
&\qquad + 2(Cz(n))w(n) + \frac{2}{c}w(n)^2 + 2\mathbf{G}(1)w(n)v(n) \\
&= -(L^*z(n) + Ww(n))^2 + 2(Cz(n))w(n) \\
&\qquad + \frac{2}{c}w(n)^2 - 2\mathbf{G}(1)kd(n)[\mathbf{G}(1)v(n)^2 + (Cz(n))v(n)] \\
&= -(L^*z(n) + Ww(n))^2 + \frac{2}{c}w(n)^2 - 2kd(n)(\mathbf{G}(1)v(n) + Cz(n))^2 \\
&= -(L^*z(n) + Ww(n))^2 - 2\left(kd(n) - \frac{k^2d(n)^2}{c}\right)(\mathbf{G}(1)v(n) + Cz(n))^2.
\end{aligned}$$

Summing from $n = 0$ to $n = \infty$ then gives

$$2\sum_{n=0}^{\infty}\left(kd(n) - \frac{k^2d(n)^2}{c}\right)(\mathbf{G}(1)v(n) + Cz(n))^2 \leq V(0) < \infty. \tag{3.8}$$

Now, since $c > k\lambda$ and $d(n) \in [0, \lambda]$, we have

$$kd(n) - \frac{k^2d(n)^2}{c} = kd(n)\left(1 - \frac{kd(n)}{c}\right) \geq kd(n)\left(1 - \frac{k\lambda}{c}\right) \geq k\frac{\delta}{\lambda}d(n)^2 \quad \forall\, n \in \mathbb{Z}_+,$$

where $\delta := 1 - k\lambda/c > 0$. Therefore, (3.8) implies that

$$d(Cz + \mathbf{G}(1)v) \in l^2(\mathbb{Z}_+), \tag{3.9}$$

and hence

$$w \in l^2(\mathbb{Z}_+). \tag{3.10}$$

Appealing to the fact that A is power-stable, we may conclude from (3.6a) and (3.10) that

$$z \in l^2(\mathbb{Z}_+). \tag{3.11}$$

Consequently, $Cz \in l^2(\mathbb{Z}_+)$, and hence, by (3.9) and the boundedness of d,

$$dv \in l^2(\mathbb{Z}_+). \tag{3.12}$$

From (3.11) and (3.12) we obtain that

$$(Cz)dv \in l^1(\mathbb{Z}_+). \tag{3.13}$$

Using (3.8), (3.11)–(3.13), and the boundedness of d it follows that

$$dv^2 \in l^1(\mathbb{Z}_+). \tag{3.14}$$

It follows from (3.6b) that, for all $m \in \mathbb{Z}_+$,

$$v(m+1)^2 = v(0)^2 + \sum_{n=0}^{m} w(n)^2 + 2\sum_{n=0}^{m} v(n)w(n). \tag{3.15}$$

Combining (3.15) with (3.10), (3.13), and (3.14) and recalling that $w = -kd(Cz + \mathbf{G}(1)v)$, we see that there exists a number $\nu \in \mathbb{R}_+$ such that

$$\lim_{n\to\infty} v(n)^2 = \nu. \tag{3.16}$$

In order to prove statement (1), it is sufficient to show that $\nu = 0$. Seeking a contradiction, suppose that $\nu > 0$. By (3.10), $\lim_{n\to\infty} w(n) = 0$, and thus we may conclude from (3.6b) that

$$\lim_{n\to\infty} (v(n+1) - v(n)) = 0. \tag{3.17}$$

Since $\nu > 0$, (3.16) and (3.17) yield that $v(n)$ does not change sign for sufficiently large n, and so

$$\lim_{n\to\infty} v(n) = \sqrt{\nu} \qquad \text{or} \qquad \lim_{n\to\infty} v(n) = -\sqrt{\nu}.$$

Assuming that $\lim_{n\to\infty} v(n) = -\sqrt{\nu}$ (the case $\lim_{n\to\infty} v(n) = \sqrt{\nu}$ can be dealt with in an entirely analogous fashion), we obtain that

$$\Phi_\infty := \lim_{n\to\infty} (\Phi(u))(n) < \tilde{r}, \tag{3.18}$$

and thus

$$\lim_{n\to\infty} x(n) = (I - A)^{-1} B\Phi_\infty. \tag{3.19}$$

It then follows from (3.3b), (3.18), and (3.19) that

(3.20)
$$\lim_{n\to\infty}(u(n+1)-u(n)) = k(r - C(I-A)^{-1}B\Phi_\infty - D\Phi_\infty) = k\mathbf{G}(1)(\tilde{r}-\Phi_\infty) > 0.$$

Therefore, $\lim_{n\to\infty} u(n) = \infty$, and u is ultimately nondecreasing, so by (D4) and the assumption that $\tilde{r} \in \operatorname{clos}(\operatorname{NVS}\Phi)$ we obtain

$$\Phi_\infty = \lim_{n\to\infty}(\Phi(u))(n) = \sup(\operatorname{NVS}\Phi) \geq \tilde{r},$$

contradicting (3.18). Therefore, $\lim_{n\to\infty} v(n) = 0$, and, consequently, $\lim_{n\to\infty}(\Phi(u))(n) = \tilde{r}$, which is statement (1).

Statement (2) follows immediately from statement (1). Statement (3) is an easy consequence of statements (1) and (2). Finally, to prove statement (4), let $\tilde{r} \in \operatorname{int}(\operatorname{clos}(\operatorname{NVS}\Phi))$. Then the boundedness of u follows immediately from statement (1) and (D5). □

We see from the proof of Theorem 3.1 that (D5) is needed only for statement (4).

One of the conditions imposed in Theorem 3.1 is that $r/\mathbf{G}(1) \in \operatorname{clos}(\operatorname{NVS}\Phi)$. The following proposition shows that this condition is close to being necessary for tracking insofar as, if tracking of r is achievable while maintaining the boundedness of $\Phi(u)$, then $r/\mathbf{G}(1) \in \operatorname{clos}(\operatorname{NVS}\Phi)$.

PROPOSITION 3.2. *Let $\lambda > 0$ and $r \in \mathbb{R}$. Suppose that $\Phi \in \mathcal{D}(\lambda)$, A is power-stable, and $\mathbf{G}(1) > 0$. If there exist an initial condition $x_0 \in X$ and a function $u \in F(\mathbb{Z}_+, \mathbb{R})$ such that $\Phi(u)$ is bounded and*

$$\lim_{n\to\infty}[Cx(n) + D(\Phi(u))(n)] = r\,,$$

where $x \in F(\mathbb{Z}_+, X)$ is given by (3.2a)*, then $r/\mathbf{G}(1) \in \operatorname{clos}(\operatorname{NVS}\Phi)$.*

Proof. Since $\Phi(u)$ is bounded and A is power-stable, x is bounded. Let $n \in \mathbb{Z}_+$, and define $y : \mathbb{Z}_+ \to \mathbb{R}$ by (3.2b); then

$$y(n) = C(x(n) - (I-A)^{-1}B(\Phi(u))(n)) + \mathbf{G}(1)(\Phi(u))(n),$$

and therefore

$$C(A-I)^{-1}(x(n+1)-x(n)) = y(n) - \mathbf{G}(1)(\Phi(u))(n).$$

For $p, m \in \mathbb{Z}_+$ with $p > m$, summing the above from m to $p-1$ gives

$$C(A-I)^{-1}(x(p)-x(m)) = \sum_{k=m}^{p-1}(y(k) - \mathbf{G}(1)(\Phi(u))(k)). \tag{3.21}$$

Seeking a contradiction, let us suppose that $r/\mathbf{G}(1) \notin \operatorname{clos}(\operatorname{NVS}\Phi)$. Since $\lim_{n\to\infty} y(n) = r$ and $\operatorname{clos}(\operatorname{NVS}\Phi)$ is an interval (see Remark 2.1, part (3)), there exist $\varepsilon > 0$, $\beta \in \{-1, 1\}$, and $m \in \mathbb{Z}_+$ such that

$$\beta(y(n) - \mathbf{G}(1)(\Phi(u))(n)) \geq \varepsilon \quad \forall\, n \geq m.$$

Combining the above with (3.21), it follows that

$$\beta C(A-I)^{-1}(x(n)-x(m)) = \sum_{k=m}^{n-1}\beta(y(k) - \mathbf{G}(1)(\Phi(u))(k)) \geq \varepsilon(n-m) \quad \forall\, n > m.$$

Therefore, $\lim_{n\to\infty} \beta C(A-I)^{-1}x(n) = \infty$, contradicting the boundedness of x. □

4. A class of continuous-time hysteresis operators and their discretizations. We call a function $f : \mathbb{R}_+ \to \mathbb{R}_+$ a *time transformation* if f is continuous and nondecreasing and satisfies $f(0) = 0$ and $\lim_{t\to\infty} f(t) = \infty$, i.e., f is continuous, nondecreasing, and surjective. We denote the set of all time transformations $f : \mathbb{R}_+ \to \mathbb{R}_+$ by $\mathcal{T}$. For all $\tau \in \mathbb{R}_+$, we define a (continuous-time) projection operator $\mathbf{Q}^c_\tau : F(\mathbb{R}_+, \mathbb{R}) \to F(\mathbb{R}_+, \mathbb{R})$ by

$$(\mathbf{Q}^c_\tau v)(t) = \begin{cases} v(t) & \text{for } 0 \le t \le \tau, \\ v(\tau) & \text{for } t > \tau. \end{cases}$$

In the following, let $\mathcal{F} \subset F(\mathbb{R}_+, \mathbb{R})$, $\mathcal{F} \neq \emptyset$. We introduce the following two assumptions on $\mathcal{F}$:

(F1) $v \circ f \in \mathcal{F}$ for all $v \in \mathcal{F}$ and all $f \in \mathcal{T}$;

(F2) $\mathbf{Q}^c_t(\mathcal{F}) \subset \mathcal{F}$ for all $t \in \mathbb{R}_+$.

We call an operator $\Phi : \mathcal{F} \to F(\mathbb{R}_+, \mathbb{R})$ *causal* if for all $v, w \in \mathcal{F}$ and all $\tau \in \mathbb{R}_+$ with $v(t) = w(t)$ for all $t \in [0, \tau]$ it follows that $(\Phi(v))(t) = (\Phi(w))(t)$ for all $t \in [0, \tau]$. An operator $\Phi : \mathcal{F} \to F(\mathbb{R}_+, \mathbb{R})$ is called *rate independent* if $\mathcal{F}$ satisfies (F1) and

$$(\Phi(v \circ f))(t) = (\Phi(v))(f(t)) \quad \forall\, v \in \mathcal{F}, \quad \forall\, f \in \mathcal{T}, \quad \forall\, t \in \mathbb{R}_+.$$

A functional $\varphi : \mathcal{F} \to \mathbb{R}$ is called *rate independent* if $\mathcal{F}$ satisfies (F1) and

$$\varphi(u \circ f) = \varphi(u) \quad \forall\, u \in \mathcal{F}, \quad \forall\, f \in \mathcal{T}.$$

DEFINITION 4.1. *Let $\mathcal{F} \subset F(\mathbb{R}_+, \mathbb{R})$, $\mathcal{F} \neq \emptyset$. An operator $\Phi : \mathcal{F} \to F(\mathbb{R}_+, \mathbb{R})$ is called a hysteresis operator if $\mathcal{F}$ satisfies* (F1) *and Φ is causal and rate independent.*

For $\mathcal{F} \subset F(\mathbb{R}_+, \mathbb{R})$, $\mathcal{F} \neq \emptyset$, let $\mathcal{F}^{\mathrm{uc}}$ denote the set of all ultimately constant $u \in \mathcal{F}$, i.e.,

$$\mathcal{F}^{\mathrm{uc}} = \{u \in \mathcal{F} \,|\, u \text{ is ultimately constant}\}.$$

Clearly, if $\mathcal{F}$ satisfies (F2), then $\mathcal{F}^{\mathrm{uc}} \neq \emptyset$. Moreover, if $\mathcal{F}$ satisfies (F1), then so does $\mathcal{F}^{\mathrm{uc}}$. The proof of the following proposition can be found in [10].

PROPOSITION 4.2. *Let $\mathcal{F} \subset F(\mathbb{R}_+, \mathbb{R})$, $\mathcal{F} \neq \emptyset$, and assume that* (F1) *and* (F2) *are satisfied. If $\Phi : \mathcal{F} \to F(\mathbb{R}_+, \mathbb{R})$ is a hysteresis operator, then the following statements hold:*

(1) *for all $v \in \mathcal{F}$ and all $\tau \in \mathbb{R}_+$*

$$(\Phi(\mathbf{Q}^c_\tau v))(t) = (\Phi(v))(\tau) \quad \forall\, t \ge \tau;$$

(2) *the functional*

$$\varphi : \mathcal{F}^{\mathrm{uc}} \to \mathbb{R}, \quad v \mapsto \lim_{t\to\infty} (\Phi(v))(t) \tag{4.1}$$

is rate independent and satisfies

$$(\Phi(v))(t) = \varphi(\mathbf{Q}^c_t v) \quad \forall\, v \in \mathcal{F}, \quad \forall\, t \in \mathbb{R}_+. \tag{4.2}$$

Conversely, if $\varphi : \mathcal{F}^{\mathrm{uc}} \to \mathbb{R}$ is a rate independent functional, then $\Phi : \mathcal{F} \to F(\mathbb{R}_+, \mathbb{R})$ given by (4.2) *is a hysteresis operator and satisfies*

$$\lim_{t\to\infty} (\Phi(v))(t) = \varphi(v) \quad \forall\, v \in \mathcal{F}^{\mathrm{uc}}. \tag{4.3}$$

For a hysteresis operator $\Phi : \mathcal{F} \to F(\mathbb{R}_+, \mathbb{R})$, we call the rate independent functional φ defined by (4.1) the *representing functional* of Φ.

For any $v \in F(\mathbb{R}_+, \mathbb{R})$ and any $t \in \mathbb{R}_+$, we define

$$M(v,t) := \{\tau \in (t,\infty) \,|\, v \text{ is monotone on } (t,\tau)\}.$$

If v is piecewise monotone, then $M(v,t) \neq \emptyset$ for all $t \in \mathbb{R}_+$ and the *standard monotonicity partition* $t_0 < t_1 < t_2 < \cdots$ of v is defined recursively by setting $t_0 = 0$ and $t_{i+1} = \sup M(v,t_i)$ for all $i \in \mathbb{Z}_+$ such that $M(v,t_i)$ is bounded. If v is piecewise monotone and ultimately constant, then the standard monotonicity partition of v is finite.

The space of all piecewise monotone $v \in C(\mathbb{R}_+, \mathbb{R})$ is denoted by $C_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R})$. We define $C^{\mathrm{uc}}_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R})$ to be the space of all ultimately constant $v \in C_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R})$. Let $F^{\mathrm{uc}}(\mathbb{Z}_+, \mathbb{R})$ denote the space of ultimately constant $v : \mathbb{Z}_+ \to \mathbb{R}$. We define the *restriction operator* $R : C^{\mathrm{uc}}_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R}) \to F^{\mathrm{uc}}(\mathbb{Z}_+, \mathbb{R})$ by

$$(R(v))(k) = \begin{cases} v(t_k) & \text{for } k \in [0,m] \cap \mathbb{Z}_+, \\ \lim_{t\to\infty} v(t) & \text{for } k \in \mathbb{Z}_+ \setminus [0,m], \end{cases}$$

where $0 = t_0 < t_1 < \cdots < t_m$ is the standard monotonicity partition of v.

In the following, we want to extend hysteresis operators defined on $C_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R})$ to spaces of piecewise continuous functions. This requires some preparation. For $\tau > 0$, we define the *prolongation operator* $P_\tau : F(\mathbb{Z}_+, \mathbb{R}) \to C_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R})$ by letting $P_\tau u$ be the linear interpolate for the values $(P_\tau u)(i\tau) = u(i)$. Let $NPC(\mathbb{R}_+, \mathbb{R}) \subset PC(\mathbb{R}_+, \mathbb{R})$ denote the space of all *normalized piecewise continuous* functions $v : \mathbb{R}_+ \to \mathbb{R}$; that is, v is piecewise continuous and is right-continuous or left-continuous at each point $t \in \mathbb{R}_+$. The space of all piecewise monotone functions $v \in NPC(\mathbb{R}_+, \mathbb{R})$ is denoted by $NPC_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R})$, while $NPC^{\mathrm{uc}}_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R})$ denotes the space of ultimately constant $v \in NPC_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R})$. We note that $NPC_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R})$ and $NPC^{\mathrm{uc}}_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R})$ both satisfy (F1) and (F2). For $v \in NPC^{\mathrm{uc}}_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R})$, we define $v(\infty) := \lim_{\tau\to\infty} v(\tau)$.

Let $v \in NPC^{\mathrm{uc}}_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R})$, and let $0 = t_0 < t_1 < \cdots < t_m$ be the standard monotonicity partition of v. We define the operator $\rho : NPC^{\mathrm{uc}}_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R}) \to F^{\mathrm{uc}}(\mathbb{Z}_+, \mathbb{R})$ by

$$\rho(v) = (v(t_0), v(t_1-), v(t_1+), v(t_2-), v(t_2+), \dots, v(t_m-), v(t_m+), v(\infty), v(\infty), \dots).$$

For $\tau > 0$, define

$$R_e : NPC^{\mathrm{uc}}_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R}) \to F^{\mathrm{uc}}(\mathbb{Z}_+, \mathbb{R}), \quad v \mapsto R((P_\tau \circ \rho)(v)).$$

The operator R_e is an extension of R, and the definition of R_e is independent of τ (see [10]). The function v, shown in Figure 4, is a normalized piecewise continuous function which is piecewise monotone and ultimately constant, so $v \in NPC^{\mathrm{uc}}_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R})$. It has standard monotonicity partition $0 = t_0 < t_1 < t_2 < t_3 < t_4$,

$$\rho(v) = (v_0, v_7, v_6, v_4, v_4, v_5, v_3, v_7, v_2, v_1, v_1, v_1, \dots),$$

and

$$R_e(v) = (v_0, v_7, v_4, v_5, v_3, v_7, v_1, v_1, v_1, \dots).$$

For any rate independent $\varphi : C^{\mathrm{uc}}_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R}) \to \mathbb{R}$, we define

$$\varphi_e : NPC^{\mathrm{uc}}_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R}) \to \mathbb{R}, \quad v \mapsto \varphi((P_\tau \circ R_e)(v)), \tag{4.4}$$

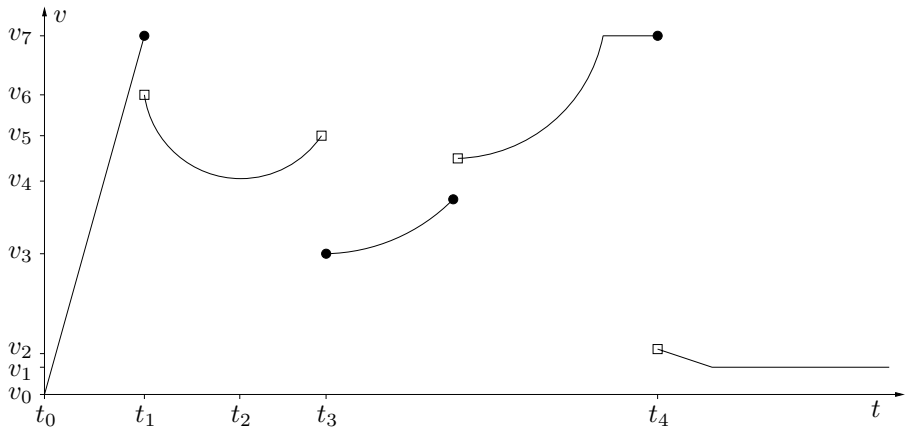


FIG. 4. *Example of a function in* $NPC^{\mathrm{uc}}_{\mathrm{pm}}(\mathbb{R}_+,\mathbb{R})$.

where $\tau > 0$. From [10] we know that φ_e does not depend on τ and that φ_e is a rate independent extension of φ.

Let $\Phi : C_{\mathrm{pm}}(\mathbb{R}_+,\mathbb{R}) \to C(\mathbb{R}_+,\mathbb{R})$ be a hysteresis operator, and let $\varphi : C^{\mathrm{uc}}_{\mathrm{pm}}(\mathbb{R}_+,\mathbb{R}) \to \mathbb{R}$ be the representing functional of Φ. Define

$$\Phi_e : NPC_{\mathrm{pm}}(\mathbb{R}_+,\mathbb{R}) \to F(\mathbb{R}_+,\mathbb{R})$$

by setting

$$(\Phi_e(v))(t) = \varphi_e(\mathbf{Q}^c_t\, v) \quad \forall\, t \in \mathbb{R}_+, \tag{4.5}$$

where φ_e is the extension of φ to $NPC^{\mathrm{uc}}_{\mathrm{pm}}(\mathbb{R}_+,\mathbb{R})$ given by (4.4).

Define $\mathcal{S}_\tau$ to be the set of all right-continuous step functions $v : \mathbb{R}_+ \to \mathbb{R}$ of step length $\tau > 0$. The following result was proved in [10].

PROPOSITION 4.3. *Let* $\Phi : C_{\mathrm{pm}}(\mathbb{R}_+,\mathbb{R}) \to C(\mathbb{R}_+,\mathbb{R})$ *be a hysteresis operator. Then*

(1) Φ_e *is an extension of* Φ;
(2) Φ_e *is a hysteresis operator with representing functional* φ_e;
(3) *for* $v, w \in NPC_{\mathrm{pm}}(\mathbb{R}_+,\mathbb{R})$ *and* $t \in \mathbb{R}_+$

$$R_e(\mathbf{Q}^c_t\, v) = R_e(\mathbf{Q}^c_t\, w) \quad \Longrightarrow \quad (\Phi_e(v))(t) = (\Phi_e(w))(t);$$

(4) $\Phi_e(NPC_{\mathrm{pm}}(\mathbb{R}_+,\mathbb{R})) \subset NPC(\mathbb{R}_+,\mathbb{R})$;
(5) $\Phi_e(\mathcal{S}_\tau) \subset \mathcal{S}_\tau$.

For given $v \in \mathcal{S}_\tau$ we introduce continuous piecewise monotone "approximations" $v_k \in C_{\mathrm{pm}}(\mathbb{R}_+,\mathbb{R})$ $(k = 1, 2, \dots)$ as follows: let $\varepsilon_k \in (0,\tau)$ with $\lim_{k\to\infty} \varepsilon_k = 0$, and define the following:

(i) if $t \in [(i+1)\tau - \varepsilon_k, (i+1)\tau)$, $i \in \mathbb{Z}_+$,

$$v_k(t) = v(i\tau) + \frac{v((i+1)\tau) - v(i\tau)}{\varepsilon_k}(t - (i+1)\tau + \varepsilon_k);$$

(ii) $v_k(t) = v(t)$ otherwise.

(See Figure 5 for an illustration.)

The following result shows that, for a given $v \in \mathcal{S}_\tau$, the functions $\Phi(v_k)$ approximate $\Phi_e(v)$ pointwise.

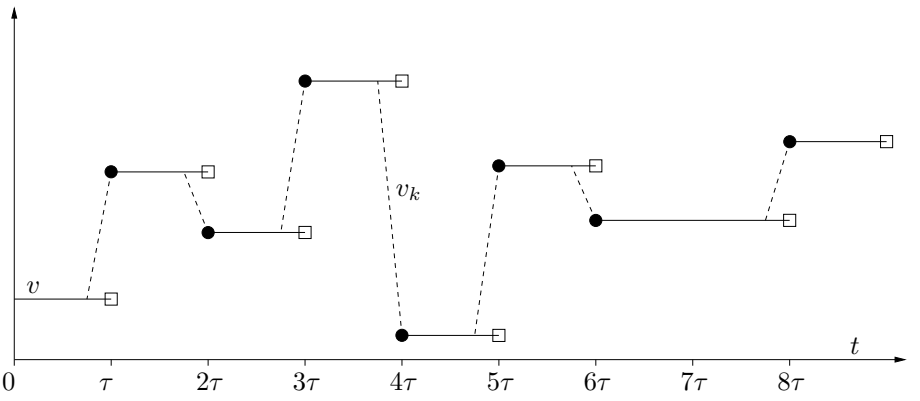


FIG. 5. *Example of $v \in \mathcal{S}_\tau$ and its approximation v_k.*

PROPOSITION 4.4. *Let $\Phi : C_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R}) \to C(\mathbb{R}_+, \mathbb{R})$ be a hysteresis operator. Then*

$$(\Phi_e(v))(t) = \lim_{k\to\infty} (\Phi(v_k))(t) \quad \forall\, v \in \mathcal{S}_\tau, \quad \forall\, t \in \mathbb{R}_+.$$

Proof. Let $t \in \mathbb{R}_+$, and let $m \in \mathbb{Z}_+$ and $s \in [0, \tau)$ be such that $t = m\tau + s$. Choose k sufficiently large so that

$$t - m\tau = s < (1 - \varepsilon_k)\tau. \tag{4.6}$$

Then, using Proposition 4.3, part (5), we have

$$(\Phi_e(v))(t) = (\Phi_e(v))(m\tau) = \varphi_e(\mathbf{Q}^c_{m\tau}\, v) = \varphi((P_\tau \circ R_e)(\mathbf{Q}^c_{m\tau}\, v)). \tag{4.7}$$

Now $R_e(\mathbf{Q}^c_{m\tau}\, v) = R(\mathbf{Q}^c_{m\tau}\, v_k) = R_e(\mathbf{Q}^c_{m\tau}\, v_k)$, and hence, using (4.6), (4.7), Proposition 4.2, part (1), and Proposition 4.3, we obtain for all sufficiently large k

$$\begin{aligned}(\Phi_e(v))(t) &= \varphi((P_\tau \circ R_e)(\mathbf{Q}^c_{m\tau}\, v_k)) = \varphi_e(\mathbf{Q}^c_{m\tau}\, v_k) = (\Phi_e(v_k))(m\tau) \\ &= (\Phi(v_k))(m\tau) = (\Phi(v_k))(t). \qquad \square\end{aligned}$$

A function $v \in F(\mathbb{R}_+, \mathbb{R})$ is called *ultimately nondecreasing* if there exists $T \in \mathbb{R}_+$ such that v is nondecreasing on $[T, \infty)$. Let $\mathcal{F} \subset F(\mathbb{R}_+, \mathbb{R})$, $\mathcal{F} \neq \emptyset$. The *numerical value set* $\mathrm{NVS}\,\Phi$ of an operator $\Phi : \mathcal{F} \to F(\mathbb{R}_+, \mathbb{R})$ is defined by

$$\mathrm{NVS}\,\Phi := \{(\Phi(v))(t) \,|\, v \in \mathcal{F},\, t \in \mathbb{R}_+\}.$$

For $\alpha \geq 0$, $u \in C_{\mathrm{pm}}([0, \alpha], \mathbb{R})$, and $\delta_1, \delta_2 > 0$, we define $\mathcal{C}_{\mathrm{pm}}(u; \delta_1, \delta_2)$ to be the set of all $v \in C_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R})$ such that

$$v(t) = u(t)\ \ \forall\, t \in [0, \alpha] \qquad \text{and} \qquad |v(t) - u(\alpha)| \leq \delta_1\ \ \forall\, t \in [\alpha, \alpha + \delta_2].$$

We introduce the following seven assumptions on the operator $\Phi : C_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R}) \to C(\mathbb{R}_+, \mathbb{R})$:

(C1) Φ is causal;
(C2) Φ is rate independent;
(C3) $\Phi(AC(\mathbb{R}_+, \mathbb{R}) \cap C_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R})) \subset AC(\mathbb{R}_+, \mathbb{R})$;

(C4) Φ is monotone in the sense that, for all $v \in AC(\mathbb{R}_+, \mathbb{R}) \cap C_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R})$ with $\Phi(v) \in AC(\mathbb{R}_+, \mathbb{R})$,

$$\frac{d}{dt}(\Phi(v))(t)\,\dot{v}(t) \geq 0, \quad \text{a.e. } t \in \mathbb{R}_+;$$

(C5) there exists $\lambda > 0$ such that for all $\alpha \in \mathbb{R}_+$, $u \in C_{\mathrm{pm}}([0,\alpha], \mathbb{R})$, there exist numbers $\delta_1, \delta_2 > 0$ such that for all $v, w \in \mathcal{C}_{\mathrm{pm}}(u; \delta_1, \delta_2)$

$$\sup_{t\in[\alpha,\alpha+\delta_2]} |(\Phi(v))(t) - (\Phi(w))(t)| \leq \lambda \sup_{t\in[\alpha,\alpha+\delta_2]} |v(t) - w(t)|;$$

(C6) if $v \in C_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R})$ is ultimately nondecreasing and $\lim_{t\to\infty} v(t) = \infty$, then $\Phi(v)(t)$ and $\Phi(-v)(t)$ converge to $\sup \mathrm{NVS}\,\Phi$ and $\inf \mathrm{NVS}\,\Phi$, respectively, as $t \to \infty$;

(C7) if, for $v \in C_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R})$, $L := \lim_{t\to\infty}(\Phi(v))(t)$ exists with $L \in \operatorname{int} \mathrm{NVS}\,\Phi$, then v is bounded.

REMARK 4.5. *We note that if* Φ *satisfies* (C1) *and* (C6)*, then* $\mathrm{NVS}\,\Phi$ *is an interval.*

If $\Phi : C_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R}) \to C(\mathbb{R}_+, \mathbb{R})$ satisfies (C5), then any number $l > 0$ such that (C5) holds for $\lambda = l$ is called a *Lipschitz constant* of Φ.

DEFINITION 4.6. *Let* $\lambda > 0$*. The set of all operators* $\Phi : C_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R}) \to C(\mathbb{R}_+, \mathbb{R})$ *satisfying* (C1)–(C7) *and having Lipschitz constant* λ *is denoted by* $\mathcal{C}(\lambda)$*.*

We consider four examples of hysteresis operators which satisfy (C1)–(C7).

Static nonlinearities. For $f \in F(\mathbb{R}, \mathbb{R})$, the corresponding static nonlinearity

$$F(\mathbb{R}_+, \mathbb{R}) \to F(\mathbb{R}_+, \mathbb{R}), \quad u \mapsto f \circ u$$

is in $\mathcal{C}(\lambda)$, provided that f is nondecreasing and globally Lipschitz with Lipschitz constant λ.

Relay. Relay (also called *passive* or *positive*) hysteresis, has been discussed in a mathematically rigorous context in a number of references; see, for example, [11] and [20]. To give a formal definition of relay, let $a_1, a_2 \in \mathbb{R}$ with $a_1 < a_2$, and let $\rho_1 : [a_1, \infty) \to \mathbb{R}$ and $\rho_2 : (-\infty, a_2] \to \mathbb{R}$ be nondecreasing and globally Lipschitz (both with Lipschitz constant $\lambda > 0$) and such that $\rho_1(a_1) = \rho_2(a_1)$ and $\rho_1(a_2) = \rho_2(a_2)$. For $v \in C_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R})$ and $t \geq 0$ define

$$S(v,t) := v^{-1}(\{a_1, a_2\}) \cap [0,t], \quad \tau(v,t) := \begin{cases} \max S(v,t) & \text{if } S(v,t) \neq \emptyset, \\ -1 & \text{if } S(v,t) = \emptyset. \end{cases}$$

Following Macki, Nistri, and Zecca [20], for each $\xi \in \mathbb{R}$, we define an operator $\mathcal{R}_\xi : C_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R}) \to C(\mathbb{R}_+, \mathbb{R})$ by

$$(\mathcal{R}_\xi(v))(t) = \begin{cases} \rho_2(v(t)) & \text{if } v(t) \leq a_1, \\ \rho_1(v(t)) & \text{if } v(t) \geq a_2, \\ \rho_2(v(t)) & \text{if } v(t) \in (a_1, a_2),\ \tau(v,t) \neq -1, \text{ and } v(\tau(v,t)) = a_1, \\ \rho_1(v(t)) & \text{if } v(t) \in (a_1, a_2),\ \tau(v,t) \neq -1, \text{ and } v(\tau(v,t)) = a_2, \\ \rho_1(v(t)) & \text{if } v(t) \in (a_1, a_2),\ \tau(v,t) = -1, \text{ and } \xi > 0, \\ \rho_2(v(t)) & \text{if } v(t) \in (a_1, a_2),\ \tau(v,t) = -1, \text{ and } \xi \leq 0. \end{cases}$$

The number ξ plays the role of an "initial state" which determines the output value $(\mathcal{R}_\xi(v))(t)$ if $v(s) \in (a_1, a_2)$ for all $s \in [0,t]$. The relay hysteresis operator $\mathcal{R}_\xi$ is

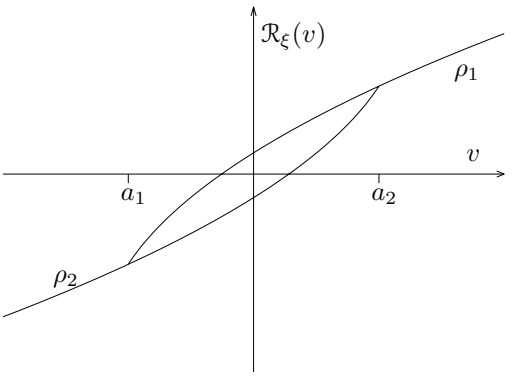


FIG. 6. *Relay hysteresis.*

illustrated in Figure 6. It is trivial that $\mathcal{R}_\xi$ is causal and rate independent (i.e., (C1) and (C2) hold). From [11] we know that $\mathcal{R}_\xi$ satisfies (C3)–(C7) and that λ is a Lipschitz constant of $\mathcal{R}_\xi$. It follows that $\mathcal{R}_\xi \in \mathcal{C}(\lambda)$.

Backlash. A discussion of the continuous-time *backlash* operator (also called the *play* operator) can be found in a number of references; see, for example, [1], [2], [8], and [11]. For all $h \in \mathbb{R}_+$ and all $\xi \in \mathbb{R}$, we define the continuous-time backlash operator $\mathcal{B}_{h,\xi} : C_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R}) \to C(\mathbb{R}_+, \mathbb{R})$ by

$$(\mathcal{B}_{h,\xi}(v))(t) = \begin{cases} b_h(v(0), \xi) & \text{for } t = 0\,, \\ b_h(v(t), (\mathcal{B}_{h,\xi}(v))(t_i)) & \text{for } t_i < t \le t_{i+1}, \quad i \in \mathbb{Z}_+, \end{cases}$$

where the function $b_h : \mathbb{R}^2 \to \mathbb{R}$ is given by (2.1) and $0 = t_0 < t_1 < t_2 < \cdots$ is such that $\lim_{n\to\infty} t_n = \infty$ and v is monotone on each interval (t_i, t_{i+1}). We remark that ξ plays the role of an "initial state." It is not difficult to show that the definition is independent of the choice of the partition (t_i); see [11]. The diagram illustrating how $\mathcal{B}_{h,\xi}$ acts is the same as in the discrete-time case; see Figure 3. It is trivial that $\mathcal{B}_{h,\xi}$ is causal (i.e., $\mathcal{B}_{h,\xi}$ satisfies (C1)). Furthermore, it is well known and easy to check that $\mathcal{B}_{h,\xi}$ is rate independent (i.e., $\mathcal{B}_{h,\xi}$ satisfies (C2)). From [11] we know that $\mathcal{B}_{h,\xi}$ satisfies (C3)–(C7) and that $\lambda = 1$ is a Lipschitz constant of $\mathcal{B}_{h,\xi}$. Therefore, we may conclude that $\mathcal{B}_{h,\xi} \in \mathcal{C}\,(1)$.

Elastic-plastic. The *elastic-plastic* operator (also called the *stop* operator) has been discussed in a mathematically rigorous context in a number of references; see, for example, [1], [2], [8], and [11]. To give a formal definition of the elastic-plastic operator, define for each $h \in \mathbb{R}_+$ the function $e_h : \mathbb{R} \to \mathbb{R}$ by

$$e_h(w) = \min\{h, \max\{-h, w\}\}.$$

For all $h \in \mathbb{R}_+$ and all $\xi \in \mathbb{R}$, we define an operator $\mathcal{E}_{h,\xi} : C_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R}) \to C(\mathbb{R}_+, \mathbb{R})$ by

$$(\mathcal{E}_{h,\xi}(v))(t) = \begin{cases} e_h(v(0) - \xi) & \text{for } t = 0, \\ e_h(v(t) - v(t_i) + (\mathcal{E}_{h,\xi}(v))(t_i)) & \text{for } t_i < t \le t_{i+1}, \quad i \in \mathbb{Z}_+, \end{cases}$$

where $0 = t_0 < t_1 < t_2 < \cdots$ is such that $\lim_{n\to\infty} t_n = \infty$ and v is monotone on each interval (t_i, t_{i+1}). Again, ξ plays the role of an "initial state." The backlash and

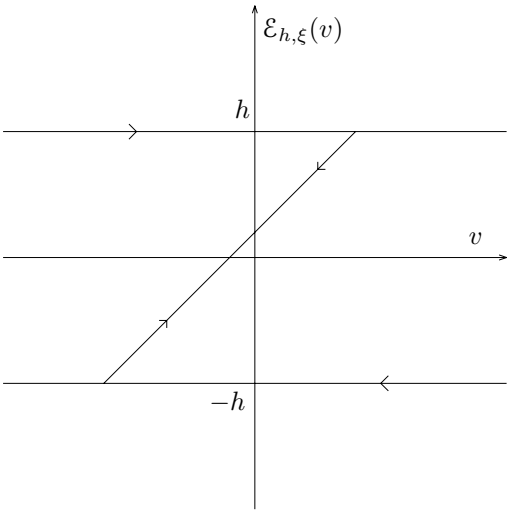


FIG. 7. *Elastic-plastic hysteresis.*

elastic-plastic operators are closely related:

$$\mathcal{E}_{h,\xi}(v) + \mathcal{B}_{h,\xi}(v) = v \quad \forall\, v \in C_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R}); \tag{4.8}$$

see [2, p. 44]. The elastic-plastic operator $\mathcal{E}_{h,\xi}$ is illustrated in Figure 7. Since $\mathcal{B}_{h,\xi}$ is causal and rate independent, it follows from (4.8) that $\mathcal{E}_{h,\xi}$ is causal and rate independent; i.e., (C1) and (C2) hold. From [11] we know that $\mathcal{E}_{h,\xi}$ satisfies (C3)–(C7) and that $\lambda = 2$ is a Lipschitz constant of $\mathcal{E}_{h,\xi}$. Therefore, $\mathcal{E}_{h,\xi} \in \mathcal{C}(2)$.

We remark that a large class of Prandtl and Preisach hysteresis operators satisfy (C1)–(C7); see [11] for details.

The following lemma will be needed later in this section.

LEMMA 4.7. *Let $\Phi \in \mathcal{C}(\lambda)$; then for every $v \in C_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R}) \cap AC(\mathbb{R}_+, \mathbb{R})$ and $t_2 > t_1 \geq 0$, there exists a constant $\eta \in [0, \lambda]$ such that*

$$v \text{ affine linear on } [t_1, t_2] \quad \Longrightarrow \quad (\Phi(v))(t_2) - (\Phi(v))(t_1) = \eta(v(t_2) - v(t_1)).$$

Proof. Let $\Phi \in \mathcal{C}(\lambda)$, $v \in C_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R}) \cap AC(\mathbb{R}_+, \mathbb{R})$, and $t_2 > t_1 \geq 0$, and assume that v is affine linear on $[t_1, t_2]$. By Proposition 4.2, part (1) and [11, Lemma 3.2, part (c)] there exists a measurable function $d : \mathbb{R}_+ \to [0, \lambda]$ such that

$$(\Phi(v))(t_2) - (\Phi(v))(t_1) = \int_{t_1}^{t_2} d(t)\dot{v}(t)\, dt. \tag{4.9}$$

Since v is affine linear on $[t_1, t_2]$, $\dot{v} \equiv (v(t_2) - v(t_1))/(t_2 - t_1)$ on (t_1, t_2). Combining this with (4.9) gives

$$(\Phi(v))(t_2) - (\Phi(v))(t_1) = \frac{v(t_2) - v(t_1)}{t_2 - t_1} \int_{t_1}^{t_2} d(t)\, dt = \eta(v(t_2) - v(t_1)),$$

where $\eta = 1/(t_2 - t_1) \int_{t_1}^{t_2} d(t)\, dt \in [0, \lambda]$. ∎

In the following section, we want to apply the discrete-time result of section 3 to the sampled-data low-gain integral control problem. Therefore, we need to investigate

the properties of the discretization of a given $\Phi^c \in \mathcal{C}(\lambda)$ obtained by standard hold and sampling operations. To this end, define the hold operator $H : F(\mathbb{Z}_+, \mathbb{R}) \to \mathcal{S}_\tau$ by

$$(Hu)(n\tau + t) = u(n) \quad \forall\, u \in F(\mathbb{Z}_+, \mathbb{R}), \quad \forall\, n \in \mathbb{Z}_+, \quad \forall\, t \in [0, \tau), \tag{4.10}$$

and the sampling operator $S : PC(\mathbb{R}_+, \mathbb{R}) \to F(\mathbb{Z}_+, \mathbb{R})$ by

$$(Sv)(n) = v(n\tau) \quad \forall\, v \in PC(\mathbb{R}_+, \mathbb{R}), \quad \forall\, n \in \mathbb{Z}_+, \tag{4.11}$$

where $\tau > 0$ denotes the sampling period.

Let $\Phi^c : C_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R}) \to C(\mathbb{R}_+, \mathbb{R})$ be a continuous-time hysteresis operator. We define a discrete-time operator $\Phi : F(\mathbb{Z}_+, \mathbb{R}) \to F(\mathbb{Z}_+, \mathbb{R})$ by

$$\Phi := S\, \Phi_e^c\, H, \tag{4.12}$$

where Φ_e^c denotes the extension of Φ^c to $NPC_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R})$ given by (4.5). We remark that as a simple consequence of the rate independence of Φ_e^c (cf. Proposition 4.3) the definition of Φ is independent of the choice of the sampling period τ. For $v \in \mathcal{S}_\tau$, we have $HSv = v$, and so, by Proposition 4.3, part (5),

$$\Phi_e^c\, H = H\, \Phi. \tag{4.13}$$

The proof of the following result can be found in [10].

Lemma 4.8. *Let $\Phi^c : C_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R}) \to C(\mathbb{R}_+, \mathbb{R})$ be a hysteresis operator, and define the operator $\Phi : F(\mathbb{Z}_+, \mathbb{R}) \to F(\mathbb{Z}_+, \mathbb{R})$ by* (4.12). *Then*

$$(\Phi(u))(n) = (\Phi^c(P_\tau u))(n\tau) \quad \forall\, u \in F(\mathbb{Z}_+, \mathbb{R}), \quad \forall\, n \in \mathbb{Z}_+, \tag{4.14}$$

and $\mathrm{NVS}\, \Phi = \mathrm{NVS}\, \Phi^c$.

The following proposition is the main result of this section.

Proposition 4.9. *Let $\Phi^c \in \mathcal{C}(\lambda)$, and define $\Phi : F(\mathbb{Z}_+, \mathbb{R}) \to F(\mathbb{Z}_+, \mathbb{R})$ by* (4.12). *Then $\Phi \in \mathcal{D}(\lambda)$.*

Proof. Note that (D1) (causality) follows from (4.14) and the causality of Φ^c, while (D2) follows from Proposition 4.2, part (1) and (4.14). Furthermore, let $u \in F(\mathbb{Z}_+, \mathbb{R})$, $n \in \mathbb{Z}_+$, and suppose that $u(n+1) \neq u(n)$. Then by (4.14) and Lemma 4.7 there exists a constant $\eta \in [0, \lambda]$ such that

$$\begin{aligned} (\Phi(u))(n+1) - (\Phi(u))(n) &= (\Phi^c(P_\tau\, u))((n+1)\tau) - (\Phi^c(P_\tau\, u))(n\tau) \\ &= \eta[(P_\tau\, u)((n+1)\tau) - (P_\tau\, u)(n\tau)] \\ &= \eta[u(n+1) - u(n)], \end{aligned}$$

and thus (D3) holds. Since $\mathrm{NVS}\, \Phi = \mathrm{NVS}\, \Phi^c$ (by Lemma 4.8) and Φ^c satisfies (C6), (D4) follows from an application of (4.14). Finally, to show that (D5) is satisfied, let $u \in F(\mathbb{Z}_+, \mathbb{R})$ be such that $\lim_{n\to\infty}(\Phi(u))(n)$ exists and

$$L := \lim_{n\to\infty} (\Phi(u))(n) \in \mathrm{int}\,(\mathrm{clos}\,(\mathrm{NVS}\, \Phi)). \tag{4.15}$$

By (4.14),

$$\lim_{n\to\infty} (\Phi^c(P_\tau\, u))(n\tau) = L. \tag{4.16}$$

Clearly $P_\tau u$ is monotone on $[n\tau,(n+1)\tau]$ for each $n \in \mathbb{Z}_+$, and therefore, by the fact that Φ^c satisfies (C4), $\Phi^c(P_\tau u)$ is monotone on $[n\tau,(n+1)\tau]$ for each $n \in \mathbb{Z}_+$. Combining this with (4.16) shows that

$$\lim_{t\to\infty}(\Phi^c(P_\tau u))(t) = L.$$

By Remark 4.5, $\mathrm{NVS}\Phi^c$ is an interval; since, by Lemma 4.8, $\mathrm{NVS}\Phi = \mathrm{NVS}\Phi^c$, it follows from (4.15) that $L \in \operatorname{int}(\mathrm{NVS}\,\Phi^c)$. Now Φ^c satisfies (C7), and so we may conclude that $P_\tau u$ and thus u are bounded. □

5. Sampled-data integral control in the presence of input hysteresis. In the following, the underlying linear system is assumed to be a single-input, single-output, continuous-time, regular system Σ (for details on regular systems see Weiss [25], [26], [27], [28]) with state-space X (a real Hilbert space) and with generating operators (A^c, B^c, C^c, D^c). This means, in particular, that A^c generates a strongly continuous semigroup $\mathbf{T} = (\mathbf{T}_t)_{t\ge 0}$, $C^c \in L(X_1, \mathbb{R})$ is an admissible observation operator for $\mathbf{T}$, and $B^c \in L(\mathbb{R}, X_{-1})$ is an admissible control operator for $\mathbf{T}$. Here X_1 denotes the domain of A^c endowed with the graph norm, and X_{-1} denotes the completion of X with respect to the norm $\|x\|_{-1} = \|(s_0 I - A^c)^{-1}x\|$, where s_0 is any fixed element in the resolvent set of A^c. The norm on X is denoted by $\|\cdot\|$, while $\|\cdot\|_1$ and $\|\cdot\|_{-1}$ denote the norms on X_1 and X_{-1}, respectively. Then $X_1 \hookrightarrow X \hookrightarrow X_{-1}$, and $\mathbf{T}$ restricts (respectively, extends) to a strongly continuous semigroup on X_1 (respectively, X_{-1}). The exponential growth constant

$$\omega(\mathbf{T}) := \lim_{t\to\infty}\frac{1}{t}\ln\|\mathbf{T}_t\|$$

is the same on all three spaces. The generator of $\mathbf{T}$ on X_{-1} is an extension of A^c to X (which is bounded as an operator from X to X_{-1}). We shall use the same symbol $\mathbf{T}$ (respectively, A^c) for the original semigroup (respectively, its generator) and the associated restrictions and extensions. With this convention, we may write $A^c \in L(X, X_{-1})$. Considered as a generator on X_{-1}, the domain of A^c is X.

We regard a regular system Σ as synonymous with its generating operators and simply write $\Sigma = (A^c, B^c, C^c, D^c)$. The regular system is said to be *exponentially stable* if the semigroup $\mathbf{T}$ is exponentially stable, that is, $\omega(\mathbf{T}) < 0$. The control operator B^c (respectively, observation operator C^c) is said to be *bounded* if $B^c \in L(\mathbb{R}, X)$ (respectively, $C^c \in L(X, \mathbb{R})$); otherwise, B^c (respectively, C^c) is said to be *unbounded*. In terms of the generating operators (A^c, B^c, C^c, D^c), the transfer function $\mathbf{G}^c(s)$ can be expressed as

$$\mathbf{G}^c(s) = C^c_L(sI - A^c)^{-1}B^c + D^c,$$

where C^c_L denotes the so-called Lebesgue extension of C^c. The transfer function $\mathbf{G}^c(s)$ is bounded and holomorphic in any half-plane $\operatorname{Re} s > \alpha$ with $\alpha > \omega(\mathbf{T})$. Moreover,

$$\lim_{s\to\infty,\, s\in\mathbb{R}} \mathbf{G}^c(s) = D^c.$$

For any $x_0 \in X$ and $u^c \in L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R})$, the state and output functions $x^c(\cdot)$ and $y^c(\cdot)$, respectively, satisfy the equations

$$\dot{x}^c(t) = A^c x^c(t) + B^c u^c(t), \quad x^c(0) = x_0, \tag{5.1a}$$

$$y^c(t) = C^c_L x^c(t) + D^c u^c(t) \tag{5.1b}$$

for almost all $t \geq 0$. The derivative on the left-hand side of (5.1a), of course, has to be understood in X_{-1}. In other words, if we consider the initial-value problem (5.1a) in the space X_{-1}, then for any $x_0 \in X$ and $u^c \in L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R})$, (5.1a) has unique strong solution (in the sense of Pazy [24, p. 109]) given by the variation of parameters formula

$$x^c(t) = \mathbf{T}_t x_0 + \int_0^t \mathbf{T}_{t-\tau} B^c u^c(\tau)\, d\tau. \tag{5.2}$$

For future reference we state the following lemma, the proof of which can be found in [15].

LEMMA 5.1. *Assume that* $\mathbf{T}$ *is exponentially stable and that* $B^c \in L(\mathbb{R}, X_{-1})$ *is an admissible control operator for* $\mathbf{T}$. *If* $u^c \in L^\infty(\mathbb{R}_+, \mathbb{R})$ *is such that* $\lim_{t\to\infty} u^c(t) = u_\infty$ *exists, then for all* $x_0 \in X$ *the state trajectory* x^c *given by* (5.2) *satisfies*

$$\lim_{t\to\infty} \|x^c(t) + (A^c)^{-1} B^c u_\infty\| = 0.$$

The aim in this section is to show that for an exponentially stable, regular, linear, infinite-dimensional, continuous-time, single-input, single-output plant with transfer function $\mathbf{G}^c(s)$, subject to a continuous-time dynamic input nonlinearity $\Phi^c \in \mathcal{C}(\lambda)$, the output $y^c(t)$ of the sampled-data closed-loop system, shown in Figure 2, converges to the reference value r as $t \to \infty$, provided that $\mathbf{G}^c(0) > 0$, r is feasible in some natural sense, and $k > 0$ is sufficiently small.

Let $u \in F(\mathbb{Z}_+, \mathbb{R})$, and apply the continuous-time signal

$$u^c = Hu \tag{5.3}$$

(where H is the standard hold operator defined in (4.10)) to the continuous-time system given by (5.1). Then the state $x^c(n\tau + t)$ satisfies

$$x^c(n\tau + t) = \mathbf{T}_t x^c(n\tau) + (\mathbf{T}_t - I)(A^c)^{-1} B^c u(n) \quad \forall\, n \in \mathbb{Z}_+, \quad \forall\, t \in [0, \tau). \tag{5.4}$$

Accordingly, we define $x : \mathbb{Z}_+ \to X$ by

$$x(n) = x^c(n\tau). \tag{5.5}$$

Clearly, $\mathbf{T}_\tau \in L(X)$ and $(\mathbf{T}_\tau - I)(A^c)^{-1} B^c \in L(\mathbb{R}, X)$ define appropriate state-space operators for the state evolution of the discretization of (5.1a). However, in general, regularity guarantees only that $y^c \in L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R})$ so that, even with piecewise constant input functions, standard sampling of the output is not defined. Moreover, even if the output function is continuous (in which case standard sampling is defined), in general the resulting discrete-time system will not have a bounded observation operator. We therefore distinguish two cases: bounded and unbounded continuous-time observation.

Bounded observation. Assume that $C^c = C^c_L \in L(X, \mathbb{R})$. If $x_0 \in X$ and u^c is given by (5.3), then the output y^c given by (5.1b) is piecewise continuous, the discontinuities being at $n\tau$. It is clear that y^c is right-continuous at $n\tau$ for all $n \in \mathbb{Z}_+$. We define

$$y := Sy^c \tag{5.6}$$

(where S is the standard sampling operator defined in (4.11)) and

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} := \begin{pmatrix} \mathbf{T}_\tau & (\mathbf{T}_\tau - I)(A^c)^{-1} B^c \\ C^c & D^c \end{pmatrix}. \tag{5.7}$$

The proof of the following proposition is an immediate consequence of Proposition 4.1 in [16].

PROPOSITION 5.2. *Suppose that* $\mathbf{T}_t$ *is exponentially stable and that the observation operator* C^c *is bounded. Let* $\tau > 0$ *and* $u \in F(\mathbb{Z}_+, \mathbb{R})$. *If* u^c *given by* (5.3) *is applied to* (5.1), *then* x *and* y *given by* (5.5) *and* (5.6), *respectively, satisfy* (3.1), *where* (A, B, C, D) *is given by* (5.7). *Moreover,* A *is power-stable, and, setting* $\mathbf{G}(z) = C(zI - A)^{-1}B + D$, *we have that*

$$\mathbf{G}(1) = C(I - A)^{-1}B + D = \mathbf{G}^c(0). \tag{5.8}$$

Let $\Phi^c \in \mathcal{C}(\lambda)$, and let Φ^c_e denote the extension of Φ^c to $NPC_{\mathrm{pm}}(\mathbb{R}_+, \mathbb{R})$ given by (4.5). Consider the continuous-time system (5.1) with continuous-time input nonlinearity Φ^c_e,

$$\dot{x}^c = A^c x^c + B^c \Phi^c_e(u^c), \quad x^c(0) = x_0 \in X, \tag{5.9a}$$

$$y^c = C^c_L x^c + D^c \Phi^c_e(u^c), \tag{5.9b}$$

controlled by the sampled-data integrator

$$u^c(t) = u(n) \quad \text{for } t \in [n\tau, (n+1)\tau), \quad n \in \mathbb{Z}_+, \tag{5.10a}$$

$$y(n) = y^c(n\tau), \quad n \in \mathbb{Z}_+, \tag{5.10b}$$

$$u(n+1) = u(n) + k(r - y(n)), \quad u(0) = u_0 \in \mathbb{R}, \quad n \in \mathbb{Z}_+. \tag{5.10c}$$

THEOREM 5.3. *Let* $\lambda > 0$. *Assume that* $\Phi^c \in \mathcal{C}(\lambda)$, C^c *is bounded,* $\mathbf{T}_t$ *is exponentially stable,* $\mathbf{G}^c(0) > 0$, *and* $r \in \mathbb{R}$ *is such that* $\tilde{r}^c := r/\mathbf{G}^c(0) \in \operatorname{clos}(\operatorname{NVS}\Phi^c)$. *Let* $K > 0$ *be defined by* (3.5), *where* $\mathbf{G}(z) = C(zI - A)^{-1}B + D$, *with* (A, B, C, D) *given by* (5.7). *Then, for all* $k \in (0, K/\lambda)$ *and all* $(x_0, u_0) \in X \times \mathbb{R}$, *the unique solution* $(x^c(\cdot), u^c(\cdot))$ *of the closed-loop system given by* (5.9) *and* (5.10) *satisfies the following:*

(1) $\lim_{t\to\infty}(\Phi^c_e(u^c))(t) = \tilde{r}^c$;
(2) $\lim_{t\to\infty} \|x^c(t) + (A^c)^{-1}B^c\tilde{r}^c\| = 0$;
(3) $\lim_{t\to\infty} y^c(t) = r$;
(4) *if* $\tilde{r}^c \in \operatorname{int}(\operatorname{NVS}\Phi^c)$, *then* u^c *is bounded.*

Proof. Let $(x^c(\cdot), u^c(\cdot))$ be the unique solution of the closed-loop system given by (5.9) and (5.10). Define $\Phi \in \mathcal{D}(\lambda)$ by (4.12), and so

$$(\Phi^c_e(u^c))(n\tau) = (\Phi^c_e(Hu))(n\tau) = (\Phi(u))(n) \quad \forall\, n \in \mathbb{Z}_+.$$

Note that by Lemma 4.8

$$\operatorname{NVS}\Phi = \operatorname{NVS}\Phi^c. \tag{5.11}$$

Defining $x \in F(\mathbb{Z}_+, \mathbb{R})$ by (5.5), it follows from Proposition 5.2 that (x, u) satisfies the closed-loop discrete-time equations (3.3), where (A, B, C, D) is given by (5.7). Therefore, using Theorem 3.1, Propositions 4.9 and 5.2, and (5.11), we see that for all $k \in (0, K/\lambda)$

$$\lim_{n\to\infty}(\Phi(u))(n) = \tilde{r}^c. \tag{5.12}$$

This implies that for all $k \in (0, K/\lambda)$, $\lim_{t\to\infty}(H(\Phi(u)))(t) = \tilde{r}^c$, and so by (4.13), $\lim_{t\to\infty}(\Phi^c_e(u^c))(t) = \tilde{r}^c$, which is statement (1). Statement (2) is a consequence of statement (1) and Lemma 5.1. Statement (3) follows easily from statements (1) and (2) and the boundedness of C^c. Finally, to prove statement (4), assume that $\tilde{r}^c \in \operatorname{int}\operatorname{NVS}\Phi^c$. Then, by (5.11), $\tilde{r}^c \in \operatorname{int}\operatorname{NVS}\Phi$. Boundedness of u and thus boundedness of u^c now follow immediately from (5.12) and the fact that (D5) holds for Φ (by Proposition 4.9). ∎

Unbounded observation. As mentioned earlier, in this case, we cannot define a sampled output via (5.6). Instead, we introduce a generalized sampling operation. In the following, let $w \in L^2([0,\tau],\mathbb{R})$ be a function satisfying the conditions

$$\text{(5.13)} \qquad \text{(i)} \quad \int_0^\tau w(t)\,dt = 1 \quad \text{and} \quad \text{(ii)} \quad \int_0^\tau w(t)\mathbf{T}_t\,x\,dt \in X_1 \quad \forall\, x \in X.$$

While condition (5.13) (ii) is difficult to check for general w, it is easy to show (using integration by parts) that (5.13) (ii) holds if there exists a partition $0 = t_0 < t_1 < \cdots < t_m = \tau$ such that $w|_{(t_{i-1},t_i)} \in W^{1,1}((t_{i-1},t_i),\mathbb{R})$ for $i = 1, 2, \ldots, m$.

We define a generalized sampling operation by

$$\text{(5.14)} \qquad y(n) = \int_0^\tau w(t)y^c(n\tau + t)\,dt \quad \forall\, n \in \mathbb{Z}_+.$$

Introducing the linear operator

$$L : X \to X_1, \quad x \mapsto \int_0^\tau w(t)\mathbf{T}_t\,x\,dt,$$

we define

$$\text{(5.15)} \qquad \begin{pmatrix} A & B \\ C & D \end{pmatrix} := \begin{pmatrix} \mathbf{T}_\tau & (\mathbf{T}_\tau - I)(A^c)^{-1}B^c \\ C^cL & C^cL(A^c)^{-1}B^c + \mathbf{G}^c(0) \end{pmatrix}.$$

The following result is an immediate consequence of Proposition 3.4 in [12].

PROPOSITION 5.4. *Suppose that $\mathbf{T}_t$ is exponentially stable. Let $\tau > 0$ and $u \in F(\mathbb{Z}_+,\mathbb{R})$. If u^c given by* (5.3) *is applied to* (5.1)*, then x and y given by* (5.5) *and* (5.14)*, respectively, satisfy* (3.1)*, where (A,B,C,D) is given by* (5.15)*. Moreover, A is power-stable, $C \in L(X,\mathbb{R})$, and, setting $\mathbf{G}(z) = C(zI - A)^{-1}B + D$,* (5.8) *is satisfied.*

Consider the following sampled-data low-gain controller for (5.9):

$$\text{(5.16a)} \qquad u^c(t) = u(n) \quad \text{for } t \in [n\tau,(n+1)\tau), \quad n \in \mathbb{Z}_+,$$

$$\text{(5.16b)} \qquad y(n) = \int_0^\tau w(t)y^c(n\tau + t)\,dt, \quad n \in \mathbb{Z}_+,$$

$$\text{(5.16c)} \qquad u(n+1) = u(n) + k(r - y(n)), \quad u(0) = u_0 \in \mathbb{R}, \quad n \in \mathbb{Z}_+.$$

THEOREM 5.5. *Let $\lambda > 0$. Assume that $\Phi^c \in \mathcal{C}(\lambda)$, $\mathfrak{L}^{-1}(\mathbf{G}^c)$ is a finite signed Borel measure on $\mathbb{R}_+$, $\mathbf{T}_t$ is exponentially stable, $\mathbf{G}^c(0) > 0$, and $r \in \mathbb{R}$ is such that $\tilde{r}^c := r/\mathbf{G}^c(0) \in \operatorname{clos}(\operatorname{NVS}\Phi^c)$. Let $K > 0$ be defined by* (3.5)*, where $\mathbf{G}(z) = C(zI - A)^{-1}B + D$, with (A,B,C,D) given by* (5.15)*. Then, for all $k \in (0, K/\lambda)$ and all $(x_0,u_0) \in X \times \mathbb{R}$, the unique solution $(x^c(\cdot),u^c(\cdot))$ of the closed-loop system given by* (5.9) *and* (5.16) *satisfies the following:*

(1) $\lim_{t\to\infty}(\Phi_e^c(u^c))(t) = \tilde{r}^c$;
(2) $\lim_{t\to\infty}\|x^c(t) + (A^c)^{-1}B^c\tilde{r}^c\| = 0$;
(3) $\lim_{t\to\infty}[r - y^c(t) + C_L^c\mathbf{T}_t x_0] = 0$;
(4) *if $\tilde{r}^c \in \operatorname{int}(\operatorname{NVS}\Phi^c)$, then u^c is bounded.*

REMARK 5.6. (1) *Since $C_L^c\mathbf{T}_t x_0$ converges exponentially to 0 as $t \to \infty$ for all $x_0 \in X_1$, it follows from statement* (3) *that the error $e^c(t) = r - y^c(t)$ converges to 0 for all $x_0 \in X_1$. If C^c is bounded, then this statement is true for all $x_0 \in X$. If C^c is unbounded and $x_0 \notin X_1$, then $e^c(t)$ does not necessarily converge to 0 as $t \to \infty$.*

However, it follows from the proof below that $e^c(t)$ *is small for large* t *in the sense that* $e^c(t) = e_1^c(t) + e_2^c(t)$*, where the function* e_1^c *is bounded with* $\lim_{t\to\infty} e_1^c(t) = 0$ *and the function* $t \mapsto e_2^c(t)\exp(\alpha t)$ *is in* $L^2(\mathbb{R}_+, \mathbb{R})$ *for some* $\alpha > 0$.

(2) *The assumption that* $\mathfrak{L}^{-1}(\mathbf{G}^c)$ *is a finite signed Borel measure on* $\mathbb{R}_+$ *is not very restrictive and seems to be satisfied in all practical examples of exponentially stable regular systems. In particular, this assumption is satisfied if* B^c *or* C^c *is bounded (see* [13, Lemma 2.3]*).*

Proof of Theorem 5.5. By using Proposition 5.4 instead of Proposition 5.2, all statements with the exception of (3) follow exactly as in the proof of Theorem 5.3. Due to the unboundedness of C^c, we cannot use statement (2) in order to show that $y^c(t)$ converges to r as $t \to \infty$. However, we have

$$y^c(t) = C_L^c \mathbf{T}_t x_0 + (\mathfrak{L}^{-1}(\mathbf{G}^c) \star \Phi_e^c(u^c))(t). \tag{5.17}$$

By assumption, $\mathfrak{L}^{-1}(\mathbf{G}^c)$ is a finite signed Borel measure, and since $\lim_{t\to\infty}(\Phi_e^c(u^c))(t) = \tilde{r}^c$ (by statement (1)), it follows from [5, Theorem 6.1, part (ii), p. 96] that

$$\lim_{t\to\infty} [\mathfrak{L}^{-1}(\mathbf{G}^c) \star \Phi_e^c(u^c)](t) = \mathbf{G}^c(0)\tilde{r}^c = r.$$

Combining this with (5.17) shows that statement (3) holds. □

6. Example: Sampled-data control of a diffusion process with output delay subject to input hysteresis. Consider a diffusion process (with diffusion coefficient $\kappa > 0$ and with Dirichlet boundary conditions) on the one-dimensional spatial domain $[0, 1]$, with scalar nonlinear pointwise control action (applied at point $x_1 \in (0, 1)$, via an operator $\Phi^c \in \mathcal{C}(\lambda)$) and delayed (delay $T \geq 0$) scalar observation generated by a spatial averaging of the delayed state over an ε-neighborhood of a point $x_2 \in (0, 1)$ with $x_2 > x_1$.

We formally write this single-input, single-output system as

$$\begin{aligned} z_t(t, x) &= \kappa z_{xx}(t, x) + \delta(x - x_1)(\Phi_e^c(u^c))(t)\,, \\ y^c(t) &= \frac{1}{2\varepsilon}\int_{x_2-\varepsilon}^{x_2+\varepsilon} z(t - T, x)\,dx, \end{aligned}$$

with boundary conditions

$$z(t, 0) = 0 = z(t, 1) \quad \forall\, t > 0.$$

For simplicity, we assume zero initial conditions

$$z(t, x) = 0 \quad \forall\,(t, x) \in [-T, 0] \times [0, 1].$$

With input $(\Phi_e^c(u^c))(\cdot)$ and output $y^c(\cdot)$, this example qualifies as a regular linear system with bounded observation and with transfer function given by

$$\mathbf{G}^c(s) = \frac{e^{-sT}\sinh\left(x_1\sqrt{s/\kappa}\right)\left[\cosh\left((1 - x_2 + \varepsilon)\sqrt{s/\kappa}\right) - \cosh\left((1 - x_2 - \varepsilon)\sqrt{s/\kappa}\right)\right]}{2\varepsilon s \sinh\sqrt{s/\kappa}}.$$

Since the observation is bounded, we may sample the output using the standard sampling operation given by (5.6). Further analysis (invoking application of the maximum principle for the heat equation, which, for brevity, we omit here) confirms the physical intuition that the impulse response $\mathfrak{L}^{-1}(\mathbf{G}^c)$ is nonnegative-valued. Consequently,

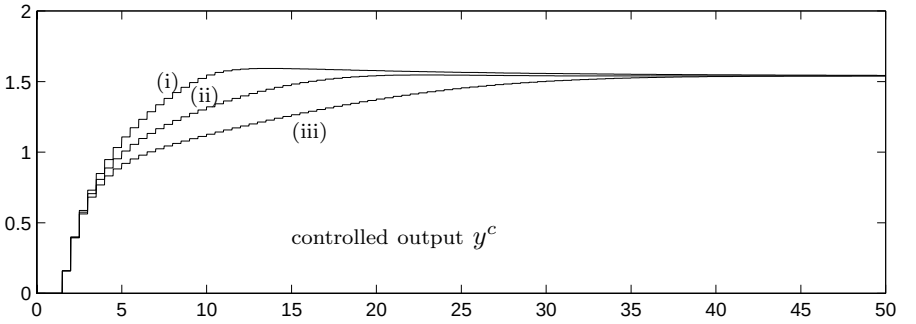


FIG. 8. *Controlled output.*

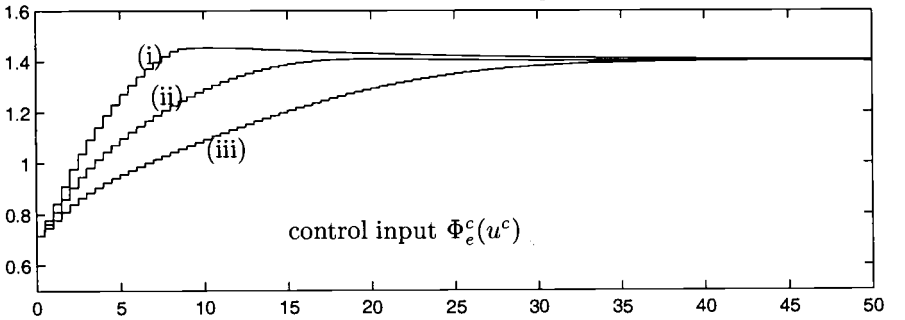


FIG. 9. *Control input.*

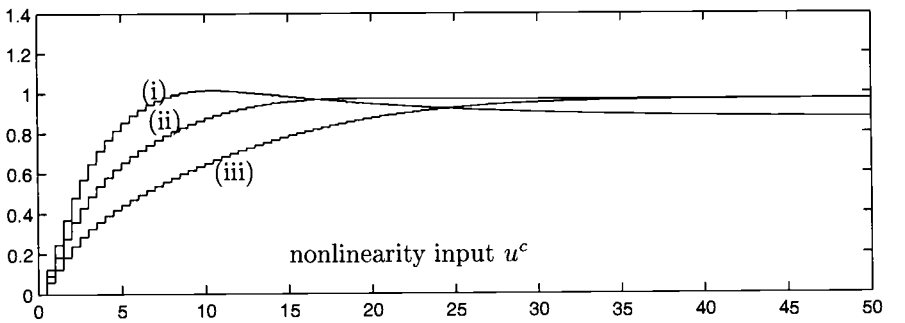


FIG. 10. *Input of relay hysteresis nonlinearity.*

the corresponding step-response is nondecreasing, and therefore we may apply a result by Özdemir and Townley [23] (see Remark 3.7 in [23]) to obtain the following lower bound for K:

$$K \geq \frac{1}{|(\mathbf{G}^c)'(0)|/\tau + 3\mathbf{G}^c(0)/2} =: K_L. \tag{6.1}$$

A simple calculation yields that

$$\mathbf{G}^c(0) = \frac{x_1(1-x_2)}{\kappa}, \quad (\mathbf{G}^c)'(0) = -\frac{x_1(1-x_2)(6T\kappa + 1 - \varepsilon^2 - x_1^2 - (1-x_2)^2)}{6\kappa^2},$$

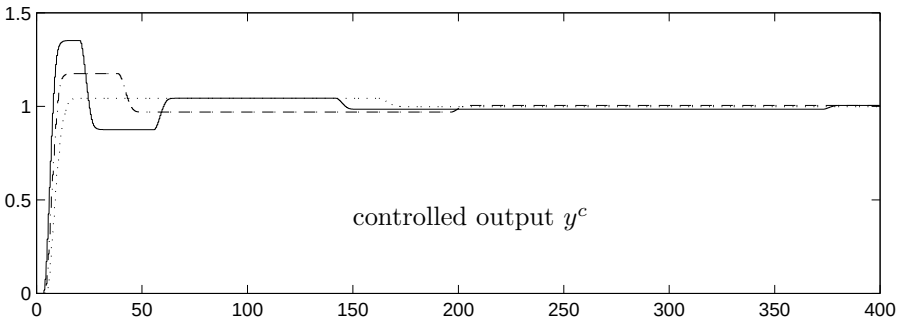


FIG. 11. *Controlled output.*

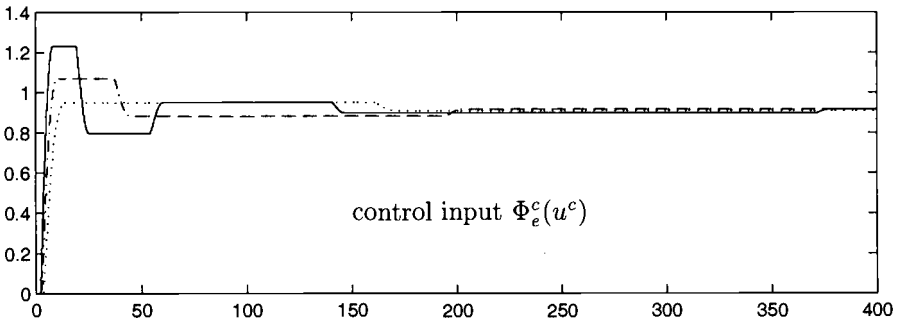


FIG. 12. *Control input.*

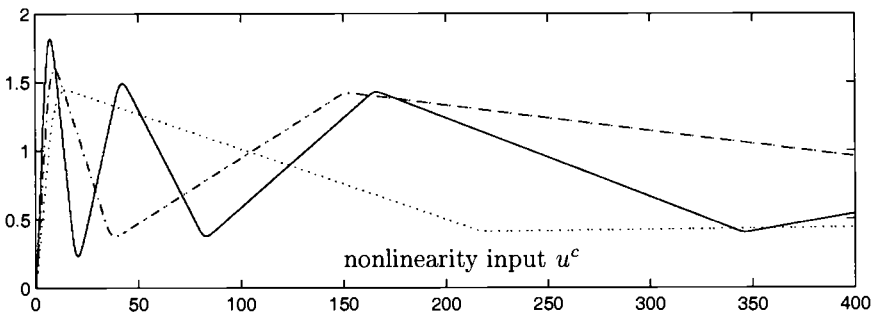


FIG. 13. *Input of backlash hysteresis nonlinearity.*

and therefore, using (6.1),

$$K \geq K_L = \frac{6\kappa^2\tau}{x_1(1-x_2)(6T\kappa + 1 - \varepsilon^2 - x_1^2 - (1-x_2)^2 + 9\kappa\tau)}.$$

By Theorem 5.3, for all $k \in (0, K_L/\lambda) \subset (0, K/\lambda)$, the sampled-data control (5.10) guarantees asymptotic tracking of all reference values r which are feasible in the sense that $r/\mathbf{G}^c(0) \in \operatorname{clos}(\operatorname{NVS}\Phi^c)$. For purposes of illustration, we adopt the following values:

$$\kappa = 0.1, \quad x_1 = \frac{1}{3}, \quad x_2 = \frac{2}{3}, \quad T = 1, \quad \tau = 0.5, \quad \varepsilon = 0.01.$$

For these specific values we obtain $K_L \approx 0.147$.

We consider relay and backlash hysteresis operators:

(a) Let $\Phi^c = \mathcal{R}_0$ be a relay hysteresis operator, where $a_1 = -1$, $a_2 = 1$, $\rho_1(v) = \sqrt{v + 1.1}$, and $\rho_2(v) = \sqrt{0.1} + \sqrt{2.1} - \sqrt{1.1 - v}$ (see section 4 for the definition and [11] for more details). Then $\Phi^c \in \mathcal{C}\,(1.6)$, and $\mathrm{NVS}\,\Phi^c = \mathrm{im}\,\rho_1 \cup \mathrm{im}\,\rho_2 = \mathbb{R}$. For reference value $r = 1.54$

$$\tilde{r}^c = \frac{r}{\mathbf{G}^c(0)} = \frac{r\kappa}{x_1(1 - x_2)} = 1.386 \in \mathrm{int}\,(\mathrm{NVS}\,\Phi^c).$$

In each of the following three cases of admissible controller gain,

$$\text{(i) } k = 0.08, \quad \text{(ii) } k = 0.06, \quad \text{(iii) } k = 0.04,$$

Figure 8 depicts the output behavior of the system under integral control, Figure 9 depicts the corresponding control input, and Figure 10 shows the input u^c of the relay nonlinearity. We see from Figure 10 that for (i), $\lim_{t\to\infty} u^c(t) = \rho_1^{-1}(\Phi_r^c)$, and for (ii) and (iii), $\lim_{t\to\infty} u^c(t) = \rho_2^{-1}(\Phi_r^c)$.

(b) Let $\Phi^c = \mathcal{B}_{0.5,\,0}$ be a backlash hysteresis operator (see section 4 for the definition and [11] for more details). Then $\Phi^c \in \mathcal{C}\,(1)$, and $\mathrm{NVS}\,\Phi^c = \mathbb{R}$. For reference value $r = 1$

$$\tilde{r}^c = \frac{r}{\mathbf{G}^c(0)} = \frac{r\kappa}{x_1(1 - x_2)} = 0.9 \in \mathrm{int}\,(\mathrm{NVS}\,\Phi^c).$$

In each of the following three cases of admissible controller gain,

$$\text{(i) } k = 0.145 \text{ (solid line)}, \quad \text{(ii) } k = 0.11 \text{ (dashdot line)}, \quad \text{(iii) } k = 0.08 \text{ (dotted line)},$$

Figure 11 depicts the output behavior of the system under sampled-data control, Figure 12 depicts the corresponding control input, and Figure 13 shows the input u^c of the backlash nonlinearity. We remark that the convergence of $u^c(t)$ as $t \to \infty$ is not guaranteed by Theorem 5.5, and in fact it seems that u^c does not converge in all three cases.

Figures 8–13 were generated using SIMULINK Simulation Software within MATLAB, wherein a truncated eigenfunction expansion, of order 10, was adopted to model the diffusion process.

Appendix. The infinite-dimensional discrete-time positive-real lemma. The following result is a version of the discrete-time infinite-dimensional positive-real lemma.

LEMMA A.1. *For a real Hilbert space X, let $A \in L(X)$, $B \in L(\mathbb{R}, X)$, $C \in L(X, \mathbb{R})$, and $D \in \mathbb{R}$, and set $\mathbf{G}(z) := C(zI - A)^{-1}B + D$. Assume that A is power-stable and*

$$\mathrm{Re}\,\mathbf{G}(e^{i\theta}) > 0 \quad \forall\, \theta \in [0, 2\pi).$$

Then there exist $P \in L(X)$, $P = P^* \geq 0$, $L \in L(\mathbb{R}, X)$, and $W \in \mathbb{R}$ such that

$$\begin{aligned} A^*PA - P &= -LL^*, \\ A^*PB &= C^* - WL, \\ W^2 &= 2D - B^*PB. \end{aligned}$$

Although Lemma A.1 should be well known, we were not able to locate it in the literature. Lemma A.1 can be obtained from Lemma A.2 (an infinite-dimensional

version of the continuous-time positive-real lemma stated below) combined with standard fractional transformation techniques (as used in [6] for the finite-dimensional case). For the sake of brevity, we omit the lengthy but straightforward details, which can be found in [21].

LEMMA A.2. *For a real Hilbert space X, let $A^c \in L(X)$, $B^c \in L(\mathbb{R}, X)$, $C^c \in L(X, \mathbb{R})$, and $D^c \in \mathbb{R}$; let $\sigma(A^c)$ denote the spectrum of A^c, and set $\mathbf{G}^c(s) := C^c(sI - A^c)^{-1}B^c + D^c$. Assume that $\sigma(A^c) \subset \{s \in \mathbb{C} \,|\, \operatorname{Re} s < 0\}$ and*

$$\operatorname{Re} \mathbf{G}^c(i\omega) > 0 \quad \forall\, \omega \in \mathbb{R} \cup \{\pm\infty\}. \tag{A.1}$$

Then there exist $P^c \in L(X)$, $P^c = (P^c)^* \geq 0$, $L^c \in L(\mathbb{R}, X)$, and $W^c > 0$ such that

$$P^c A^c + (A^c)^* P^c = -L^c (L^c)^*, \tag{A.2a}$$

$$P^c B^c = (C^c)^* - W^c L^c, \tag{A.2b}$$

$$2D^c = (W^c)^2. \tag{A.2c}$$

In a different form, Lemma A.2 is due to Yakubovich [30] (see also Wexler [29]). For completeness, we include a proof which is based on the positive-real Riccati equation theory developed in van Keulen [7].

Proof of Lemma A.2. By (A.1) we have that $D^c > 0$; defining $W^c := \sqrt{2D^c}$ gives (A.2c). Furthermore, again by (A.1), it follows from [7] (see Theorem 3.10 and Remark 3.14 in [7]) that there exists $Q^c \in L(X)$, $Q^c = (Q^c)^*$, such that

$$Q^c A^c + (A^c)^* Q^c = (1/W^c)^2 ((B^c)^* Q^c + C^c)^* \, ((B^c)^* Q^c + C^c).$$

Setting

$$P^c := -Q^c, \quad L^c := (1/W^c)(C^c - (B^c)^* P^c)^*$$

yields (A.2a) and (A.2b). Since $\sigma(A^c) \subset \{s \in \mathbb{C} \,|\, \operatorname{Re} s < 0\}$, it follows from (A.2a) by a routine argument that $P^c \geq 0$. ∎

REFERENCES

[1] M. BROKATE, *Hysteresis operators*, in Phase Transitions and Hysteresis, A. Visintin, ed., Springer-Verlag, Berlin, 1994, pp. 1–38.

[2] M. BROKATE AND J. SPREKELS, *Hysteresis and Phase Transitions*, Springer-Verlag, New York, 1996.

[3] G. W. M. COPPUS, S. L. SHA, AND R. K. WOOD, *Robust multivariable control of a binary distillation column*, IEE Proc., Part D, 130 (1983), pp. 201–208.

[4] E. J. DAVISON, *Multivariable tuning regulators: The feedforward and robust control of a general servomechanism problem*, IEEE Trans. Automat. Control, 21 (1976), pp. 35–47.

[5] G. GRIPENBERG, S.-O. LONDEN, AND O. STAFFANS, *Volterra Integral and Functional Equations*, Cambridge University Press, Cambridge, UK, 1990.

[6] L. HITZ AND B. D. O. ANDERSON, *Discrete positive-real functions and their application to system stability*, Proc. IEE, 116 (1969), pp. 153–155.

[7] B. A. M. VAN KEULEN, *H^∞-Control for Infinite-Dimensional Systems: A State-Space Approach*, Birkhäuser Boston, Boston, 1993.

[8] M. A. KRASNOSEL'SKIĬ AND A. V. POKROVSKIĬ, *Systems with Hysteresis*, Springer-Verlag, Berlin, 1989.

[9] H. LOGEMANN AND R. F. CURTAIN, *Absolute stability results for well-posed infinite-dimensional systems with applications to low-gain integral control*, ESAIM Control Optim. Calc. Var., 5 (2000), pp. 395–424.

[10] H. LOGEMANN AND A. D. MAWBY, *Extending Hysteresis Operators to Spaces of Piecewise Continuous Functions*, Mathematics Preprint 00/14, University of Bath, Bath, UK, 2000; also available online from http://www.maths.bath.ac.uk/MATHEMATICS/preprints.html and via anonymous ftp from ftp.maths.bath.ac.uk from the directory pub/preprints.

[11] H. Logemann and A. D. Mawby, *Low-gain integral control of infinite-dimensional regular linear systems subject to input hysteresis*, in Advances in Mathematical Systems Theory, F. Colonius et al., eds., Birkhäuser Boston, Boston, 2001, pp. 255–293.

[12] H. Logemann and E. P. Ryan, *Time-varying and adaptive discrete-time low-gain control of infinite-dimensional linear systems with input nonlinearities*, Math. Control Signals Systems, 13 (2000), pp. 293–317.

[13] H. Logemann and E. P. Ryan, *Time-varying and adaptive integral control of infinite-dimensional regular linear systems with input nonlinearities*, SIAM J. Control Optim., 38 (2000), pp. 1120–1144.

[14] H. Logemann, E. P. Ryan, and S. Townley, *Integral control of linear systems with actuator nonlinearities: Lower bounds for the maximal regulating gain*, IEEE Trans. Automat. Control, 44 (1999), pp. 1315–1319.

[15] H. Logemann, E. P. Ryan, and S. Townley, *Integral control of infinite-dimensional linear systems subject to input saturation*, SIAM J. Control Optim., 36 (1998), pp. 1940–1961.

[16] H. Logemann and S. Townley, *Discrete-time low-gain control of uncertain infinite-dimensional systems*, IEEE Trans. Automat. Control, 42 (1997), pp. 22–37.

[17] H. Logemann and S. Townley, *Low-gain control of uncertain regular linear systems*, SIAM J. Control Optim., 35 (1997), pp. 78–116.

[18] J. Lunze, *Robust Multivariable Feedback Control*, Prentice Hall, London, 1988.

[19] J. Lunze, *Experimentelle Erprobung einer Einstellregel für PI-Mehrgrößenregler bei der Herstellung von Ammoniumnitrat-Harnstoff-Lösung*, Messen Steuern Regeln, 30 (1987), pp. 2–6.

[20] J. W. Macki, P. Nistri, and P. Zecca, *Mathematical models for hysteresis*, SIAM Rev., 35 (1993), pp. 94–123.

[21] A. D. Mawby, *Integral Control of Infinite-Dimensional Linear Systems Subject to Input Hysteresis*, Ph.D. thesis, Department of Mathematical Sciences, University of Bath, Bath, UK, 2000.

[22] M. Morari, *Robust stability of systems with integral control*, IEEE Trans. Automat. Control, 30 (1985), pp. 574–577.

[23] N. Özdemir and S. Townley, *Robust Sampled-Data Integral Control by Variable and Adaptive Sampling*, Preprint, School of Mathematical Sciences, University of Exeter, Exeter, UK, 2000.

[24] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[25] G. Weiss, *Transfer functions of regular linear systems.* I. *Characterization of regularity*, Trans. Amer. Math. Soc., 342 (1994), pp. 827–854.

[26] G. Weiss, *Admissibility of unbounded control operators*, SIAM J. Control Optim., 27 (1989), pp. 527–545.

[27] G. Weiss, *Admissible observation operators for linear semigroups*, Israel J. Math., 65 (1989), pp. 17–43.

[28] G. Weiss, *The representation of regular linear systems on Hilbert spaces*, in Distributed Parameter System, F. Kappel, K. Kunisch, and W. Schappacher, eds., Birkhäuser Verlag, Basel, 1989, pp. 401–416.

[29] D. Wexler, *On frequency domain stability for evolution equations in Hilbert spaces via the algebraic Riccati equation*, SIAM J. Math. Anal., 11 (1980), pp. 969–983.

[30] V. A. Yakubovich, *A frequency domain theorem for the case in which the state and control spaces are Hilbert spaces with an application to some problems in the synthesis of optimal controls* I, Siberian Math. J., 15 (1974), pp. 457–476.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 1, pp. 141–163

ASYMPTOTIC PROPERTIES OF INPUT-OUTPUT OPERATORS NORM ASSOCIATED WITH SINGULARLY PERTURBED SYSTEMS WITH MULTIPLICATIVE WHITE NOISE*

VASILE DRAGAN†, TOADER MOROZAN†, AND PENG SHI‡

Abstract. In this paper, we study the problem of the asymptotic property of the norm of input-output operators related to a class of singularly perturbed stochastic linear systems. The system is under perturbation of multiplicative white noise. By using reduction order and boundary layer techniques, it is shown that the norm of the operator of the perturbed system is less than a given number γ when the small perturbation ε tends to zero if both the related norms of the reduced subsystem and the boundary layer subsystem are less than γ. Furthermore, a stabilizing robust controller is designed, which is independent of perturbation ε.

Key words. singularly perturbed system, input-output operator, stability, Riccati equation

AMS subject classifications. 93E03, 93E15, 93E20

PII. S0363012900369447

1. Introduction.

Singularly perturbed control systems (SPCS) evolving in discrete time scale arise in many applications as well as in the construction of the difference approximations of SPCS evolving in continuous time. A great amount of effort has been made on SPCS in the past three decades; see, e.g., [34] and the references therein. A popular approach adopted to deal with these systems is based on the so-called reduced technique [41]. The composite design based on separate designs for slow and fast subsystems has been systematically reviewed in [47]. Moreover, a number of averaging-type methods allowing us to treat general SPCS in continuous time were developed recently (see [1, 2, 16, 17, 18, 19, 24, 26, 52]). These methods are much more adaptable to the discrete time scale. For example, a full analogy between the averaging procedures in problems of optimal control of SPCS evolving in continuous and discrete times was established in [25]. The research on singularly perturbed systems in the H_∞ sense is of great practical importance and has attracted a lot of interest in the last few years; see, e.g., [28, 48, 51]. The state-space solution of the H_∞ control problem [3] was used to approximate the solution of singularly perturbed H_∞ control using slow and fast subproblems [48]; see also [51]. A sequential procedure was described in [28] to decompose the problem into slow and fast subproblems, and a composite compensator was provided. Recently, the H_∞-optimal control of singularly perturbed linear systems, under either perfect state measurements or imperfect state measurements, for both finite and infinite horizons, has been investigated in [42, 43] via a differential game theoretic approach. In the meantime, [4] studied the asymptotic expansions for game theoretic Riccati equations and showed how they may be used in singularly perturbed H_∞ control. More recently, [13, 14] considered a construction of high-order approximations to a controller that guarantees a desired performance level on the basis of the exact decomposition of the full-order Riccati

*Received by the editors March 24, 2000; accepted for publication (in revised form) December 1, 2001; published electronically April 26, 2002.
http://www.siam.org/journals/sicon/41-1/36944.html

†Institute of Mathematics of the Romanian Academy, P.O.Box 1-764, RO-70700, Bucharest, Romania (vdragan@stoilow.imar.ro, tmorozan@stoilow.imar.ro).

‡Land Operations Division, Defence Science and Technology Organisation, P.O. Box 1500, Edinburgh SA 5111, Australia (peng.shi@dsto.defence.gov.au).

equations to the reduced-order slow and fast equations. The problems of H_∞-norms and disturbance attenuation for systems with fast transients have been tackled by [5], and it has been shown that, for a singularly perturbed system, the H_∞ of the transfer function tends to the largest of the H_∞-norms for the boundary layer system and for the reduced slow model. More recently, a singularly perturbed zero-sum dynamic game with full information has been considered in [21], and it has been demonstrated that, when the singular perturbations parameter tends to zero, similar to singularly perturbed differential games [20], the upper (lower) value function of the dynamic game has a limit which coincides with a viscosity solution of a Hamilton–Jacobi–Isaacs-type equation. A composite linear controller has been designed in [49] based on the slow and fast problems such that both robust stability and a prescribed H_∞ performance for the full-order system are achieved, irrespective of the uncertainties. The problem of H_∞ control for singularly perturbed linear continuous-time systems with Markovian jump parameters has been studied in [12], in which the asymptotic structure of the composite mode-dependent controller is characterized.

It is worthwhile to mention that an important issue in the theory of SPCS is a justification of a so-called reduction technique approach (RTA). According to this approach, the fast variables are replaced by their steady states obtained with "frozen" slow variables and controls, and the slow dynamics is approximated by the corresponding reduced-order system. Although the RTA may fail to provide a proper approximation for the SPCS in a general case [17, 18, 19], its application was very successful in many important special cases (see [23, 35, 31, 32, 44, 45] and the references therein). In the differential game context, the efficiency of the RTA was established for singularly perturbed linear quadratic games in [22, 29] and for the singularly perturbed H^∞ problem with linear dynamics in [42, 43].

On the other hand, the control of stochastic systems with multiplicative white noise has received much attention in the past half century. For the results concerning the stability for stochastic systems with state-dependent noise, we refer readers to, for example, [7, 10, 36, 37] and the references therein. The linear quadratic problem associated to a linear stochastic system with multiplicative white noise was investigated; see, for example, [53, 54]. Robust stabilization for the above class of stochastic system was intensively studied in [9, 8] and the references therein.

In this paper, we investigate the asymptotic behavior of the input-output operator norm of the singularly perturbed linear continuous-time systems with multiplicative white noise. We consider the norms of both the slow/reduced subsystem and the fast/boundary layer subsystem. We demonstrate that when the perturbation ε goes to zero, then the input-output norm of the original system is less than the maximum of the norms corresponding to the both subsystems. A robust controller, which is free of perturbation ε, is designed to stabilizing the singularly perturbed stochastic systems.

2. Problem formulation. Let us consider the linear controlled system described by its differential equations

$$dx_1(t) = [A_{11}x_1(t) + A_{12}x_2(t) + B_1u(t)]dt + \sum_{j=1}^{N}[A_{11}^j x_1(t) + A_{12}^j x_2(t)]dw_j(t),$$

(2.1)

$$\varepsilon dx_2(t) = [A_{21}x_1(t) + A_{22}x_2(t) + B_2u(t)]dt + \varepsilon^\nu \sum_{j=1}^{N}[A_{21}^j x_1(t) + A_{22}^j x_2(t)]dw_j(t)$$

and the output

$$y(t) = C_1 x_1(t) + C_2 x_2(t) + Du(t), \tag{2.2}$$

where $x_i \in \mathbf{R}^{n_i}, i = 1, 2$, are state vectors, $u \in \mathbf{R}^m$ is the input vector, $A_{lk}, A_{lk}^j, B_l, C_k$, $l, k = 1, 2, j = 1, 2, \ldots, N, D$ are real matrices with corresponding dimensions, $\varepsilon > 0$ is a small parameter, and $\nu > \frac{1}{2}$. $w(t) = (w_1(t), w_2(t), \ldots, w_N(t)), t \geq 0$, is a standard Wiener process on a given probability space $(\Omega, \mathcal{F}, \mathcal{P})$.

We also consider the uncontrolled system associated to (2.1):

$$\begin{aligned} dx_1(t) &= [A_{11}x_1(t) + A_{12}x_2(t)]dt + \sum_{j=1}^{N} [A_{11}^j x_1(t) + A_{12}^j x_2(t)]dw_j(t), \\ \varepsilon dx_2(t) &= [A_{21}x_1(t) + A_{22}x_2(t)]dt + \varepsilon^\nu \sum_{j=1}^{N} [A_{21}^j x_1(t) + A_{22}^j x_2(t)]dw_j(t). \end{aligned} \tag{2.3}$$

For each $t_0 \geq 0$, we denote $\mathcal{X}(t_0)$ the set of n-dimensional random vectors ($n = n_1 + n_2$), which are $\mathcal{F}_{t_0}$-measurable, and

$$E[|x_0|^2] < \infty.$$

Obviously, $\mathbf{R}^n \subset \mathcal{X}(t_0)$. Let $\varepsilon > 0$ be fixed. Then, for each $t_0 \geq 0$ and $x_0 = \binom{x_{10}}{x_{20}} \in \mathcal{X}(t_0)$, system (2.3) has a unique solution

$$x(t, t_0, x_0, \varepsilon) = \begin{pmatrix} x_1(t, t_0, x_0, \varepsilon) \\ x_2(t, t_0, x_0, \varepsilon) \end{pmatrix},$$

which verifies the initial condition

$$x(t_0, t_0, x_0, \varepsilon) = x_0$$

(see, e.g., [15, 40, 54]). Moreover, the dependence $x_0 \to x(\cdot, t_0, x_0, \varepsilon)$ is linear.

Let $\Phi(t, t_0, \varepsilon) = (\Phi_1(t, t_0, \varepsilon) \quad \Phi_2(t, t_0, \varepsilon) \ldots \Phi_n(t, t_0, \varepsilon))$ be the $n \times n$ matrix having the columns $\Phi_j(t, t_0, \varepsilon) = x(t, t_0, e_j, \varepsilon), t \geq t_0 > 0, \varepsilon > 0, j = 1, 2, \ldots, n$. $e_j = (0, \ldots, 0, 1, 0, \ldots, 0)^*$ is a vector of the canonic basis of $\mathbf{R}^n$. Hence $t \to \Phi(t, t_0, \varepsilon)$ is a matrix solution of system (2.3) which verifies

$$\Phi(t_0, t_0, \varepsilon) = I_n.$$

From the uniqueness of the solution, it follows that

$$x(t, t_0, x_0, \varepsilon) = \Phi(t, t_0, \varepsilon)x_0, \quad (\forall) t \geq t_0 > 0, \ \varepsilon > 0, \ x_0 \in \mathcal{X}(t_0).$$

Throughout this paper, the matrix $\Phi(t, t_0, \varepsilon)$ will be termed as the *fundamental matrix solution* of system (2.3).

We recall the following definition, which will be referred to throughout the paper.

DEFINITION 2.1. *We say that the zero solution (that is, $x(t) = 0$ for all $t \in [0, \infty)$) of system* (2.3) *is exponentially stable in mean square (ESMS) (alternatively, system* (2.3) *generates a mean square stable evolution) if there exist constants $\alpha > 0$ and $\beta \geq 1$ such that $E|\Phi(t, t_0, \varepsilon)x_o|^2 \leq \beta e^{-\alpha(t-t_0)}|x_0|^2$ for all $t \geq t_0 \geq 0, x_o \in \mathbf{R}^{n_1+n_2}$.*

For each $t \geq 0$, we denote $\mathcal{F}_t \subset \mathcal{F}$ the smallest σ-algebra containing all sets $S \in \mathcal{F}$ with $\mathcal{P}(S) = 0$ and with respect to which all functions $w_j(s), 0 \leq s \leq t, 1 \leq j \leq N$ are measurable.

Let $L^2_w\{[0,\infty) \times \Omega, \mathbf{R}^d\}$ be the set of all functions $u \in L^2\{[0,\infty) \times \Omega, \mathbf{R}^d\}$ with the additional property that $u(t)$ are $\mathcal{F}_t$-measurable for all $t \geq 0$.

Since $\mathcal{F}_t$ contains all sets $S \in \mathcal{F}$ with $\mathcal{P}(S) = 0$, it can be easily proved that $L^2_w\{[0,\infty) \times \Omega, \mathbf{R}^d\}$ is closed in $L^2\{[0,\infty) \times \Omega, \mathbf{R}^d\}$, and hence it is itself a real Hilbert space with the inner product

$$\langle u, v \rangle_{L^2_w} = E \int_0^\infty u^*(t) v(t) dt.$$

Let us suppose that the zero solution of system (2.3) is ESMS.

If $u \in L^2_w\{[0,\infty) \times \Omega, \mathbf{R}^m\}$, we denote by

$$x(t, \varepsilon, u) = \begin{pmatrix} x_1(t, \varepsilon, u) \\ x_2(t, \varepsilon, u) \end{pmatrix}$$

the solution of system (2.1) with initial zero condition (i.e., $x_1(0, \varepsilon, u) = 0, x_2(0, \varepsilon, u) = 0$).

Applying Proposition 1 in [38], we deduce that $x(t, \varepsilon, u) \in L^2_w\{[0,\infty) \times \Omega, \mathbf{R}^{n_1+n_2}\}$ and $\lim_{t\to\infty} E|x(t, \varepsilon, u)|^2 = 0$.

Moreover, there exists a $\gamma > 0$ such that

$$||x(\cdot, \varepsilon, u)||^2 = E \int_0^\infty |x(t, \varepsilon, u)|^2 dt \leq \gamma^2 E \int_0^\infty |u(t)|^2 dt = \gamma^2 ||u||^2. \tag{2.4}$$

Thus the linear operator $\mathbf{T}_\varepsilon$ is well defined. $\mathbf{T}_\varepsilon : L^2_w([0,\infty) \times \Omega, \mathbf{R}^m) \to L^2_w([0,\infty) \times \Omega, \mathbf{R}^p)$ by

$$(\mathbf{T}_\varepsilon u)(t) = (C_1 \quad C_2) x(t, \varepsilon, u) + Du(t) \quad \forall t \geq 0. \tag{2.5}$$

From (2.4), we have that T_ε is a linear bounded operator and it will be called an input-output operator associated to system (2.1)–(2.2), and system (2.1)–(2.2) will be termed the "state-space realization" of the operator $\mathbf{T}_\varepsilon$.

REMARK 2.1. *When system* (2.1) *is a deterministic one (i.e.,* $A^j_{lk} = 0, l, k = 1, 2, j = 1, 2, \ldots, N$*), the transfer matrix function* G *is the frequency domain version of the input-output operator.*

However, in stochastic framework, we are not able to define, in a standard way, a transfer matrix function associated to system (2.1)–(2.2)*, and, therefore, we consider input-output operators instead of transfer matrices even if the coefficients of the given system are time invariant.*

Our aim in this paper is to investigate the asymptotic behavior of the norm of operator $\mathbf{T}_\varepsilon$ when ε approaches zero.

We shall extend the results in [5, 50] to the case of controlled systems described by Ito differential equations of type (2.1)–(2.2).

To this end, we associate to system (2.1)–(2.2) two systems with lower dimensions not depending upon the small parameter ε, namely the reduced subsystem and the boundary layer subsystem.

Setting $\varepsilon = 0$ and assuming that A_{22} is an invertible matrix, we can associate the following reduced subsystem to system (2.1)–(2.2):

$$dx_1(t) = [A_r x_1(t) + B_r u(t)]dt + \sum_{j=1}^{N} A_r^j x_1(t) dw_j(t),$$

$$y_r(t) = C_r x_1(t) + D_r u(t), \tag{2.6}$$

where

$$A_r = A_{11} - A_{12}A_{22}^{-1}A_{21},\ A_r^j = A_{11}^j - A_{12}^j A_{22}^{-1}A_{21},$$
$$B_r = B_1 - A_{12}A_{22}^{-1}A_{21},\ C_r = C_1 - C_2 A_{22}^{-1}A_{21},$$
$$D_r = D - C_2 A_{22}^{-1}B_2.$$

The unforced reduced subsystem is as follows:

$$dx_1(t) = A_r x_1(t)dt + \sum_{j=1}^{N} A_r^j x_1(t) dw_j(t). \tag{2.7}$$

If the zero solution of system (2.7) is ESMS, then we can associate to system (2.6) the corresponding input-output operator

$$\mathbf{T}_r : L_w^2[(0,\infty)\times\Omega, \mathbf{R}^m] \to L_w^2[(0,\infty)\times\Omega, \mathbf{R}^p]$$

by

$$(\mathbf{T}_r u)(t) = C_r x_1(t,u) + D_r u(t),$$

where $x_1(t,u)$ is a solution of the system (2.6) with initial condition $x_1(0,u) = 0$.

For the given system (2.1), we associate the so-called boundary layer system, described by

$$x'(\tau) = A_{22}x_2(\tau) + B_2 u(\tau),$$

$$y(\tau) = C_2 x_2(\tau) + Du(\tau), \tag{2.8}$$

with $\tau = \frac{t}{\varepsilon}$, which is a deterministic system.

The transfer matrix function corresponding to system (2.8) is

$$G_f(s) = C_2(sI_{n_2} - A_{22})^{-1}B_2 + D.$$

Again, our objective in this paper is to prove that $||\mathbf{T}_\varepsilon||$ tends to $\max\{||\mathbf{T}_r||, ||G_f||_\infty\}$.

3. Some preliminary results.

3.1. A Klimusev–Krasovski-type result. In this subsection, we extend the results of Klimusev–Krakovski [30] to the singularly perturbed systems of Ito differential equations of type (2.3).

THEOREM 3.1. *Assume that all eigenvalues of matrix A_{22} are located in the half plane of* $\mathrm{Re}(s) < 0$ *and the zero solution of the reduced subsystem* (2.7) *is ESMS.*

Then there exists an $\varepsilon_0 > 0$ *such that, for any arbitrary* $\varepsilon \in (0, \varepsilon_0]$, *the zero solution of the full system* (2.3) *is ESMS.*

Moreover, if

$$\begin{pmatrix} \Phi_{11}(t,t_0,\varepsilon) & \Phi_{12}(t,t_0,\varepsilon) \\ \Phi_{21}(t,t_0,\varepsilon) & \Phi_{22}(t,t_0,\varepsilon) \end{pmatrix}$$

is the partition of the fundamental matrix solution $\Phi(t,t_0,\varepsilon)$ *of system* (2.3), *then the following estimates hold:*

$$E|\Phi_{11}(t,t_0,\varepsilon)|^2 \le \beta_1 e^{-\alpha_1(t-t_0)},$$

$$E|\Phi_{12}(t,t_0,\varepsilon)|^2 \le \beta_1 \varepsilon e^{-\alpha_1(t-t_0)},$$

$$E|\Phi_{21}(t,t_0,\varepsilon)|^2 \le \beta_2 e^{-\alpha_1(t-t_0)},$$

$$E|\Phi_{22}(t,t_0,\varepsilon)|^2 \le \beta_2 (e^{\frac{-\alpha_2(t-t_0)}{\varepsilon}} + \varepsilon e^{-\alpha_1(t-t_0)})$$

for all $t \ge t_0 \ge 0$, $\varepsilon \in (0, \varepsilon_0]$, *where* $\alpha_i > 0$ *and* $\beta_i \ge 1, i = 1, 2$, *are independent of* ε, t, t_0.

Proof. Let us consider the nonlinear algebraic equation

$$A_{21} + A_{22}S_{21} = \varepsilon S_{21}(A_{11} + A_{12}S_{21}) \tag{3.1}$$

with unknown $S_{21} \in \mathbf{R}^{n_2 \times n_1}$.

By an implicit function argument (see, for example, [6]), we deduce that there exist an $\varepsilon_1 > 0$ and an analytic function $S_{21} : (-\varepsilon_1, \varepsilon_1) \to \mathbf{R}^{n_2 \times n_1}$ having the asymptotic structure $S_{21}(\varepsilon) = -A_{22}^{-1}A_{21} + O(\varepsilon)$, which solves (3.1).

Also, we consider the linear algebraic equation

$$S_{12}(A_{22} - \varepsilon S_{21}(\varepsilon)A_{12}) - \varepsilon(A_{11} + A_{12}S_{21}(\varepsilon)) = A_{12} \tag{3.2}$$

with unknown $S_{12} \in \mathbf{R}^{n_1 \times n_2}$. By an implicit function argument again, we conclude that there exist an $\varepsilon_2 \in (o, \varepsilon_1]$ and an analytic function $S_{12} : (-\varepsilon_2, \varepsilon_2) \to \mathbf{R}^{n_1 \times n_2}$ having the asymptotic structure $S_{12}(\varepsilon) = A_{12}A_{22}^{-1} + O(\varepsilon)$, which solves (3.2).

If $\binom{x_1(t,\varepsilon)}{x_2(t,\varepsilon)}$ is a solution of system (2.3), we define

$$\begin{pmatrix} \xi_1(t,\varepsilon) \\ \xi_2(t,\varepsilon) \end{pmatrix} = \begin{pmatrix} I_{n_1} & -\varepsilon S_{12}(\varepsilon) \\ 0 & I_{n_2} \end{pmatrix} \begin{pmatrix} I_{n_1} & 0 \\ -S_{21}(\varepsilon) & I_{n_2} \end{pmatrix} \begin{pmatrix} x_1(t,\varepsilon) \\ x_2(t,\varepsilon) \end{pmatrix},$$
$$t \ge 0, \quad \varepsilon \in (0, \varepsilon_2].$$

It is easy to see that the process $\binom{\xi_1(t,\varepsilon)}{\xi_2(t,\varepsilon)}, t \ge 0$, is a solution of the following system of Ito differential equations:

$$\begin{aligned} d\xi_1(t) &= \hat{A}_{11}(\varepsilon)\xi_1(t)dt + \sum_{j=1}^{N}[\hat{A}_{11}^j(\varepsilon)\xi_1(t) + \hat{A}_{12}^j(\varepsilon)\xi_2(t)]dw_j(t), \\ \varepsilon d\xi_2(t) &= \hat{A}_{22}(\varepsilon)\xi_2(t)dt + \varepsilon^{\nu_1}\sum_{j=1}^{N}[\hat{A}_{21}^j(\varepsilon)\xi_1(t) + \hat{A}_{22}^j(\varepsilon)\xi_2(t)]dw_j(t), \end{aligned} \tag{3.3}$$

where

$$\hat{A}_1(\varepsilon) = A_r + O(\varepsilon), \quad \hat{A}_{11}^j(\varepsilon) = A_r^j + O(\varepsilon),$$

$$\hat{A}_{12}^j(\varepsilon) = A_{12}^j + O(\varepsilon), \quad \hat{A}_{22}(\varepsilon) = A_{22} + O(\varepsilon),$$

$$\hat{A}_{21}^j(\varepsilon) = A_{21}^j - A_{22}^j A_{22}^{-1} A_{21} + O(\varepsilon),$$

$$\hat{A}_{22}^j(\varepsilon) = A_{22}^j + O(\varepsilon), \qquad \nu_1 = \min\{\nu, 1\}.$$

Applying Lemma A.1 (in the appendix) for

$$f_0(t) = 0, f_j(t) = \hat{A}_{12}^j(\varepsilon)\xi_2(t,\varepsilon), \quad j = 1, 2, \ldots, N,$$

we obtain from the first equation of system (3.3) that

$$E|\xi_1(t,\varepsilon)|^2 \le \hat{\beta}_1 \left[e^{-\hat{\alpha}_1(t-t_0)} E|\xi_1(t_0)|^2 + \sum_{j=1}^{N} \int_{t_0}^{t} e^{-\hat{\alpha}_1(t-s)} E|\xi_2(s,\varepsilon)|^2 ds \right] \tag{3.4}$$

for all $t \ge t_0 > 0$, where $\hat{\alpha}_1 > 0$ and $\hat{\beta}_1 \ge 1$.

Using Lemma A.2 (in the appendix) for

$$g_0(t) = 0, \; g_i(t) = \hat{A}_{21}^j(\varepsilon)\xi_1(t,\varepsilon), \quad j = 1, 2, \ldots, N,$$

one obtains

$$E|\xi_2(t,\varepsilon)|^2 \le \hat{\beta}_2 \left[e^{-\alpha_2 \frac{(t-t_0)}{\varepsilon}} E|\xi_2(t_0,\varepsilon)|^2 + \varepsilon^{2\nu_1 - 2} \sum_{j=1}^{N} \int_{t_0}^{t} e^{-\alpha_2 \frac{t-s}{\varepsilon}} E|\xi_1(s,\varepsilon)|^2 ds \right]. \tag{3.5}$$

Replacing (3.5) with (3.4) and changing the order of integration we deduce

$$E|\xi_1(t,\varepsilon)|^2 \le \hat{\beta}_1 e^{-\hat{\alpha}_1(t-t_0)} E|\xi_1(t_0,\varepsilon)|^2 + \hat{\beta}_1\hat{\beta}_2 \int_{t_0}^{t} e^{-\hat{\alpha}_1(t-s)} e^{-\alpha_2 \frac{s-t_o}{\varepsilon}} ds$$
$$\cdot E|\xi_2(t_0,\varepsilon)|^2 + \varepsilon^{2\nu_1-2}\hat{\beta}_1\hat{\beta}_2 \sum_{j=1}^{N} \int_{t_0}^{t} \int_{s}^{t} e^{-\hat{\alpha}_1(t-\sigma)} e^{-\alpha_2 \frac{\sigma-s}{\varepsilon}} d\sigma E|\xi_1(s,\varepsilon)|^2 ds.$$

Hence

$$E|\xi_1(t,\varepsilon)|^2 \le \hat{\beta}_3 e^{-\hat{\alpha}_1(t-t_0)} [E|\xi_1(t_0,\varepsilon)|^2 + \varepsilon E|\xi_2(t_0,\varepsilon)|^2]$$
$$+ \hat{\beta}_4 \varepsilon^{2\nu_1 - 1} \sum_{j=1}^{N} \int_{t_0}^{t} e^{-\hat{\alpha}_1(t-s)} E|\xi_1(s,\varepsilon)|^2 ds$$

with $\hat{\beta}_3 > 0$ and $\hat{\beta}_4 > 0$. By using the Gronwall lemma [6], we obtain from the above inequality

$$E|\xi_1(t,\varepsilon)|^2 \le \hat{\beta}_3 e^{-(\hat{\alpha}_1 - \varepsilon^{2\nu_1} N\hat{\beta}_4)(t-t_0)} [E|\xi_1(t_0,\varepsilon)|^2 + \varepsilon E|\xi_2(t_0,\varepsilon)|^2] \quad \forall t \ge t_0 \ge 0.$$

Thus, for all $\varepsilon \in (0, \frac{\frac{\hat{\alpha}_1}{N} - \alpha_1}{\hat{\beta}_4})$, we conclude that

$$E|\xi_1(t,\varepsilon)|^2 \le \hat{\beta}_3 e^{-N\alpha_1(t-t_0)}[E|\xi_1(t_0,\varepsilon)|^2 + \varepsilon E|\xi_2(t_0,\varepsilon)|^2] \quad \forall t \ge t_0 \ge 0. \tag{3.6}$$

Now, taking into account (3.6), we deduce from (3.5) that

$$\begin{aligned} E|\xi_2(t,\varepsilon)|^2 &\le \hat{\beta}_5 e^{-\alpha_2 \frac{(t-t_0)}{\varepsilon}} E|\xi_2(t_0,\varepsilon)|^2 \\ &\quad + \hat{\beta}_6 \varepsilon^{2\nu_1 - 1} e^{-\alpha_1(t-t_0)}[E|\xi_1(t_0,\varepsilon)|^2 + \varepsilon E|\xi_2(t_0,\varepsilon)|^2]. \end{aligned} \tag{3.7}$$

Since $x_1(t,\varepsilon) = \xi_1(t,\varepsilon) + \varepsilon S_{12}(\varepsilon)\xi_2(t,\varepsilon)$, one has

$$E|x_1(t,\varepsilon)|^2 \le \beta_1 e^{-\alpha_1(t-t_0)}[E|x_1(t_0,\varepsilon)|^2 + \varepsilon E|x_2(t_0,\varepsilon)|^2] \tag{3.8}$$

for some $\beta_1 \ge 1$ not depending upon ε.

On the other hand, since $x_2(t,\varepsilon) = \xi_2(t,\varepsilon) + S_{21}(\varepsilon)x_1(t,\varepsilon)$, and taking into account (3.7) and (3.8), we have

$$E|x_2(t,\varepsilon)|^2 \le \beta_2[e^{-\alpha_2(t-t_0)}E|\xi_2(t_0,\varepsilon)|^2 + e^{-\alpha_1(t-t_0)}(E|x_1(t_0,\varepsilon)|^2 + \varepsilon E|x_2(t_0,\varepsilon)|^2)] \tag{3.9}$$

for all $t \ge t_0 \ge 0$, for some $\beta_2 \ge 1$ not depending upon ε.

The inequalities (3.8) and (3.9) ensure the exponential stability of the zero solution of the full system (2.3). The estimates for the block components $\Phi_{ij}(t,t_0,\varepsilon)$ of the fundamental matrix $\Phi(t,t_0,\varepsilon)$ are directly obtained from (3.8) and (3.9). Thus the proof is finished. □

It should be noted that Theorem 3.1 can be used to design a stabilizing state feedback controller for the full system of type (2.1) based on separately designing the stabilizing feedback gain for two subsystems of lower dimension which do not depend on the small parameter ε. To be more precise, let us consider the controlled system

$$\begin{aligned} dx_1(t) &= [A_{11}x_1(t) + A_{12}x_2(t) + B_1u(t)]dt + \sum_{j=1}^{N}[A_{11}^j x_1(t) + A_{12}^j x_2(t)]dw_j(t), \\ \varepsilon dx_2(t) &= [A_{21}x_1(t) + A_{22}x_2(t) + B_2u(t)]dt + \varepsilon^\nu \sum_{j=1}^{N}[A_{21}^j x_1(t) + A_{22}^j x_2(t)]dw_j(t), \end{aligned} \tag{3.10}$$

where $\varepsilon > 0, \nu > \frac{1}{2}, u \in \mathbf{R}^m$ are the vectors of the control parameters. If we take $u(t) = F_1x_1(t) + F_2x_2(t)$, the resulting closed loop system is

$$\begin{aligned} dx_1(t) &= [(A_{11} + B_1F_1)x_1(t) + (A_{12} + B_1F_2)x_2(t)]dt \\ &\quad + \sum_{j=1}^{N}[A_{11}^j x_1(t) + A_{12}^j x_2(t)]dw_j(t), \\ \varepsilon dx_2(t) &= [(A_{21} + B_2F_1)x_1(t) + (A_{22} + B_2F_2)x_2(t)]dt \\ &\quad + \varepsilon^\nu \sum_{j=1}^{N}[A_{21}^j x_1(t) + A_{22}^j x_2(t)]dw_j(t). \end{aligned} \tag{3.11}$$

Applying Theorem 3.1 to system (3.11), we may construct a stabilizing control

$$u(t) = F_1x_1(t) + F_2x_2(t)$$

for system (3.10). This is done in the following corollary.

COROLLARY 3.2. *Assume that A_{22} is invertible and $\tilde{F} \in \mathbf{R}^{m\times n_1}$ is chosen such that the zero solution of the closed loop system*

$$dx_1(t) = [A_r + B_r\tilde{F}]x_1(t)dt + \sum_{j=1}^{N}[A_r^j + B_r^j\tilde{F}]x_1(t)dw_j(t) \tag{3.12}$$

is ESMS and $F_2 \in \mathbf{R}^{m\times n_2}$ is chosen such that $A_{22}+B_2F_2$ is a Hurwitz matrix (A_r, B_r being defined as in the case of system (2.6)). Set $F_1 = (I_m + F_2A_{22}^{-1}B_2)\tilde{F} + F_2A_{22}^{-1}A_{21}$; then the control

$$u(t) = F_1x_1(t) + F_2x_2(t) \tag{3.13}$$

stabilizes system (3.10) for arbitrary $\varepsilon > 0$ small enough. (It means that there exists an $\varepsilon_0 > 0$ such that the zero solution of the corresponding closed loop system (3.11) is ESMS for any arbitrary $\varepsilon \in (0, \varepsilon_0)$.)

Proof. The closed loop system obtained by coupling the control (3.13) to system (3.10) is

$$\begin{aligned} dx_1(t) &= [(A_{11} + B_1F_1)x_1(t) + (A_{12} + B_1F_2)x_2(t)]dt \\ &\quad + \sum_{j=1}^{N}[A_{11}^jx_1(t) + A_{12}^jx_2(t)]dw_j(t), \\ \varepsilon dx_2(t) &= [(A_{21} + B_2F_1)x_1(t) + (A_{22} + B_2F_2)x_2(t)]dt \\ &\quad + \varepsilon_\nu \sum_{j=1}^{N}[A_{21}^jx_1(t) + A_{22}^jx_2(t)]dw_j(t). \end{aligned} \tag{3.14}$$

Setting $\varepsilon = 0$ in (3.14), we obtain the reduced subsystem

$$\begin{aligned} dx_1(t) &= [A_{11} + B_1F_1 - (A_{12} + B_1F_2)(A_{22} + B_2F_2)^{-1}(A_{21} + B_2F_1)]x_1(t)dt \\ &\quad + \sum_{j=1}^{N}[A_{11}^j - A_{12}^j(A_{22} + B_2F_2)^{-1}(A_{21} + B_2F_1)]x_1(t)dw_j(t). \end{aligned} \tag{3.15}$$

After some algebraic computation, similar to that in the deterministic framework (e.g., [33, 6]), we can show that (3.15) is just (3.12). Now the conclusion is immediate from Theorem 3.1. □

REMARK 3.1. *The designing of the matrix gain $\tilde{F}$ in order to guarantee the ESMS of the zero solution of the system (3.12) may be done by using, for instance, the result of Theorem 1 in [46], and the designing of the matrix gain F_2 may be done by using any known methods for the stabilization of a linear time invariant finite dimensional system in the deterministic framework.*

3.2. Representation formula for the stabilizing solution of a class of algebraic Riccati equations. Let us consider the controlled system described by the differential Ito equation

$$dx(t) = [Ax(t) + Bu(t)]dt + \sum_{i=1}^{N} A^ix(t)dw_i(t) \tag{3.16}$$

and the output

$$y = Cx(t) + Du(t). \tag{3.17}$$

If the uncontrolled system

$$dx(t) = Ax(t)dt + \sum_{i=1}^{N} A^i x(t)dw_i(t) \tag{3.18}$$

associated to (3.16) generates an exponentially stable evolution, then system (3.16)–(3.17) defines an input-output operator $T : L_w^2\{(0,\infty) \times \Omega, \mathbf{R}^m\} \to: L_w^2\{(0,\infty) \times \Omega, \mathbf{R}^p\}$ by $(\mathbf{T}u)(t) = Cx_u(t) + Du(t), t \geq 0$, where $x_u(\cdot) \in L_w^2\{(0,\infty) \times \Omega, \mathbf{R}^n\}$ is the solution of (3.14) with initial condition $x_u(0) = 0$.

The result of the following theorem gives a stochastic version of the well-known bounded real lemma.

THEOREM 3.3 (see [38, 39]). *Under the considered assumptions, the following statements are equivalent:*

(i) *the uncontrolled system* (3.18) *defines a mean square exponentially stable evolution, and the input-output operator* $\mathbf{T}$ *associated to* (3.16)–(3.17) *verifies*

$$||\mathbf{T}|| < \gamma;$$

(ii) $D^*D < \gamma^2 I_m$, *and the algebraic Riccati-type equation*

$$\begin{aligned} A^*X + XA + \sum_{i=1}^{N} (A^i)^* X A^i \\ + (XB + C^*D)(\gamma^2 I_m - D^*D)^{-1}(B^*X + D^*C) + C^*C = 0 \end{aligned} \tag{3.19}$$

has a unique stabilizing solution $\tilde{X} = \tilde{X}^* \geq 0$.

Recall that $\tilde{X}$ is a "stabilizing solution" of (3.19) if the system

$$dx(t) = (A + B\tilde{F})x(t)dt + \sum_{i=1}^{N} A^i x(t)dw_i(t)$$

defines a mean square exponentially stable evolution with

$$\tilde{F} = (\gamma^2 I_m - D^*D)^{-1}(B^*\tilde{X} + D^*C).$$

Combining the results in Proposition 3 and Theorem 1 in Morozan [38, 39], we obtain a useful representation formula of the stabilizing solution $\tilde{X}$ of the Riccati-type equation (3.19).

THEOREM 3.4. *Suppose that the statement* (i) *in Theorem* 3.3 *holds. Then the stabilizing solution of* (3.19) *has the representation formula*

$$\tilde{X} = \mathcal{P}^0 - \mathcal{P}\mathcal{R}_\gamma^{-1}\mathcal{P}^* \tag{3.20}$$

with $\mathcal{P}^0 : \mathbf{R}^n \to \mathbf{R}^n, \mathcal{P} : L_w^2((0,\infty) \times \Omega, \mathbf{R}^m) \to \mathbf{R}^n, \mathcal{R}_\gamma : L_w^2((0,\infty) \times \Omega, \mathbf{R}^m) \to L_w^2((o,\infty) \times \Omega, \mathbf{R}^m)$ *by*

$$\mathcal{P}^0 = E\int_0^\infty \Phi^*(t,0)C^*C\Phi(t,0)dt,$$

$$\mathcal{P}u = E\int_0^\infty \Phi^*(t,0)C^*\left[C\int_0^t \Phi(t,s)Bu(s)ds + Du(t)\right]dt$$
$$= E\int_0^\infty \Phi^*(t,0)C^*(Tu)(t)dt \ \forall u \in L^2_w((0,\infty)\times\Omega,\mathbf{R}^m),$$
$$\mathcal{R}_\gamma = T^*T - \gamma^2 I,$$

where $\Phi(t,0)$ *is the fundamental random matrix associated to the uncontrolled system* (3.18).

Proof. Consider the quadratic cost function

$$J_\gamma(x_0,u) = E\int_0^\infty [|y_u(t)|^2 - \gamma^2|u(t)|^2]dt.$$

Since $||\mathbf{T}|| < \gamma$, applying Proposition 3 in [38], we deduce that $u \to J_\gamma(x_0,u)$ is a continuous and a concave function.

Setting $y_u(t) = C\Phi(t,0)x_0 + [Tu](t)$, we can easily show that

$$\max\{J_\gamma(x_0,u)|u \in L^2_w((0,\infty)\times\Omega,\mathbf{R}^m)\} = x_0^*(\mathcal{P}^0 - \mathcal{P}\mathcal{R}_\gamma^{-1}\mathcal{P}^*)x_0. \tag{3.21}$$

On the other hand, applying Theorem 1 in [38], we get

$$J_\gamma(x_0,u) = x_0^*\tilde{X}x_0 - E\int_0^\infty (u(t)-\tilde{F}x(t))^*(\gamma^2 I_m - D^*D)(u(t)-\tilde{F}x(t))dt$$
$$\forall u \in L^2_w((0,\infty)\times\Omega,\mathbf{R}^m).$$

Hence one has

$$\max\{J_\gamma(x_0,u)|u \in L^2_w((0,\infty)\times\Omega,\mathbf{R}^m)\} = x_0^*\tilde{X}x_0. \tag{3.22}$$

Thus, from (3.21) and (3.22), (3.20) follows, and the proof is complete. □

4. Main results. We make the following assumptions throughout this section and thereafter:

$\mathbf{H_1}$. The linear unforced system (2.7) defines an exponentially stable evolution in mean square.

$\mathbf{H_2}$. All eigenvalues of matrix A_{22} are located in the half plane of $\mathrm{Re}(s) < 0$.

PROPOSITION 4.1. *Under assumptions* $\mathbf{H_1}$–$\mathbf{H_2}$, *if there exists a sequence* $\{\varepsilon_k\}_{k\in\mathbf{N}}$ *such that* $\lim_{k\to\infty}\varepsilon_k = 0$ *and* $||\mathbf{T}_{\varepsilon_k}|| < \gamma$ *for all* $k \in \mathbf{N}$, *then* $\gamma \geq \max\{||\mathbf{T}_r||, ||\mathbf{G}_f||_\infty\}$.

Proof. Let $\gamma' < \gamma$ be fixed. Then $||\mathbf{T}_{\varepsilon_k}|| < \gamma'$ for all $k \in \mathbf{N}$. Set

$$A(\varepsilon_k) = \begin{pmatrix} A_{11} & A_{12} \\ \varepsilon^{-1}A_{21} & \varepsilon^{-1}A_{22} \end{pmatrix},\ A^i(\varepsilon_k) = \begin{pmatrix} A^i_{11} & A^i_{12} \\ \varepsilon_k^{\nu-1}A^i_{21} & \varepsilon_k^{\nu-1}A^i_{22} \end{pmatrix},$$
$$B(\varepsilon_k) = \begin{pmatrix} B_1 \\ \varepsilon_k^{-1}B_2 \end{pmatrix}.$$

Applying Theorem 3.1, we have that for all k large enough, the system

$$dx(t) = A(\varepsilon_k)x(t)dt + \sum_{j=1}^N A^j(\varepsilon_k)x(t)dw_j(t)$$

defines an exponentially stable evolution. Using statements (i) and (ii) in Theorem 3.3, we conclude that the algebraic Riccati-type equation

$$A^*(\varepsilon_k)X + XA(\varepsilon_k) + \sum_{j=1}^{N} A^{j*}(\varepsilon_k)XA^j(\varepsilon_k) + (XB(\varepsilon_k) + C^*D)(\gamma^2 I_m - D^*D)^{-1} \cdot (B^*(\varepsilon_k)X + D^*C) + C^*C = 0 \tag{4.1}$$

has a stabilizing solution $X(\varepsilon_k) = X^*(\varepsilon_k) \geq 0$.

From Theorem 3.4, we deduce that the stabilizing solution $X(\varepsilon_k, \gamma)$ of (4.1) has the representation formula

$$X(\varepsilon_k, \gamma') = \mathcal{P}^0(\varepsilon_k) - \mathcal{P}(\varepsilon_k)\mathcal{R}^{-1}_{\gamma'}(\varepsilon_k)\mathcal{P}^*(\varepsilon_k).$$

Let

$$\begin{pmatrix} X_{11}(\varepsilon_k, \gamma') & X_{12}(\varepsilon_k, \gamma') \\ X^*_{12}(\varepsilon_k, \gamma') & X_{22}(\varepsilon_k, \gamma') \end{pmatrix}$$

be the partition of the solution $X(\varepsilon_k, \gamma')$ compatible with the partition of the coefficient matrix of system (2.1).

Taking into account the estimations in Theorem 3.1, we shall deduce estimations for $X_{ij}(\varepsilon_k, \gamma')$.

First, notice that

$$-\mathcal{R}_{\gamma'}(\varepsilon_k) = (\gamma'^2 - \gamma^2)I - \mathcal{R}_{\gamma}(\varepsilon_k) \geq (\gamma'^2 - \gamma^2)I.$$

Since $\mathcal{R}_{\gamma'}(\varepsilon_k) = \mathcal{R}^*_{\gamma'}(\varepsilon_k)$, where $\mathcal{R}^*_{\gamma'}(\varepsilon_k)$ stands for the adjoint operator of $\mathcal{R}_{\gamma'}(\varepsilon_k)$, we obtain that $\mathcal{R}_{\gamma'}(\varepsilon_k)$ is invertible on $L^2_w((0,\infty) \times \Omega, \mathbf{R}^m)$ with bounded inverse. Moreover, we have

$$||\mathcal{R}^{-1}_{\gamma'}(\varepsilon_k)|| \leq (\gamma'^2 - \gamma^2)^{-1/2}$$

for all $k \in \mathbf{N}$ large enough. On the other hand, we may write

$$\mathcal{P}(\varepsilon_k) = \begin{pmatrix} \mathcal{P}_1(\varepsilon_k) \\ \mathcal{P}_2(\varepsilon_k) \end{pmatrix},\ \mathcal{P}_j(\varepsilon_k) : L^2_w((0,\infty) \times \Omega; \mathbf{R}^m) \to \mathbf{R}^{n_j}, \quad j = 1, 2,$$

$$\mathcal{P}_j(\varepsilon_k)u = E\int_0^\infty [\Phi^*_{1j}(t, 0, \varepsilon_k)C_1^* + \Phi^*_{2j}(t, 0, \varepsilon_k)C_2^*]y(t, \varepsilon_k)dt,$$

where

$$y(t, \varepsilon_k) = \begin{pmatrix} C_1 & C_2 \end{pmatrix} \int_0^t \Phi(t, 0, \varepsilon_k)B(\varepsilon_k)u(s)ds + Du(t).$$

Using the estimates in Theorem 3.1, we obtain that there exist constants $c_1 > 0$ and $c_2 > 0$ not depending on k but possibly depending on γ', such that

$$\|\mathcal{P}_1(\varepsilon_k)\| \leq c_1, \quad \|\mathcal{P}_2(\varepsilon_k)\| \leq c_2\varepsilon_k, \ \ (\forall)\ k \geq 1.$$

With these inequalities, we may conclude that the stabilizing solution $X(\varepsilon_k)$ of (4.1) has the following asymptotic structure:

$$X(\varepsilon_k) = \begin{pmatrix} X_{11}(\varepsilon_k) & \varepsilon_k X_{12}(\varepsilon_k) \\ \varepsilon_k X^*_{12}(\varepsilon_k) & \varepsilon_k X_{22}(\varepsilon_k) \end{pmatrix}$$

with $|X_{ij}(\varepsilon_k)| \le c_3(\gamma')$, where $c_3(\gamma') > 0$ does not depend on k.

We define

$$\begin{aligned} F(\varepsilon_k) &= (F_1(\varepsilon_k) \qquad F_2(\varepsilon_k)) \\ &= (\gamma'^2 I_m - D^*D)^{-1}\left[\begin{pmatrix} B_1^* & \dfrac{1}{\varepsilon_k}B_2^* \end{pmatrix} X(\varepsilon_k) + D^*(C_1 \qquad C_2)\right]. \end{aligned}$$

With this notation, we have that (4.1) is equivalent to a system with unknowns $X_{11}(\varepsilon_k), X_{12}(\varepsilon_k), X_{22}(\varepsilon_k), F_1(\varepsilon_k), F_2(\varepsilon_k)$, and thus the rest of the proof can be carried out along the same lines as that in the deterministic framework [4]. □

Based on the result in Proposition 4.1, we can easily show the following theorem.

THEOREM 4.2. *Under assumptions* $\mathbf{H_1}$–$\mathbf{H_2}$*, the norm of the input-output operator defined by system* (2.1)–(2.2) *verifies*

$$\liminf_{\varepsilon \searrow 0} \|\mathbf{T}_\varepsilon\| \ge \max\{\|\mathbf{T}_r\|, \|\mathbf{G}_f\|_\infty\}.$$

In the remainder of this section, we will consider a controlled system described by

$$\begin{aligned} dx_1(t) &= [A_{11}x_1(t) + A_{12}x_2(t) + B_1u(t)]dt \\ &\quad + \sum_{i=1}^{N}[A_{11}^i x_1(t) + \varepsilon^\mu A_{12}^i x_2(t)]dw_i(t), \\ \varepsilon dx_2(t) &= [A_{21}x_1(t) + A_{22}x_2(t) + B_2u(t)]dt \\ &\quad + \varepsilon^\nu \sum_{i=1}^{N}[A_{21}^i x_1(t) + A_{22}^i x_2(t)]dw_i(t) \end{aligned} \tag{4.2}$$

and the output

$$y(t) = C_1x_1(t) + C_2x_2(t) + Du(t), \tag{4.3}$$

where $x_i, A_{ij}, B_i, C_j, D, \nu$, and ε are as in system (2.1)–(2.2) and $\mu > 0$ is independent of ε.

When $\mu = 0$, system (4.2)–(4.3) is just the system (2.1)–(2.2).

In this case (when $\mu > 0$), the reduced subsystem obtained by setting $\varepsilon = 0$ in (4.2)–(4.3) is (2.6), where $A_r^j = A_{11}^j, j = 1, 2, \ldots, N$. The corresponding boundary layer subsystem is (2.8).

It is easy to see that the result of Theorem 4.2 also holds for the input-output operator corresponding to system (4.2)–(4.3). In the case of system (4.2)–(4.3), we may derive a result concerning the superior limit of the norm of operator $\mathbf{T}_\varepsilon$ defined by this system.

First, we present the following result.

PROPOSITION 4.3. *Assume that* $\mathbf{H_1}$–$\mathbf{H_2}$ *hold for system* (4.2)–(4.3). *Then, for all*

$$\gamma > \max\{\|\mathbf{T}_r\|, \|\mathbf{G}_f\|_\infty\}, \tag{4.4}$$

there exists an $\varepsilon_0(\gamma) > 0$ *such that for arbitrary* $\varepsilon \in (0, \varepsilon_0(\gamma))$ *we have* $\|\mathbf{T}_\varepsilon\| < \gamma$.

REMARK 4.1. *In Proposition* 4.3, *that assumption* $\mathbf{H_1}$ *holds for system* (4.2)–(4.3) *means that the assumption is true when the control input* $u(t)$ *is zero (i.e., unforced situation).*

Proof. We show that, under the considered assumptions, there exists an $\varepsilon_0(\gamma) > 0$ such that for arbitrary $\varepsilon \in (0, \varepsilon_0(\gamma))$ the algebraic Riccati equation

$$\begin{aligned} &A^*(\varepsilon)X + XA(\varepsilon) + \sum_{i=1}^{N}(A^i)^*(\varepsilon)XA^i(\varepsilon) \\ &\quad + (XB(\varepsilon) + C^*D)(\gamma^2 I_m - D^*D)^{-1}(B^*(\varepsilon)X + D^*C) + C^*C = 0 \end{aligned} \tag{4.5}$$

has a stabilizing solution $\tilde{X}(\varepsilon) = \tilde{X}^*(\varepsilon) \geq 0$. Then, applying the results in [38, 39], we may conclude that $\|\mathbf{T}_\varepsilon\| < \gamma, \varepsilon \in (0, \varepsilon_0(\gamma))$.

If (4.4) holds, then $\|\mathbf{G}_f\|_\infty < \gamma$, and we get $D^*D < \gamma^2 I_m$, and thus (4.5) is well defined.

Set

$$F = (F_1 \qquad F_2) = (\gamma^2 I_m - D^*D)^{-1}(B^*(\varepsilon)X + D^*C),\ X = \begin{pmatrix} X_{11} & \varepsilon X_{12} \\ \varepsilon X_{12}^* & \varepsilon X_{22} \end{pmatrix}.$$

We obtain that (4.5) is equivalent to the following system:

$$\begin{aligned}
&B_1^*X_{11} + B_2^*X_{12}^* + D^*C_1 = (\gamma^2 I_m - D^*D)F_1, \\
&\varepsilon B_1^*X_{12} + B_2^*X_{22} + D^*C_2 = (\gamma^2 I_m - D^*D)F_2, \\
&A_{11}^*X_{11} + A_{21}^*X_{12}^* + X11A_{11} + X_{12}A_{21} + \sum_{i=1}^{N}[(A_{11}^i)^*X_{11}A_{11}^i \\
&\qquad + \varepsilon^\nu (A_{21}^i)^*X_{12}^*A_{11}^i + \varepsilon^\nu (A_{11}^i)^*X_{12}A_{21}^i + \varepsilon^{2\nu-1}(A_{21}^i)^*X_{22}A_{21}^i] \\
&\qquad + F_1^*(\gamma^2 I_m - D^*D)F_1 + C_1^*C_1 = 0, \\
&\varepsilon A_{11}^*X_{12} + A_{21}^*X_{22} + X_{11}A_{12} + X_{12}A_{22} + \sum_{i=1}^{N}[\varepsilon^\mu (A_{11}^i)^*X_{11}A_{12}^i \\
&\qquad + \varepsilon^{\mu+\nu}(A_{21}^i)^*X_{12}^*A_{12}^i + \varepsilon^\nu (A_{11}^i)^*X_{12}A_{22}^i + \varepsilon^{2\nu-1}(A_{21}^i)^*X_{22}A_{22}^i] \\
&\qquad + F_1^*(\gamma^2 I_m - D^*D)F_2 + C_1^*C_2 = 0, \\
&\varepsilon A_{12}^*X_{12} + A_{22}^*X_{22} + \varepsilon X_{12}^*A_{12} + X_{22}A_{22} + \sum_{i=1}^{N}[\varepsilon^{2\mu}(A_{12}^i)^*X_{11}A_{12}^i + \varepsilon^{\mu+\nu}(A_{22}^i)^*X_{12}^*A_{12}^i \\
&\qquad + \varepsilon^{\mu+\nu}(A_{12}^i)^*X_{12}A_{22}^i + \varepsilon^{2\nu-1}(A_{22}^i)^*X_{22}A_{22}^i] \\
&\qquad + F_2^*(\gamma^2 I_m - D^*D)F_2 + C_2^*C_2 = 0
\end{aligned}$$

with unknowns $F_j \in \mathbf{R}^{m\times n_j}, X_{ij} \in \mathbf{R}^{n_i\times n_j}, X_{ii} = X_{ii}^*, i, j = 1, 2$. We associate the following implicit function problem:

$$\begin{aligned}
&B_1^*X_{11} + B_2^*X_{12}^* + D^*C_1 = (\gamma^2 I_m - D^*D)F_1, \\
&\eta_1 B_1^*X_{12} + B_2^*X_{22} + D^*C_2 = (\gamma^2 I_m - D^*D)F_2, \\
&A_{11}^*X_{11} + A_{21}^*X_{12}^* + X11A_{11} + X_{12}A_{21} + \sum_{i=1}^{N}[(A_{11}^i)^*X_{11}A_{11}^i + \eta_2(A_{21}^i)^*X_{12}^*A_{11}^i \\
&\qquad + \eta_2(A_{11}^i)^*X_{12}A_{21}^i + \eta_3(A_{21}^i)^*X_{22}A_{21}^i] \\
&\qquad + F_1^*(\gamma^2 I_m - D^*D)F_1 + C_1^*C_1 = 0, \\
&\eta_1 A_{11}^*X_{12} + A_{21}^*X_{22} + X_{11}A_{12} + X_{12}A_{22} + \sum_{i=1}^{N}[\eta_4(A_{11}^i)^*X_{11}A_{12}^i + \eta_2\eta_4(A_{21}^i)^*X_{12}^*A_{12}^i
\end{aligned} \tag{4.7}$$

$$
\begin{aligned}
&+ \eta_2 (A_{11}^i)^* X_{12} A_{22}^i + \eta_3 (A_{21}^i)^* X_{22} A_{22}^i] \\
&+ F_1^* (\gamma^2 I_m - D^* D) F_2 + C_1^* C_2 = 0, \\
\eta_1 A_{12}^* X_{12} + A_{22}^* X_{22} + \eta_1 X_{12}^* A_{12} + X_{22} A_{22} + \sum_{i=1}^{N} &[\eta_4^2 (A_{12}^i)^* X_{11} A_{12}^i \\
&+ \eta_2 \eta_4 (A_{22}^i)^* X_{12}^* A_{12}^i + \eta_2 \eta_4 (A_{12}^i)^* X_{12} A_{22}^i + \eta_3 (A_{22}^i)^* X_{22} A_{22}^i] \\
&+ F_2^* (\gamma^2 I_m - D^* D) F_2 + C_2^* C_2 = 0
\end{aligned}
$$

with unknowns $F_j, X_{ij}, i, j = 1, 2$ and the free parameters $\eta_1, \eta_2, \eta_3, \eta_4$.

The rest of the proof follows in the same way as in the deterministic case [4]. □

From Proposition 4.3, we get the following theorem.

THEOREM 4.4. *Under* $\mathbf{H_1}$–$\mathbf{H_2}$, *the norm of the input-output operator defined by system* (4.2)–(4.3) *verifies*

$$\limsup_{\varepsilon \searrow 0} \|\mathbf{T}_\varepsilon\| \leq \max\{\|\mathbf{T}_r\|, \|\mathbf{G}_f\|_\infty\}.$$

Combining the results of Theorem 4.2 and Theorem 4.4, we have the following result.

THEOREM 4.5. *Under assumptions* $\mathbf{H_1}$–$\mathbf{H_2}$, *the norm of the input-output operator defined by system* (4.2)–(4.3) *verifies*

$$\lim_{\varepsilon \searrow 0} \|\mathbf{T}_\varepsilon\| = \max\{\|\mathbf{T}_r\|, \|\mathbf{G}_f\|_\infty\}.$$

5. Robust stabilization via an ε-independent controller for a class of singularly perturbed linear stochastic systems. In Corollary 3.2, it has been shown how the Klimusev–Krasovski-type result of Theorem 3.1 could be used to design a composite stabilizing control for a singularly perturbed linear stochastic system.

In this section, we will show how the results of Theorems 3.1 and 4.5 can be used to analyze the robustness properties of an ε-independent stabilizing controller for a linear stochastic system with two time scales.

Consider the system

$$
\begin{aligned}
dx_1(t) &= [A_{11} x_1(t) + A_{12} x_2(t) + B_1^1 v(t) + B_2^1 u(t)] dt \\
&\quad + \sum_{k=1}^{N} [A_{11}^k x_1(t) + \varepsilon^\mu A_{12}^k x_2(t)] dw_k(t), \\
\varepsilon dx_2(t) &= [A_{21} x_1(t) + A_{22} x_2(t) + B_1^2 v(t) + B_2^2 u(t)] \\
&\quad + \varepsilon^\nu \sum_{k=1}^{N} [A_{21}^k x_1(t) + A_{22}^k x_2(t)] dw_k(t), \qquad (5.1) \\
y_1(t) &= C_{11} x_1(t) + C_{12} x_2(t) + D_{11} v(t) + D_{12} u(t), \\
dy_2(t) &= [C_{21} x_1(t) + C_{22} x_2(t) + D_{21} v(t)] dt + \sum_{k=1}^{N} [C_{21}^k x_1(t) + \varepsilon^\lambda C_{22}^k x_2(t)] dw_k(t),
\end{aligned}
$$

having control inputs $u \in \mathbf{R}^{m_2}$, exogenous disturbances $v \in \mathbf{R}^{m_1}$ (which are supposed to be adapted stochastic processes), controlled outputs $y_1 \in \mathbf{R}^{p_1}$, and measured outputs $y_2 \in \mathbf{R}^{p_2}$. $\mu > 0, \lambda > 0$, and ν and ε are as in (2.1)–(2.2).

If $A_{ij}^k = 0, i, j = 1, 2, k = 1, 2, \ldots, N$, then (5.1) is a deterministic system having the measured output subjected to some random disturbances modeled by multiplicative white noise. In this section, we investigate the problem of designing of a robust stabilizing controller with a single time scale. We shall use both strict proper controllers of the form

$$\begin{aligned} dx_c(t) &= A_c x_c(t)dt + B_c dy_2(t), \\ u_2(t) &= C_c x_c(t) \end{aligned} \tag{5.2}$$

and proper controllers of the form

$$\begin{aligned} dx_c(t) &= A_c x_c(t)dt + B_c dy_2(t), \\ du_2 &= C_c x_c(t)dt + D_c dy_2(t). \end{aligned} \tag{5.3}$$

The closed loop system by coupling the controller (5.2) to system (5.1) is

$$\begin{aligned} dx_1(t) &= [A_{11}x_1(t) + B_2^1 C_c x_c(t) + A_{12}x_2(t) + B_1^1 v(t)]dt \\ &\quad + \sum_{k=1}^{N} [A_{11}^k x_1(t) + \varepsilon^\mu A_{12}^k x_2(t)]dw_k(t), \\ dx_c(t) &= [B_c C_{21} x_1(t) + A_c x_c(t) + B_c C_{22} x_2(t) + B_c D_{21} v(t)]dt \\ &\quad + \sum_{k=1}^{N} [B_c C_{21}^k x_1(t) + \varepsilon^\lambda B_c C_{22}^k x_2(t)]dw_k(t), \\ \varepsilon dx_2(t) &= [A_{21}x_1(t) + B_{22}C_c x_c(t) + A_{22}x_2(t) + B_1^2 v(t)]dt \\ &\quad + \varepsilon^\nu \sum_{k=1}^{N} [A_{21}^k x_1(t) + A_{22}^k x_2(t)]dw_k(t), \\ y_1(t) &= C_{11}x_1(t) + D_{12}C_c x_c(t) + C_{12}x_2(t) + D_{11}v(t). \end{aligned} \tag{5.4}$$

This system may be rewritten in the following form:

$$\begin{aligned} d\xi(t) &= [\hat{A}_{11}\xi(t) + \hat{A}_{12}x_2(t) + \hat{B}_1 v(t)]dt + \sum_{k=1}^{N} [\hat{A}_{11}^k \xi(t) + \hat{A}_{12}^k x_2(t)]dw_k(t), \\ \varepsilon dx_2(t) &= [\hat{A}_{21}\xi(t) + \hat{A}_{22}x_2(t) + \hat{B}_2 v(t)]dt + \varepsilon^\nu \sum_{k=1}^{N} [\hat{A}_{21}^k \xi(t) + \hat{A}_{22}^k x_2(t)]dw_k(t), \\ y_1(t) &= \hat{C}_1 \xi(t) + \hat{C}_2 x_2(t) + \hat{D}_{11} v(t), \end{aligned} \tag{5.5}$$

where $\xi(t) = \binom{x_1(t)}{x_c(t)}$ is the slow component of the state and

$$\hat{A}_{11} = \begin{pmatrix} A_{11} & B_2^1 C_c \\ B_c C_2^1 & A_c \end{pmatrix}, \quad \hat{A}_{12} = \begin{pmatrix} A_{12} \\ B_c C_{22} \end{pmatrix}, \quad \hat{A}_{21} = \begin{pmatrix} A_{21} & B_2^2 C_c \end{pmatrix}, \quad \hat{A}_{22} = A_{22},$$

$$\hat{A}_{11}^k = \begin{pmatrix} A_{11}^k & 0 \\ B_c C_{21}^k & 0 \end{pmatrix}, \quad \hat{A}_{12}^k = \begin{pmatrix} \varepsilon^\nu A_{12}^k \\ \varepsilon^\lambda B_c C_{22}^k \end{pmatrix}, \quad \hat{B}_1 = \begin{pmatrix} B_1^1 \\ B_c D_{21} \end{pmatrix}, \quad \hat{B}_2 = B_1^2,$$

$$\hat{C}_1 = (C_{11} \quad D_{12}C_c), \;\; \hat{C}_2 = C_{12}, \;\; \hat{D}_{11} = D_{11}.$$

In the same way, the closed loop system obtained by coupling the controller (5.3) to system (5.1) may be written as

$$\begin{aligned}
d\xi(t) &= [\check{A}_{11}\xi(t) + \check{A}_{12}x_2(t) + \check{B}_1 v(t)]dt + \sum_{k=1}^{N}[\check{A}_{11}^k\xi(t) + \check{A}_{12}^k x_2(t)]dw_k(t),\\
\varepsilon dx_2(t) &= [\check{A}_{21}\xi(t) + \check{A}_{22}x_2(t) + \check{B}_2 v(t)]dt + \sum_{k=1}^{N}[\check{A}_{21}^k\xi(t) + \check{A}_{22}x_2(t)]dw_k(t),\\
y_1(t) &= \check{C}_1\xi(t) + \check{C}_2 x_2(t) + \check{D}_{11}v(t),
\end{aligned} \tag{5.6}$$

where $\xi = \binom{x_1}{x_c}$ is the slow state, x_2 is the fast state, and the coefficient matrices are

$$\begin{aligned}
&\check{A}_{11} = \begin{pmatrix} A_{11} + B_2^1 D_c C_{21} & B_2^1 \\ B_c C_{12} & A_c \end{pmatrix}, \;\; \check{A}_{12} = \begin{pmatrix} A_{12} + B_2^1 D_c C_{22}) \\ B_c C_{22} \end{pmatrix},\\
&\check{A}_{21} = \begin{pmatrix} A_{21} + B_2^2 D_c C_{21} & B_2^2 C_c \end{pmatrix}, \;\; \check{A}_{22} = A_{22} + B_2^2 D_c C_{22},\\
&\check{A}_{11}^k = \begin{pmatrix} A_{11}^k + B_2^1 D_c C_{21}^k & 0 \\ B_c C_{21}^k & 0 \end{pmatrix}, \;\; \check{A}_{12}^k = \begin{pmatrix} \varepsilon^\nu A_{12}^k + \varepsilon^\lambda B_2^1 D_c C_{22}^k \\ \varepsilon^\lambda B_c C_{22}^k \end{pmatrix},\\
&\check{A}_{21}^k = \begin{pmatrix} \varepsilon^\nu A_{21}^k + B_2^2 D_c C_{21}^k & 0 \end{pmatrix}, \;\; \check{A}_{22}^k = \varepsilon^\nu A_{22}^k + \varepsilon^\lambda B_2^2 D_c C_{22}^k,\\
&\check{B}_1 = \begin{pmatrix} B_1^1 + B_2^1 D_c D_{21} \\ B_c D_{21} \end{pmatrix}, \;\; \check{B}_2 = B_1^2 + B_2^2 D_c D_{21},\\
&\check{C}_1 = (C_{11} + D_{12} D_c C_{21} \quad D_{12}C_c), \;\; \check{C}_2 = C_{12} + D_{12} D_c C_{22},\\
&\check{C}_{1j}^k = D_{12} D_c \check{C}_2^k, \;\; j = 1, 2, \; k = 1, 2, \ldots N, \;\; \check{D}_{11} = D_{11} + D_{12} D_c D_{21}.
\end{aligned}$$

REMARK 5.1. *The output of the closed loop system* (5.6) *is directly affected by the white noises which are presented in the measured output of the plant* (5.1). *The stochastic systems of type* (5.6) *exceed the class of stochastic systems investigated in this paper. Hence, when a proper controller of type* (5.3) *is used to stabilize a stochastic system of type* (5.1), *we shall assume that* $C_{2j}^k = 0, j = 1, 2, k = 1, \ldots, N$. *However, since systems of type* (5.6) *appear in a natural way, an investigation of this type of system would be of interest.*

Assuming that A_{22} is invertible, we associate two subsystems of lower dimensions and independent of the small parameter ε:

$$\begin{aligned}
dx_r(t) &= [A_r x_r(t) + B_{r1} v_r(t) + B_{r2} u_r(t)]dt + \sum_{k=1}^{N} A_r^k x_r(t) dw_k(t),\\
y_{r1}(t) &= C_{r1} x_r(t) + D_r^{11} v_r(t) + D_r^{12} u_r(t),\\
dy_{r2}(t) &= [C_{r2} x_r(t) + D_r^{21} v_r(t) + D_r^{22} u_r(t)]dt + \sum_{k=1}^{N} C_{r2}^k x_r(t) dw_k(t),
\end{aligned} \tag{5.7}$$

which will be termed as "reduced subsystem" or "slow subsystem" associated with (5.1), where

$$\begin{aligned}
&A_r = A_{11} - A_{12}A_{22}^{-1}A_{21}, && A_r^k = A_{11}^k, \;\; k = 1, 2, \ldots, N,\\
&B_{rj} = B_j^1 - A_{12}A_{22}^{-1}B_j^2, \;\; j = 1, 2, && C_{ri} = C_{i1} - C_{i2}A_{22}^{-1}A_{21}, \;\; i = 1, 2,\\
&C_r^k = C_{21}^k, k = 1, 2, \ldots, N, && D_r^{ij} = D_{ij} - C_{i2}A_{22}^{-1}B_j^2, \; i, j = 1, 2.
\end{aligned}$$

We also associate the subsystem

$$\begin{aligned} x_f'(\tau) &= A_{22}x_f(\tau) + B_1^2 v_f(\tau) + B_2^2 u_f(\tau),\\ y_{f1}(\tau) &= C_{12}x_f(\tau) + D_{11}v_f(\tau) + D_{12}u_f(\tau),\\ y_{f2}(\tau) &= C_{22}x_f(\tau) + D_{21}v_f(\tau), \end{aligned} \tag{5.8}$$

where $\tau = \frac{t}{\varepsilon}$, which is the "boundary layer subsystem" or "fast subsystem" corresponding to system (5.1).

The closed loop system obtained by coupling a strict proper controller of type (5.2) to the reduced subsystem (5.7) is described by

$$\begin{aligned} dx_r(t) &= [A_r x_r(t) + B_{r2}C_c x_c(t) + B_{r1}v_r(t)]dt + \sum_{k=1}^{N} A_r^k x_r(t)dw_k(t),\\ dx_c(t) &= [B_c C_{r2}x_r + (A_c + B_c D_r^{22}C_c)x_c + B_c D_r^{21}v_r]dt + \sum_{k=1}^{N} B_c C_{22}^k x_r(t)dw_k(t),\\ y_{r1}(t) &= C_{r1}x_r(t) + D_r^{12}C_c x_c(t) + D_r^{11}v_r(t). \end{aligned} \tag{5.9}$$

We denote by $\mathbf{T}_r^{cl} : \tilde{L}^2([0,\infty), \mathbf{R}^{m_1}) \to \tilde{L}^2([0,\infty); \mathbf{R}^{p_1})$ the input-output operator defined by system (5.8).

The first result in this section is presented by the following theorem.

THEOREM 5.1. *Assume:* (a) *All eigenvalues of matrix A_{22} are located in the half plane of* $\mathrm{Re}(s) < 0$*; and* (b) *A strict proper controller of type* (5.2) *was designed in order to stabilize the reduced subsystem* (5.7). *Under these assumptions, there exists an $\hat{\varepsilon} > 0$, such that the same strict proper controller stabilizes the full system* (5.1) *for any arbitrary $\varepsilon \in (0, \hat{\varepsilon})$. Moreover, the input-output operator $\mathbf{T}_\varepsilon^{cl}$ of the corresponding closed loop system* (5.5) *has the asymptotic behavior*

$$\lim_{\varepsilon \to 0} \|\mathbf{T}_\varepsilon^{cl}\| = \max\{\|\mathbf{T}_r^{cl}\|;\ \|C_{12}(sI_{n_2} - A_{22})^{-1}B_1^2 + D_{11}\|_\infty\}. \tag{5.10}$$

Proof. To show that the designed controller stabilizes system (5.1), we shall apply Theorem 3.1 to system (5.5).

To this end, we remark that the coefficient matrix of the boundary layer subsystem is just A_{22}, which is stable.

On the other hand, by direct calculation, we obtain that the reduced subsystem associated to the closed loop system (5.5) is just (5.9).

By assumption (b), the zero state equilibrium of system (5.9) (with $v_r = 0$) is ESMS.

Finally, the asymptotic behavior of the norm of the input-output operator associated to the closed loop system can be obtained by applying Theorem 4.5. □

REMARK 5.2. *Based on the stochastic version of the small gain theorem (see, e.g.,* [7]), *it follows that the level of robustness achieved by a stabilizing controller is measured by the norm of the input-output operator associated to the closed loop system.*

From Theorem 5.1, it follows that, if a strict proper controller of type (5.2) was designed to stabilize the reduced subsystem (5.7) and to achieve $\|\mathbf{T}_r^{cl}\| < \gamma$ for a prefixed tolerance $\gamma > 0$, then the same controller will achieve a level of attenuation less than $\gamma + \mathcal{O}(\varepsilon)$ for the full system (5.1) only if

$$\|C_{12}(sI_{n_2} - A_{22})^{-1}B_1^2 + D_{11}\|_\infty < \|\mathbf{T}_r^{cl}\|.$$

Hence equality (5.10) gives a measure of loss-level of robustness when a robustly stabilizing controller (5.2) designed to stabilize the subsystem (5.7) is used in the full system (5.1) with stable fast dynamics.

REMARK 5.3. *From the expressions of the coefficient matrices of the closed loop system* (5.6), *it follows that the coefficient matrix of the boundary layer subsystem may be unstable even if* A_{22} *is stable.*

Then, under a proper controller, the fast dynamics may become unstable. We recall that, if A_{22} *is stable, then* $A_{22} + B_2^2 D_c C_{22}$ *is also stable if and only if*

$$\|D_c\| < [\|C_{22}(sI_{n_2} - A_{22})^{-1}B_2^2\|_\infty]^{-1}.$$

For details, see [27].

THEOREM 5.2. *Assume that* (a) *The matrix* A_{22} *is stable and* $C_{2j}^k = 0$, $j = 1, 2$, $k = 1, \ldots, N$, *and* (b) *a controller of the form*

$$\begin{aligned} \dot{x}_c(t) &= \tilde{A}_c x_c(t) + \tilde{B}_c y_2(t), \\ \dot{u}_2(t) &= \tilde{C}_c x_c(t) + \tilde{D}_c y_2(t) \end{aligned} \tag{5.11}$$

was designed in order to stabilize the system obtained from (5.7) *taking* $D_r^{22} = 0$. *In addition, the following conditions are satisfied:*

$$\begin{aligned} &\text{(i)} \quad A_{22} - B_{22}\tilde{D}_c C_{22} \text{ is invertible, and} \\ &\text{(ii)} \quad \|\tilde{D}_c(I_{p_2} - C_{22}A_{22}^{-1}B_2^2\tilde{D}_c)^{-1}\| < (\|C_{22}(sI_{n_2} - A_{22})^{-1}B_2^2\|_\infty)^{-1}. \end{aligned} \tag{5.12}$$

Set

$$\begin{aligned} A_c &= \tilde{A}_c + \tilde{B}_c(I_{p_2} - C_{22}A_{22}^{-1}B_2^2\tilde{D}_c)^{-1}C_{22}A_{22}^{-1}B_2^2\tilde{C}_c, \\ B_c &= \tilde{B}_c(I_{p_2} - C_{22}A_{22}^{-1}B_2^2\tilde{D}_c)^{-1}, \\ C_c &= \tilde{C}_c + \tilde{D}_c(I_{p_2} - C_{22}A_{22}^{-1}B_2^2\tilde{D}_c)^{-1}C_{22}A_{22}^{-1}B_2^2\tilde{C}_c, \\ D_c &= \tilde{D}_c(I_{p_2} - C_{22}A_{22}^{-1}B_2^2\tilde{D}_c)^{-1}. \end{aligned} \tag{5.13}$$

Under these conditions there exists an $\tilde{\varepsilon} > 0$ *such that the controller of type* (5.3) *having the coefficient matrix defined in* (5.13) *stabilizes system* (5.1) *for any arbitrary* $\varepsilon \in (0, \tilde{\varepsilon})$.

Moreover, the norm of the input-output operator defined by the corresponding closed loop system verifies

$$\lim_{\varepsilon \to \infty} \|\mathbf{T}_\varepsilon^{cl}\| = \max\{\|\tilde{\mathbf{T}}_r^{cl}\|, \|\tilde{G}_f\|_\infty\}, \tag{5.14}$$

where $\tilde{\mathbf{T}}_r^{cl}$ *is the input-output operator obtained by coupling the controller* (5.11) *to the reduced system* (5.7) *with* $D_r^{22} = 0$ *and*

$$\begin{aligned} \tilde{G}_f(s) = {} & (C_{12} + D_{12}D_c C_{22})(sI_{n_2} - A_{22} - B_2^2 D_c C_{22})^{-1}(B_1^2 - B_2^2 D_c D_{21}) \\ & + D_{11} + D_{12}D_c D_{21}. \end{aligned} \tag{5.15}$$

Proof. If $A_{22} - B_2^2\tilde{D}_c C_{22}$ is invertible, then, by direct calculation, it can be checked that $I_{p_2} - C_{22}A_{22}^{-1}B_2^2\tilde{D}_c$ is an invertible matrix, and therefore the formulae (5.13) are well defined. Now, the conclusion of the theorem follows in the same way as in the case of Theorem 5.1 by applying Theorems 3.1 and 4.5. ∎

REMARK 5.4. *From Theorem* 5.2, *we may conclude that if we design a controller* (5.11) *in order to stabilize the slow subsystem associated to the given system* (5.1) *and to achieve a prescribed level of disturbance attenuation for that slow subsystem, then such a controller may deteriorate the properties concerning the stabilization and the level of disturbance attenuation of the fast subsystem, and, as a consequence, it may deteriorate properties concerning the disturbances attenuation of the closed loop system defined by* (5.1) *and* (5.13).

An alternative algorithm of designing of an ε-independent stabilizing controller for singularly perturbed systems of type (5.1) could be as follows.

Step 1. Choose a matrix D_c such that the control $u_f = D_c x_f$ stabilizes the fast subsystem (5.8) and $\|\tilde{G}_f(s)\|_\infty < \gamma$, $\tilde{G}_f(s)$ being defined in (5.15).

Step 2. Make the change $u(t) = D_c y_2 + \tilde{u}_2$ in the space of controls of system (5.1).

Thus we obtain the modified system

$$
\begin{aligned}
dx_1(t) &= [(A_{11} + B_2^1 D_c C_{21})x_1(t) + (A_{12} + B_2^2 D_c C_{22})x_2(t) \\
&\quad + (B_1^1 + B_2^1 D_c D_{21})v(t) + B_2^1 \tilde{u}_2(t)]dt + \sum_{k=1}^{N} [A_{11}^k x_1(t) + \varepsilon^\mu A_{12}^k x_2(t)]dw_k(t), \\
\varepsilon dx_2(t) &= [(A_{21} + B_2^2 D_c C_{21})x_1(t) + (A_{22} + B_2^2 D_c C_{22})x_2(t) + (B_1^2 + B_2^2 D_c D_{21})v(t) \\
&\quad + B_2^2 \tilde{u}_2(t)]dt + \varepsilon^{nu} \sum_{k=1}^{N} [A_{21}^k x_1(t) + A_{22}^k x_2(t)]dw_k(t), \qquad (5.16) \\
y_1(t) &= (C_{11} + D_{12} D_c C_{21})x_1(t) + (C_{12} + D_{12} D_c C_{22})x_2(t) \\
&\quad + (D_{11} + D_{12} D_c D_{21})v(t) + D_{12} \tilde{u}_2(t), \\
y_2(t) &= C_{21} x_1(t) + C_{22} x_2(t) + D_{21} v(t).
\end{aligned}
$$

Step 3. Design (if possible) a strict proper controller of type (5.2) which stabilizes the reduced subsystem associated to system (5.16) obtained by setting $\varepsilon = 0$. Again using Theorem 3.1, we obtain that the controller

$$
\begin{aligned}
u_2(t) &= D_c y_2(t) + \tilde{u}_2(t), \\
\tilde{u}_2(t) &= C_c x_c(t), \qquad (5.17) \\
\dot{x}_c &= A_c x_c(t) + B_c y_2(t)
\end{aligned}
$$

stabilizes system (5.1) for arbitrary $\varepsilon > 0$ small enough. By applying Theorem 4.5, we may obtain information concerning the level of disturbance attenuation provided by the controller (5.17) in system (5.1).

REMARK 5.5. *In general, the reduced subsystem associated to system* (5.16) *is proper and not strict proper, and, in this case, a strict proper controller cannot provide a good level of disturbance attenuation for the slow closed loop system. Hence the controllers of type* (5.17) *designed to improve the level of disturbance attenuation for the fast subsystem of system* (5.1) *may deteriorate the properties of disturbance attenuation of the slow close loop system. The main conclusion of this section is that a stabilizing controller achieving a satisfactory level of disturbance attenuation for both the slow part and the fast part of the closed loop system must be a controller with two time scales.*

Appendix.

LEMMA A.1. *Consider the affine system*

$$dx(t) = [A_0x(t) + F_0(t)]dt + \sum_{i=1}^{N}[A_ix(t) + F_i(t)]dw_i(t), \tag{A.1}$$

where $F_i \in L_w^2((0,\infty)\times\Omega,\mathbf{R}^n), i = 0,1,2,\ldots,N$. *If the zero solution of the linear system*

$$dx(t) = A_0x(t)dt + \sum_{i=1}^{N} A_ix(t)dw_i(t)$$

is ESMS, then there exist positive constants c *and* α *such that all the solutions of system* (A.1) *verify*

$$E|x(t)^2 \le c\left[e^{-\alpha(t-t_0)}E|x(t_0)|^2 + \sum_{i=0}^{N}\int_{t_0}^{t} e^{-\alpha(t-s)}E|F_i(s)|^2ds\right] \quad \forall t \ge t_0 \ge 0.$$

Proof. The proof can be carried out by the same approach as that used in Proposition 1 in [10] or as in Theorem 5.1 in [11], which is based on a Lyapunov functional technique. □

It should be mentioned that the proof of this inequality cannot be conducted by direct estimation from the constant variation formula, as in the deterministic framework.

LEMMA A.2. *Consider the affine system with singular perturbation*

$$\varepsilon dx(t) = [A_0x(t) + g_0(t)]dt + \varepsilon^{\nu}\sum_{i=1}^{N}[A_ix(t) + g_i(t)]dw_i(t), \tag{A.2}$$

where $\nu > \frac{1}{2}, g_i \in L_w^2((0,\infty)\times\Omega,\mathbf{R}^n), i = 0,1,2,\ldots,N$. *If* A_0 *is Hurwitz, then there exist positive constants* c *and* α *such that all solutions of system* (A.2) *verify*

$$E|x(t,\varepsilon)|^2 \le c\left[e^{\frac{-\alpha(t-t_0)}{\varepsilon}}E|x(t_0,\varepsilon)|^2 + \sum_{i=0}^{N}\int_{t_0}^{t} e^{\frac{-\alpha(t-s)}{\varepsilon}}E|g_i(s)|^2ds\right] \quad \forall t \ge t_0 \ge 0,$$

where $\varepsilon > 0$ *is small enough.*

Proof. The desired result can be carried out by using Lemma A.1. □

Acknowledgment. The authors wish to thank the referees for their constructive comments and suggestions, which have improved the presentation.

REFERENCES

[1] Z. ARTSTEIN AND V. GAITSGORY, *Tracking fast trajectories along a slow dynamics: A singular perturbations approach*, SIAM J. Control Optim., 35 (1997), pp. 1487–1507.

[2] Z. ARTSTEIN AND A. VIGODNER, *Singularly perturbed ordinary differential equations with dynamic limits*, Proc. Roy. Soc. Edinburgh Sect. A, 126 (1996), pp. 541–569.

[3] J. C. DOYLE, K. GLOVER, P. P. KHARGONEKAR, AND B. A. FRANCIS, *State space solutions to the standard* H^2 *and* H^∞ *control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.

[4] V. Dragan, *Asymptotic expansions for game-theoretic Riccati equations and stabilization with disturbance attenuation for singularly perturbed systems*, Systems Control Lett., 20 (1993), pp. 455–463.
[5] V. Dragan, *H_∞-norms and disturbance attenuation for systems with fast transients*, IEEE Trans. Automat. Control, 41 (1996), pp. 747–750.
[6] V. Dragan and A. Halanay, *Stabilization of Linear Systems*, Birkhäuser Boston, Boston, 1999.
[7] V. Dragan, A. Halanay, and A. Stoica, *A small gain theorem for linear stochastic systems*, Systems Control Lett., 30 (1997), pp. 243–251.
[8] V. Dragan, A. Halanay, and A. Stoica, *The γ-attenuation problem for systems with state dependent noise*, Stochastic Anal. Appl., 17 (1999), pp. 395–404.
[9] V. Dragan and T. Morozan, *Global solutions to a game-theoretic Riccati equation of stochastic control*, J. Differential Equations, 138 (1997), pp. 328–350.
[10] V. Dragan and T. Morozan, *Mixed input-output optimization for time varying Itô systems with state dependent noise*, Dynam. Contin., Discrete Impuls. Systems, 3 (1997), pp. 317–333.
[11] V. Dragan and T. Morozan, *Stability and robust stabilization to linear stochastic systems described by differential equations with Markovian jumping and multiplicative white noise*, Stochastic Anal. Appl., 35 (2002), to appear.
[12] V. Dragan, P. Shi, and E. K. Boukas, *Control of singularly perturbed systems with Markovian jump parameters: An H_∞ approach*, Automatica J. IFAC, 35 (1999), pp. 1369–1378.
[13] E. Fridman, *Exact decomposition of linear singularly perturbed H^∞-optimal control problem*, Kybernetika (Prague), 31 (1995), pp. 589–597.
[14] E. Fridman, *Near-optimal H^∞ control of linear singularly perturbed systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 236–240.
[15] A. Friedman, *Stochastic Differential Equations and Applications*, Academic Press, New York, 1975.
[16] V. Gaitsgory, *Use of the averaging method in control problems*, Differential Equations, 22 (1986), pp. 1290–1299.
[17] V. Gaitsgory, *Suboptimization of singularly perturbed control systems*, SIAM J. Control Optim., 30 (1992), pp. 1228–1249.
[18] V. Gaitsgory, *Control of Systems with Slow and Fast Motions*, Nauka, Moscow, 1993 (in Russian).
[19] V. Gaitsgory, *Suboptimal control of singularly perturbed systems and periodic optimization*, IEEE Trans. Automat. Control, 38 (1993), pp. 888–903.
[20] V. Gaitsgory, *Limit Hamilton-Jacobi-Isaacs equations for singularly perturbed zero-sum differential games*, J. Math. Anal. Appl., 202 (1996), pp. 862–899.
[21] V. Gaitsgory and P. Shi, *Limit Hamilton-Jacobi-Isaacs equations for singularly perturbed zero-sum dynamic (discrete time) games*, in Proceedings of the 7th International Symposium on Dynamic Games and Applications, Kanagawa, Japan, 1996, pp. 168–174.
[22] B. F. Gardner and J. B. Cruz, *Well-posedness of singular perturbed Nash games*, J. Franklin Inst., 30 (1978), pp. 355–374.
[23] F. Garofalo and G. Leitmann, *Nonlinear composite control of a class of nominally linear singularly perturbed uncertain systems*, in Deterministic Control of Uncertain Systems, A. S. I. Zinober, ed., IEE Press, London, 1990, pp. 269–288.
[24] G. Grammel, *Controllability of differential inclusions*, J. Dynam. Control Systems, 1 (1995), pp. 581–595.
[25] G. Grammel, *Singularly perturbed control systems: Recent progress*, in Proceedings of the 35th IEEE Conference on Decision and Control, Kobe, Japan, 1996, pp. 505– 510.
[26] G. Grammel, *Singularly perturbed differential inclusions: An averaging approach*, Set-Valued Anal., 4 (1996), pp. 361–374.
[27] D. Hinrichsen and A. J. Pritchard, *Stability radii of linear systems*, Systems Control Lett., 7 (1986), pp. 1–10.
[28] H. K. Khalil and F. C. Chen, *H_∞ control of two-time-scale systems*, Systems Control Lett., 19 (1992), pp. 35–42.
[29] H. K. Khalil and P. V. Kokotović, *Feedback and well-posedness of singularly perturbed Nash games*, IEEE Trans. Automat. Control, 24 (1979), pp. 699–708.
[30] A. Klimusev and N. Krasovski, *Uniform asymptotic stability of system of differential equations having small parameters at the derivatives*, Applied. Math. Mech., 26 (1962), pp. 680–690.
[31] P. V. Kokotović, *Applications of singular perturbations techniques to control problems*, SIAM Rev., 26 (1984), pp. 501–550.

[32] P. V. Kokotović, H. Khalil, and J. O'Reilly, *Singular Perturbations in Control: Analysis and Design*, Academic Press, New York, 1984.
[33] P. V. Kokotović, H. K. Khalil, and J. O'Reilly, *Singular Perturbations in Control: Analysis and Design*, Academic Press, New York, 1986.
[34] P. V. Kokotović, R. E. Khalil, and J. O'Reilly, *Singular Perturbation Methods in Control: Analysis and Design*, Academic Press, New York, 1986.
[35] P. V. Kokotović, R. E. O'Malley, and P. Sannuti, *Singular perturbations and order reduction in control theory*, Automatica J. IFAC, 12 (1976), pp. 123–132.
[36] H. J. Kushner, *Stochastic Stability and Control*, Academic Press, New York, 1967.
[37] G. Ladde and V. Lakshmikantham, *Random Differential Inequalities*, Academic Press, New York, 1980.
[38] T. Morozan, *Parametrized Riccati Equation Associated to Input Output Operators for Time Varying Stochastic Differential Equations with State Dependent Noise*, Tech. report 37/1995, Institute of Mathematics of Romanian Academy of Science, Bucharest, Romania, 1995.
[39] T. Morozan, *Parametrized Riccati equations and input-output operators for time-varying stochastic differential equations with state dependent noise*, Stud. Cerc. Mat., 51 (1999), pp. 99–115.
[40] B. Oksendal, *Stochastic Differential Equations: An Introduction with Applications*, Springer-Verlag, New York, 2000.
[41] R. E. O'Malley, *Introduction to Singular Perturbations*, Academic Press, New York, 1974.
[42] Z. Pan and T. Basar, *H^∞-optimal control for singularly perturbed systems. Part* I: *Perfect state measurements*, Automatica J. IFAC, 29 (1993), pp. 401–423.
[43] Z. Pan and T. Basar, *H^∞-optimal control for singularly perturbed systems. Part* II: *Imperfect state measurements*, IEEE Trans. Automat. Control, 39 (1994), pp. 280–299.
[44] A. A. Pervozvansky and V. Gaitsgory, *Theory of Suboptimal Decisions*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1988.
[45] M. Quincampox, *Contribution a L'etude des perturbations singulieres pour les systemes controles et les inclusions differentialles*, C. R. Acad. Sci. Paris Ser. I Math., 316 (1993), pp. 133–138.
[46] M. A. Rami and X. Y. Zhou, *Linear matrix inequalities, Riccati equations, and indefinite stochastic linear quadratic controls*, IEEE Trans. Automat. Control, 45 (2000), pp. 1131–1143.
[47] V. R. Saksena, J. O'Reilly, and P. V. Kokotović, *Singular perturbations and time-scale methods in control theory: Survey* 1976–1983, Automatica J. IFAC, 20 (1984), pp. 273–293.
[48] S. M. Shahruz, *H_∞-optimal compensators for singularly perturbed systems*, in Proceedings of the 28th IEEE Conference on Decision and Control, Tampa, FL, 1989, pp. 2397–2398.
[49] P. Shi and V. Dragan, *Asymptotic H_∞ control of singularly perturbed systems with parametric uncertainties*, IEEE Trans. Automat. Control, 44 (1999), pp. 1738–1742.
[50] H. D. Tuan and S. Hosoe, *On a state-space approach in robust control for singularly perturbed systems*, Internat. J. Control, 66 (1997), pp. 435–462.
[51] J. L. Vian and M. E. Sawan, *H_∞ control for singularly perturbed systems*, in Proceedings of the 30th IEEE Conference on Decision and Control, Brighton, England, 1991, pp. 1072–1074.
[52] A. Vigodner, *Limits of singularly perturbed control problems with statistical dynamics of fast motions*, SIAM J. Control Optim., 35 (1997), pp. 1–28.
[53] J. Willems and J. Willems, *Feedback stabilizability for stochastic systems with state and control dependent noise*, Automatica J. IFAC, 12 (1976), pp. 277–283.
[54] W. M. Wonham, *Random differential equations in control theory*, in Probabilistic Methods in Applied Mathematics, A. T. Bharucha-reid, ed., Academic Press, New York, 1971, pp. 131–213.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 1, pp. 164–180

ON THE MODELLING AND STABILIZATION OF FLOWS IN NETWORKS OF OPEN CANALS*

GUENTER LEUGERING† AND E. J. P. GEORG SCHMIDT‡

Abstract. In this paper, we present a model for the controlled flow of a fluid through a network of canals using a coupled system of St. Venant equations. We then generalize in a variety of ways recent results of Coron, d'Ándréa-Novel, and Bastin concerning the stabilizability around equilibrium of the flow through a single channel. This work is based on the theory of quasilinear hyperbolic systems and, in particular, on a delicate result of Li Ta-tsien concerning the existence and decay of global classical solutions.

Key words. St. Venant equations, hyperbolic systems, stabilizability, control of canals

AMS subject classifications. 35L45, 35L50, 35L65, 93C20

PII. S0363012900375664

1. Introduction.

The St. Venant equations are a nonlinear hyperbolic system first introduced in [10] to model the flow of fluid through a canal (or channel). They have, in particular, become a standard tool for hydraulic engineers used in the modelling of the dynamics of canals and rivers.

The book [4] provides a useful engineering reference to this topic. A recent paper [2] has considered the flow along a canal between two large bodies of water controlled by sluice gates at the ends of the canal. The settings of these gates determine the fluid velocity at the two ends, and the main result shows that suitable feedback boundary conditions can be used to exponentially stabilize a given operational state of the canal. This result depends on a subtle theorem of Greenberg and Li [5], which guarantees the exponential decay of solutions to certain hyperbolic systems in two variables subject to boundary conditions which impose damping. See also the monograph by Li [8] and [1], [3].

Our main contributions in this paper are

- to derive a model for the flow of fluid through a network of canals;
- to prove the feedback stabilizability of certain equilibrium operating states in a system of canals meeting at one "multiple node" for a broad class of feedback controls acting at the "simple nodes"; and
- to prove feedback null controllability using an absorbing feedback law.

Furthermore, we shall generalize the result of [2], allowing canals of nonrectangular cross section and using a broad class of boundary conditions. In order to prove stabilizability for a system of canals, we shall use a generalization of the "boundary damping" result in [5] to be found in [8] which applied to hyperbolic systems with arbitrarily many variables. As was the case for the previously cited papers, our

*Received by the editors July 24, 2000; accepted for publication (in revised form) November 12, 2001; published electronically April 26, 2002.
http://www.siam.org/journals/sicon/41-1/37566.html

†Fachbereich Mathematik AG 10, Technische Universität Darmstadt, 64289 Darmstadt, Germany (gleugering@mathematik.tu-darmstadt.de). The research of this author was supported by the grant DFG Le59513-1.

‡Department of Mathematics and Statistics, McGill University, 805 Sherbrooke Street West, Montreal, Canada H3A 2K6 (gschmidt@math.mcgill.ca). The research of this author was supported by the Natural Sciences and Engineering Research Council of Canada grant A727.

methods are entirely nonlinear but do remain within the realm of classical, shock-free solutions.

We end this introduction with some comments on notation. We shall denote derivatives of functions of a single scalar variable by D or sometimes by D_s, where s is the variable. In the presence of other variables, we let ∂_s denote partial differentiation with respect to s. We let $\bigtriangledown$ or $\bigtriangledown_{\mathbf{s}} = (\partial_{s_1}, \ldots, \partial_{s_n})$ denote the gradient of a function with respect to a vectorial argument $\mathbf{s} = (s_1, \ldots, s_n)$. We also introduce some norms. For an m-vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)$ and an $l \times m$ matrix A, we define

$$|\boldsymbol{\xi}|_\infty = \max\{|\xi_i| : i = 1, \ldots, m\}, \quad \| A \|_\infty = \max \left\{ \sum_{j=1}^{m} |A_{ij}| : i = 1, \ldots, l \right\}.$$

One then has

$$|A\boldsymbol{\xi}|_\infty \leq \| A \|_\infty \, |\boldsymbol{\xi}|_\infty \quad \text{and} \quad \| AB \|_\infty \leq \| A \|_\infty \| B \|_\infty,$$

where B is a second matrix having m rows. Let $C([0, L]; \mathbb{R}^n)$ and $C^1([0, L]; \mathbb{R}^n)$ denote, respectively, the spaces of continuous and continuously differentiable functions from $[0, L]$ to $\mathbb{R}^n$ with corresponding norms

$$\| \boldsymbol{\xi}(\cdot) \| = \sup\{|\boldsymbol{\xi}(x)|_\infty : x \in [0, L]\}, \quad \| \boldsymbol{\xi}(\cdot) \|_1 = \max\{\| \boldsymbol{\xi}(\cdot) \|, \| D\boldsymbol{\xi}(\cdot) \|\}.$$

2. The model. We consider first a single canal parametrized lengthwise by $x \in [0, L]$. Let $Y_b(x)$ denote the altitude above sea level of the bed of the canal at x. The variable $y = y(x) \in [0, d(x)]$ denotes the elevation above the canal bed, where $d(x)$ denotes the depth of the canal. See Figure 2.1. Let $\sigma(x, y)$ denote the width of the canal cross section at x and elevation y. Let $A(x, t)$ denote the area of the cross section at x occupied by water at time t. We assume that the water level is constant across the canal at height $h(x, t) \in [0, d(x)]$. Clearly, $A = A(x, h)$ and $h = h(x, A)$. See Figure 2.2. In particular, leaving aside t dependence for the moment,

$$A(x, h) = \int_0^h \sigma(x, y)\, dy \quad \text{and} \quad A = \int_0^{h(x,A)} \sigma(x, y)\, dy. \tag{2.1}$$

It follows that

$$\partial_h A(x, h) = b(x, h) \quad \text{and} \quad \partial_A h(x, A) = \frac{1}{b(x, A)}, \tag{2.2}$$

where $b(x, h) \triangleq \sigma(x, h)$ (or $b(x, A) \triangleq \sigma(x, h(x, A))$) is the width of the water surface at x.

It turns out to be convenient to choose $A(x, t)$, rather than $h(x, t)$, as our geometric state variable describing the distribution of water along the canal at a given time since it conveniently leads to a system of conservation laws. The derivation of the St. Venant equations depends on the assumption that the flow of water along the canal can be represented by a scalar velocity function $V(x, t)$ in the direction of the channel from 0 to L. This can be thought of either as a constant or as an average velocity over the cross section of the channel. For a more detailed physical discussion of the underlying assumptions of the St. Venant model, see [4, p. 8].

Assuming a constant density ($\rho \equiv 1$, say), the mass flow rate of the liquid along the channel is given by $Q(x, t) \triangleq A(x, t)V(x, t)$. Conservation of mass is then expressed by the conservation equation

$$\partial_t A + \partial_x [AV] = 0, \tag{2.3}$$

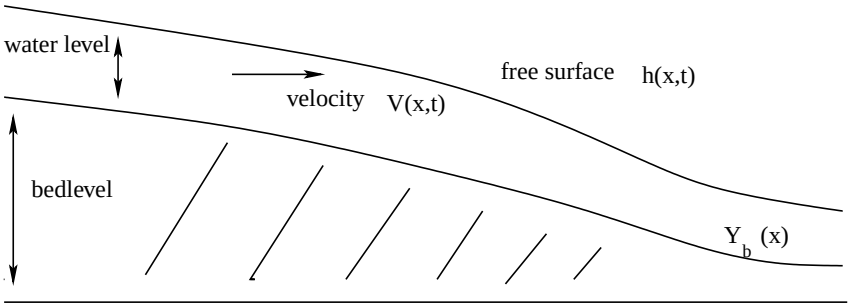


FIG. 2.1. *The configuration.*

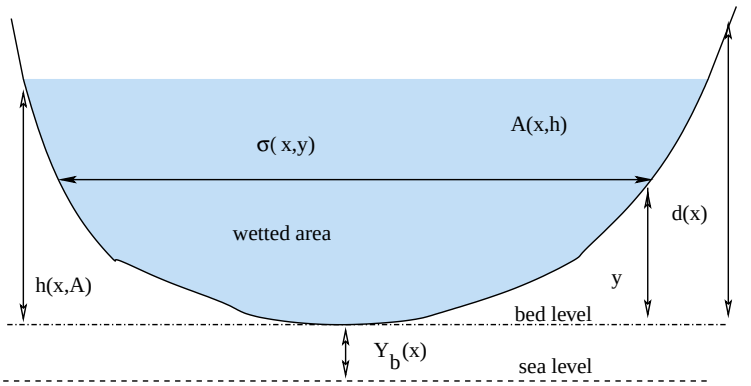


FIG. 2.2. *Cross section at x.*

or, in the weak form,

$$\int\int [A\partial_t\phi + AV\partial_x\phi]\, dx\, dt = 0 \tag{2.4}$$

for smooth test functions.

To obtain the St. Venant system, the conservation equation has to be complemented with a dynamical equation. We derive this from Hamilton's principle applied to the Lagrangian functional

$$\mathcal{L}(A,V) \triangleq \int_0^T \int_0^L \left[\frac{1}{2}AV^2 - g\int_0^{h(x,A)} [Y_b(x)+y]\sigma(x,y)\, dy\right] dxdt \tag{2.5}$$

obtained as the difference between the kinetic and potential energies, where g is the gravitational constant. Here A and $Q = AV$ are assumed to satisfy (2.3), and the same will have to be true of the variations δA and $\delta Q = (A+\delta A)(V+\delta V) - AV = A\delta V + V\delta A$. Now if we let $\delta A = \partial_x\phi$ and $\delta Q = -\partial_t\phi$, with ϕ a smooth function of x and t with support in $(0,L)\times(0,T)$, then (2.3) is indeed satisfied. So we get

$$\delta A = \partial_x\phi, \quad A\delta V = \delta Q - V\delta A = -\partial_t\phi - V\partial_x\phi.$$

Noting that

$$\int_0^{h(x,A)} Y_b(x)\sigma(x,y)\, dy = AY_b(x),$$

using (2.2), and recalling that $b(x, A) = \sigma(x, h(x, A))$, we now get

$$\begin{aligned}\frac{d}{d\lambda}\mathcal{L}(A+\lambda\delta A, V+\lambda\delta V)\Big|_{\lambda=0} &= \int_0^T \int_0^L \left[\frac{1}{2}V^2\delta A + AV\delta V - gY_b(x)\delta A - gh(x,A)\delta A\right] dxdt \\ &= \int_0^T \int_0^L \left[\left[\frac{1}{2}V^2 - gh(x,A) - gY_b(x)\right]\partial_x\phi - V[\partial_t\phi + V\partial_x\phi]\right] dxdt.\end{aligned}$$

Setting this equal to 0, we get

$$\int_0^T \int_0^L \left[\left[\frac{1}{2}V^2 + gh(x,A) + gY_b(x)\right]\partial_x\phi + V\partial_t\phi\right] dxdt = 0, \tag{2.6}$$

the weak form of

$$\partial_t V + \partial_x\left[\frac{1}{2}V^2 + gh(x,A) + gY_b(x)\right] = 0. \tag{2.7}$$

REMARK 1. *One can add an empirically motivated resistance, or friction, term to the left-hand side of the last equation. Various alternatives occur in the engineering literature (see, for example,* [4, pp. 19–22]). *Generically these are of the form* $F(x, A, V)$ *satisfying*

$$F(x, A, 0) = 0 \quad \textit{and} \quad VF(x, A, V) \geq 0. \tag{2.8}$$

We may also take into account sign conditions on the partial derivatives of F. Now we consider networks of canals. We use notation similar to that introduced in [6] for networks of strings and beams. We index the canals, and the quantities associated with the canals, by $i \in \mathcal{I} = \{1, \ldots, n\}$. We label the locations of the end points of the canals, which we shall refer to as nodes, by $j \in \{1, \ldots, m\}$. We distinguish between multiple nodes, indexed by $j \in \mathcal{J}_M$, at which various canals come together, and the simple nodes, indexed by $j \in \mathcal{J}_S$, which are endpoints of a single canal. For $j \in \mathcal{J}$, we introduce

$$\mathcal{I}_j = \{i \in \mathcal{I} : \text{ the } i\text{th canal meets the } j\text{th node}\}.$$

For $i \in \mathcal{I}_j$, we set $x_{ij} = 0$ or L_i corresponding to the end which meets the other canals at the jth node. We also set $\epsilon_{ij} = 1$ if $x_{ij} = L_i$ or $\epsilon_{ij} = -1$, if $x_{ij} = 0$.

At simple nodes, we shall later impose boundary conditions through which controls can be imposed on the network. At multiple nodes, we shall first impose the following condition expressing conservation of fluid in the flow through the node indexed by $j \in \mathcal{J}_M$:

$$\sum_{i\in\mathcal{I}_j} \epsilon_{ij} Q_i(x_{ij}, t) = \sum_{i\in\mathcal{I}_j} \epsilon_{ij} A_i(x_{ij}, t) V_i(x_{ij}, t) = 0 \quad \text{for } j \in \mathcal{J}_M. \tag{2.9}$$

We shall derive a second dynamic node condition from Hamilton's principle.

For each $i \in \mathcal{I}$, we require that $A_i(x, t)$ and $V_i(x, t)$ satisfy the continuity equation

$$\partial_t A_i + \partial_x[A_i V_i] = 0. \tag{2.10}$$

Let $\mathbf{A} = (A_1, \dots, A_n)$ and $\mathbf{V} = (V_1, \dots, V_n)$. We introduce the Lagrangian functional

$$\mathcal{L}(\mathbf{A}, \mathbf{V}) \triangleq \sum_{i \in \mathcal{I}} \int_0^T \int_0^{L_i} \left[\frac{1}{2} A_i V_i^2 - g \int_0^{h_i(x, A_i)} [Y_{bi}(x) + y] \sigma_i(x, y)\, dy \right] dx dt. \tag{2.11}$$

The components of $\mathbf{A}$ and of $\mathbf{V}$ and their variations have to satisfy (2.9) and (2.10). This can be achieved, much as before, for variations generated by

$$\delta A_i = \partial_x \phi_i, \quad A_i \delta V_i = \delta Q_i - V_i \delta A_i = -\partial_t \phi_i - V \partial_x \phi_i, \tag{2.12}$$

where the set of smooth functions $\phi_i(x, t)$ satisfy

$$\sum_{i \in \mathcal{I}_j} \epsilon_{ij} \partial_t \phi_i(x_{ij}, t) = 0 \quad \text{for each } j \in \mathcal{J}. \tag{2.13}$$

Exactly as before, we then get

$$\begin{aligned} &\frac{d}{d\lambda} \mathcal{L}(\mathbf{A} + \lambda \delta \mathbf{A}, \mathbf{V} + \lambda \delta \mathbf{V}) \Big|_{\lambda = 0} \\ &\quad = - \sum_{i \in \mathcal{I}} \int_0^T \int_0^{L_i} \left[\left(\frac{1}{2} V_i^2 + g h_i(x, A_i) + g Y_{bi}(x) \right) \partial_x \phi_i + V_i \partial_t \phi_i \right] dx dt. \end{aligned}$$

Now we set this equal to zero using the variations in two different ways. First, we restrict our attention to one index i; letting ϕ_i have support in $(0, L) \times (0, T)$ and setting variations in the other components equal to zero, we get the indexed version of (2.6) with indexed quantities, the weak form of

$$\partial_t V_i + \partial_x \left[\frac{1}{2} V_i^2 + g h_i(x, A_i) + g Y_{bi}(x) \right] = 0. \tag{2.14}$$

Second, we concentrate on the jth node, making variations only in A_i and V_i for $i \in \mathcal{I}_j$ with $j \in \mathcal{J}_M$. In fact, we choose any two indices $k, l \in \mathcal{I}_j$ and set $\phi_i \equiv 0$ for $i \neq k, l$. Supposing that $x_{kj} = 0$ and $x_{lj} = 0$, we can choose $\phi \in C_0^\infty([0, \check{L}) \times (0, T))$ with $\check{L} = \min\{L_k, L_l\}$ and set $\phi_k = \phi$ and $\phi_l = -\phi$ to get admissible variations. We introduce the notation

$$S_i \triangleq \frac{1}{2} V_i^2 + g h_i(x, A_i) + g Y_{bi}(x)$$

(a quantity which, divided by g, is called the *specific energy*) and get

$$\int_0^T \int_0^{\tilde{L}} \left[(S_k - S_l) \partial_x \phi + (V_k - V_l) \partial_t \phi \right] dx dt = 0 \tag{2.15}$$

for any $\phi \in C_0^\infty([0, \check{L}) \times (0, T))$. Assuming sufficient regularity, we can integrate by parts and use (2.14) and the support properties of $\phi(x, t)$ to get

$$\int_0^T [S_k(0, t) - S_l(0, t)] \phi(0, t)\, dt = 0.$$

Since ϕ can be chosen so that $\phi(0, t)$ is an arbitrary function in $C_0^\infty((0, T))$, we can conclude that

$$S_k(x_{kj}, t) \equiv S_l(x_{lj}, t).$$

The above argument can be modified to deal with the various cases in which x_{kj} and x_{lj} are not both zero. So we end up with the following dynamic node condition for each $j \in \mathcal{J}_M$:

$$\left[\frac{1}{2}V_i^2 + gh_i(\cdot, A_i) + gY_{bi}\right](x_{ij}, t) \text{ coincide for } i \in \mathcal{I}_j. \tag{2.16}$$

REMARK 2. *One can, in fact, derive this node condition from* (2.15) *without assuming differentiability of* A_i *and* V_i. *This is done by a mollification argument. So, in fact,*

$$\sum_{i\in\mathcal{I}} \int_0^T \int_0^{L_i} \left[\left(\frac{1}{2}V_i^2 + gh_i(x, A_i) + gY_{bi}(x)\right)\partial_x\phi_i + V_i\partial_t\phi_i\right] dxdt = 0 \tag{2.17}$$

for test functions $\phi_i \in C_0^\infty([0, L_i) \times (0, T))$ *satisfying that* (2.13) *is a suitable weak formulation of both the dynamic equations* (2.14) *and the dynamic node conditions* (2.16).

To summarize, the network is described by (2.14) and supplemented by the multiple node conditions (2.9) and (2.16) as well as boundary conditions to be imposed at the simple nodes and initial conditions prescribing $\mathbf{A}(x, 0)$ and $\mathbf{V}(x, 0)$. The discussion of the boundary conditions will be given later, taking into account issues concerning hyperbolic systems.

We end this section with a comment on energy conservation. Let

$$\begin{aligned} E(A, V) &\triangleq \frac{1}{2}AV^2 + g\int_0^{h(x,A)} [Y_b(x) + y]\sigma(x, y)\, dy \\ &= \frac{1}{2}AV^2 + gAY_b(x) + g\int_0^{h(x,A)} y\sigma(x, y)\, dy, \end{aligned} \tag{2.18}$$

where A and V satisfy (2.3) and (2.7). Then one easily verifies

$$\partial_t E + \partial_x[QS] = 0. \tag{2.19}$$

This of course also holds in indexed form for each canal, and if one introduces the total energy

$$\mathcal{E}(t) = \mathcal{E}(\mathbf{A}(\cdot, t), \mathbf{V}(\cdot, t)) \triangleq \sum_{i\in\mathcal{I}} \int_0^{L_i} E_i(A_i, V_i)(x, t)\, dx,$$

one obtains, using the multiple node conditions,

$$D_t\mathcal{E}(t) = \sum_{j\in\mathcal{J}_S} Q(x_{ij}, t)S(x_{ij}, t)$$

so that energy is conserved if, for example, there is no flow through the simple nodes of the canal network.

REMARK 3. *Friction can be introduced with a friction term* $F_i(x, A_i, V_i)$ *in the left-hand side of each equation* (2.14). *One then gets the following index-free variant of* (2.19)*:*

$$\partial_t E + \partial_x[QS] + F(x, A, V) = 0$$

as well as

$$D_t\mathcal{E}(t) = \sum_{j\in\mathcal{J}_S} Q(x_{ij},t)S(x_{ij},t) - \sum_{i\in\mathcal{I}} \int_0^{L_i} A_iV_iF_i(x,A_i,V_i)\,dx.$$

The friction terms lead to energy decay because of the assumption (2.8).

3. Equilibrium flows and their perturbations. We seek solutions $\overline{\mathbf{A}}$ and $\overline{\mathbf{V}}$ of the network system described in the previous section which depend only on x and not on t. Explicitly, they are required to satisfy

$$\begin{cases} \partial_x[\overline{A}_i\overline{V}_i] = 0 \ \text{ on } [0,L_i] & \text{for } i\in\mathcal{I}, \\ \partial_x[\frac{1}{2}\overline{V}_i^2 + gh_i(x,\overline{A}_i) + gY_{bi}(x)] = 0 \ \text{ on } [0,L_i] & \text{for } i\in\mathcal{I}, \\ \sum_{i\in\mathcal{I}_j}\epsilon_{ij}\overline{A}_i(x_{ij})\overline{V}_i(x_{ij}) = 0 & \text{for } j\in\mathcal{J}_M, \\ \frac{1}{2}\overline{V}_i(x_{ij})^2 + gh_i(x_{ij},\overline{A}_i(x_{ij})) + gY_{bi}(x) \text{ coincide} & \text{for } j\in\mathcal{J}_M, i\in\mathcal{I}_j. \end{cases} \tag{3.1}$$

It is easy to deduce that one must have

$$\overline{A}_i(x)\overline{V}_i(x) \equiv \overline{Q}_i \ \text{ with } \sum_{i\in\mathcal{I}_j}\epsilon_{ij}\overline{Q}_{ij} = 0, \tag{3.2}$$

$$\overline{S}_i(x) = \frac{1}{2}\overline{V}_i(x)^2 + gh_i(x,\overline{A}_i(x)) + gY_{bi}(x) \equiv \overline{S}, \tag{3.3}$$

where $\overline{Q}_i$ and $\overline{S}$ are constants.

We say that the fluid in the canals is *still* if $\mathbf{V} = \mathbf{0}$. In that case, the components of $\mathbf{A}(x)$ are determined from

$$h_i(x,\overline{A}_i(x)) = \frac{1}{g}[\overline{S} - gY_{bi}(x)],$$

which will have a solution respecting the depth restriction on each canal if and only if

$$0 \le \frac{1}{g}[\overline{S} - gY_{bi}(x)] \le d_i(x) \ \text{ for all } i\in\mathcal{I}.$$

The equilibria which are not still are more difficult to determine. For the purposes of this paper, we now make two restrictive assumptions, namely

- the canals are *prismatic*, which means that the cross sections of the canals do not depend on x (so d_i, h_i, σ_i, and b_i do not depend directly on x); and
- the system of canals is *level*, the beds of the canals all lying at the same constant elevation Y_b.

In this case, equilibria $\mathbf{A},\mathbf{V}$ will not depend on x. One can think of canals designed with certain operating conditions in mind. For example, one can specify a standard water height of $\overline{h}$ for the whole canal system with $\overline{h} < d_i$ for each $i\in\mathcal{I}$. For each canal, one can fix the flow direction, encoding this by $\epsilon_i = 1$ or -1 depending on whether $\overline{V}_i$ is to be positive or negative. This is to be done in such a way that at each multiple node flows occur both into and out of the node. We then try to design the cross sections of each canal in such a way that $\overline{A}_i = A_i(\overline{h})$ satisfy

$$\sum_{i\in\mathcal{I}_j}\epsilon_{ij}\epsilon_i\overline{A}_i = 0. \tag{3.4}$$

If this is possible, we can set $\overline{V}_i = \epsilon_i \overline{V}$, where $\overline{V}$ is any given positive velocity, and easily check that the conditions (3.1) are satisfied. To clarify the meaning of the above condition, we note that, when $\epsilon_{ij}\epsilon_i = 1$, the flow in the ith canal is into the jth node while, when $\epsilon_{ij}\epsilon_i = -1$, that flow is away from the node. For canals of rectangular cross section with b_i the breadth of the ith canal, one can require

$$\sum_{i \in \mathcal{I}_j} \epsilon_{ij}\epsilon_i b_i = 0, \tag{3.5}$$

independently of the value $\overline{h}$. It is almost obvious that the above process always works in the following particular network configurations:

- star configurations in which n canals all meet at one multiple node;
- tree configurations in which the direction of flow is always toward the trunk or always away from the trunk.

REMARK 4. *One can easily experiment with a variety of particular networks and direction assignments to find many other configurations in which the above process works. These may include closed paths.*

A general goal now is to stabilize the flow around such an equilibrium flow by means of suitable feedback boundary conditions at the simple nodes, which lie at the extremities of the canal network, i.e., at the points where water flows into or out of the system of canals. At present, we can do this only for star configurations. In fact, we first consider a single canal for which our results are also in large part new.

4. Stabilization and null controllability for a single canal. We consider one level, prismatic canal and study perturbations of constant equilibrium flows.

First we make use of the standard method of Riemann invariants for hyperbolic systems to be found, for example, in Taylor [9, Chapter 16]. We begin with a single canal. Evaluating the x derivatives in (2.3) and (2.7), we can rewrite these equations as a system

$$\partial_t \begin{pmatrix} A \\ V \end{pmatrix} + \begin{pmatrix} V & A \\ g/b(A) & V \end{pmatrix} \partial_x \begin{pmatrix} A \\ V \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{4.1}$$

The eigenvalues of the matrix are

$$\lambda_\pm(A,V) = V \pm \gamma(A), \quad \text{where } \gamma(A) = \sqrt{gA/b(A)} \tag{4.2}$$

with $\lambda_+ > 0$ and $\lambda_- < 0$ in the *subcritical* case that

$$V^2 < gA/b(A). \tag{4.3}$$

The corresponding left eigenvectors are

$$\mathbf{l}_\pm(A) = \frac{1}{2}(\pm\sqrt{g/[Ab(A)]}, 1).$$

Riemann invariants are then obtained by solving

$$\bigtriangledown \xi_\pm(A,V) = (\partial_A, \partial_V)\xi_\pm(A,V) = \mathbf{l}_\pm(A).$$

We can take

$$\xi_\pm(A,V) = \frac{1}{2}\left(V \pm \int_0^A q(\alpha)\, d\alpha\right) \quad \text{with} \quad q(\alpha) = \sqrt{\frac{g}{\alpha b(\alpha)}}.$$

The system (4.1) is now equivalent to

$$\partial_t \xi_\pm(A,V) + \lambda_\pm(A,V)\partial_x \xi_\pm(A,V) = 0.$$

For solutions A and V, the *Riemann invariants* $\xi_\pm$ are constant along characteristic curves $(x_\pm(t),t)$ with

$$D_t x_\pm(t) = \lambda_\pm(A(x_\pm(t),t),V(x_\pm(t),t)).$$

Now we consider perturbations $A = \overline{A} + a$ and $V = \overline{V} + v$ of an equilibrium state. We assume that the equilibrium flow is subcritical so that this will continue to be the case for small perturbations. The system becomes

$$\partial_t \begin{pmatrix} a \\ v \end{pmatrix} + \begin{pmatrix} \overline{V}+v & \overline{A}+a \\ g/b(\overline{A}+a) & \overline{V}+v \end{pmatrix} \partial_x \begin{pmatrix} a \\ v \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{4.4}$$

In terms of the perturbation variables a and v, the eigenvalues of the matrix are

$$\lambda_\pm(a,v) = \overline{V} + v \pm \beta(a), \quad \text{where } \beta(a) = \sqrt{g\frac{\overline{A}+a}{b(\overline{A}+a)}}, \tag{4.5}$$

corresponding to Riemann invariants

$$\xi_\pm(a,v) = \frac{1}{2}\left(v \pm \int_0^a p(\alpha)\,d\alpha\right) \quad \text{with} \quad p(\alpha) = q(\overline{A}+\alpha). \tag{4.6}$$

The characteristic curves $(x_\pm(t),t)$ are now determined by solving

$$D_t x_\pm(t) = \lambda_\pm(a(x_\pm(t),t),v(x_\pm(t),t)). \tag{4.7}$$

We shall often indicate the values of functions evaluated at equilibrium (i.e., with $a = 0$ and $v = 0$) by a bar. For a and v small, the characteristics will lie close to those of the subcritical equilibrium, namely $(\overline{x}_\pm(t),t)$, with

$$\overline{x}_\pm(t) = x_0 + (\overline{V} \pm \overline{\beta})(t - t_0), \tag{4.8}$$

where (x_0,t_0) are arbitrary initial data. We note that the map

$$(a,v) \mapsto (\xi_+(a,v),\xi_-(a,v))$$

is invertible with

$$v = \xi_+ + \xi_- \quad \text{and} \quad a = a(\xi_+ - \xi_-),$$

where $a(\xi)$ is defined implicitly by $\xi = \int_0^{a(\xi)} p(\alpha)\,d\alpha$. We note that $Da(\xi) = 1/p(a(\xi))$.

We next turn to the feedback stabilization of a single canal, which should drive perturbations a and v to zero exponentially in time. The system has to be complemented by initial conditions

$$a(x,0) = a^0(x), \quad v(x,0) = v^0(x). \tag{4.9}$$

In terms of the Riemann invariants, it is well known that one can impose boundary conditions of the form

$$\xi_+(0,t) = g^0(\xi_-(0,t)) \quad \text{and} \quad \xi_-(L,t) = g^L(\xi_+(L,t)).$$

In particular, one could impose the absorbing boundary conditions

$$\xi_+(0,t) = 0 \quad \text{and} \quad \xi_-(L,t) = 0, \tag{4.10}$$

equivalent to the following feedback boundary conditions in terms of a and v:

$$v(0,t) = -P(a(0,t)), \quad v(L,t) = P(a(L,t)) \quad \text{with} \quad P(a) = \int_0^a p(\alpha)\, d\alpha. \tag{4.11}$$

One could also replace $P(a)$ by Taylor approximations $P_1(a) = p(0)a$ or $P_2(a) = p(0)a + p'(0)a^2/2$, say, or, alternatively, by other functions of a.

We shall need the following lemma.

LEMMA 1. *Consider boundary conditions of the form*

$$v(0,t) = f^0(a(0,t)), \quad v(L,t) = f^L(a(L,t)), \tag{4.12}$$

where f^0 and f^L are continuously differentiable functions of a near 0 and satisfy

$$f^0(0) = 0,\ Df^0(0) \neq p(0),\ f^L(0) = 0,\ Df^L(0) \neq -p(0).$$

Then, for small enough values of the variables, the boundary conditions can be rewritten as

$$\xi_+(0,t) = g^0(\xi_-(0,t)), \quad \xi_-(L,t) = g^L(\xi_+(L,t))$$

with

$$g^0(0) = 0,\ Dg^0(0) = \frac{Df^0(0) + p(0)}{Df^0(0) - p(0)},\ g^L(0) = 0,\ Dg^L(0) = \frac{Df^L(0) - p(0)}{Df^L(0) + p(0)}.$$

Proof. We prove the assertions concerning the boundary condition at 0. From $v = \xi_+ + \xi_- = f^0(a(\xi_+ - \xi_-))$, where we suppress the argument $(0,t)$ in all the functions, one gets

$$\xi_- = \phi^0(\xi_+ - \xi_-), \quad \text{with } \phi^0(\xi) = \frac{1}{2}[-\xi + f^0(a(\xi))].$$

Now $D\phi^0(0) = \frac{1}{2}[-1 + Df^0(0)/p(0)]$, which is not zero as long as $Df^0(0) \neq p(0)$. Hence ϕ^0 has a local inverse near 0. So locally $\xi_+ - \xi_- = {\phi^0}^{-1}(\xi_-)$ or $\xi_+ = \xi_- + {\phi^0}^{-1}(\xi_-) = g^0(\xi_-)$ and

$$Dg^0(0) = 1 + \frac{1}{D\phi^0(0)} = \frac{Df^0(0) + p(0)}{Df^0(0) - p(0)}.$$

The other boundary condition is treated similarly starting from

$$\xi_+ = \phi^L(\xi_+ - \xi_-), \quad \text{with } \phi^L(\xi) = \tfrac{1}{2}[\xi + f^L(a(\xi))]. \qquad \square$$

REMARK 5. *If $Df^0(0) = -p(0)$, as is the case when $f^0(a) = -P(a)$ or a Taylor approximation to $P(a)$, one has $Dg^0(0) = 0$ implying strong damping at 0. The situation is similar at L if $Df^L(0) = p(0)$.*

In order to obtain continuously differentiable solutions of (4.4) with initial conditions (3.4) and boundary conditions (4.12), we need to impose the compatibility conditions

$$
\begin{aligned}
& v^0(0) = f^0(a^0(0)), \quad v^0(L) = f^L(a^0(L)), \\
& \frac{g}{b(\overline{A} + a^0(0))} Da^0(0) + (\overline{V} + v^0(0)) Dv^0(0) \\
& \qquad = Df^0(a^0(0)) \left[((\overline{V} + v^0(0)) Da^0(0) + (\overline{A} + a^0(0)) Dv^0(0) \right], \\
& \frac{g}{b(\overline{A} + a^0(L))} Da^0(L) + (\overline{V} + v^0(L)) Dv^0(L) \\
& \qquad = Df^L(a^0(L)) \left[(\overline{V} + v^0(L)) Da^0(L) + (\overline{A} + a^0(L)) Dv^0(L) \right].
\end{aligned} \tag{4.13}
$$

These intricate, but obviously necessary, conditions are automatically satisfied for initial data with compact support in $(0, L)$. At this point, the following generalization of Theorem 1 of [2] is a direct consequence of Theorem 2 of [5] or, alternatively, of Theorem 1.3 of [8].

THEOREM 2. *Consider the St. Venant system* (4.4) *with boundary conditions* (4.12) *and initial conditions* (4.9) *satisfying the compatibility conditions* (4.13). *Suppose that*

$$
|Dg^0(0) Dg^L(0)| = \left| \frac{Df^0(0) + p(0)}{Df^0(0) - p(0)} \frac{Df^L(0) - p(0)}{Df^L(0) + p(0)} \right| < 1. \tag{4.14}
$$

Then if $\| (a^0, v^0) \|_1$ *is sufficiently small, there exists a unique continuously differentiable solution* $(a(x,t), v(x,t))$ *to the problem which is defined for all positive* t *and satisfies an estimate*

$$
\| (a(\cdot, t), v(\cdot, t)) \|_1 < Ce^{-\alpha t} \| (a^0, v^0) \|,
$$

where C *and* α *are suitable positive constants.*

REMARK 6. *In the notation of* [2], *the boundary conditions* (2.16) *of that paper can be written*

$$
\begin{aligned}
u_0 &= -\left(\frac{\overline{V}}{2} + \lambda_0 \right) \frac{y_0 - \overline{y}}{\overline{y} + (y_0 - \overline{y})} \triangleq f^0(y_0 - \overline{y}), \\
u_L &= -\left(\frac{\overline{V}}{2} - \lambda_L \right) \frac{y_L - \overline{y}}{\overline{y} + (y_L - \overline{y})} \triangleq f^L(y_L - \overline{y}).
\end{aligned}
$$

The variables y_0, y_L, *and* $\overline{y}$ *correspond to* $A(0,t)$, $A(L,t)$, *and* $\overline{A}$, *with* $y_0 - \overline{y}$ *corresponding to* $a(0,t)$, *etc. In our notation,* $u_0 = u(0,t)$ *and* $u_L = u(L,t)$. *The quantities calculated in the proof of their Theorem* 1 *correspond directly to our calculation of* $Dg^0(0)$ *and* $Dg^L(0)$. *Our result is more general in that it allows for a broad class of boundary conditions and does not require rectangular cross sections. For this purpose, the use of variables* A *and* V *is advantageous since they lead to equations in divergence form, which is not the case if one uses the variables* h *and* V *except in the case of rectangular cross sections. We, however, do not relate our boundary conditions to Liapunov conditions as is done in* [2]. *The nice discussion of boundary conditions in* [2] *justify the assertion that control of the velocity variables can be achieved by means of adjusting sluice gate heights (through* (2.12) *of that paper).*

Next we show that the use of an absorbing feedback boundary condition takes the perturbations a and u to zero in finite time.

THEOREM 3. *Consider the St. Venant system* (4.4) *with boundary conditions*

$$v(0,t) = f^0(a(0,t)) \quad \textit{and} \quad v(L,t) = P(a(L,t)) \tag{4.15}$$

(where $Df^0(0) \neq p(0)$*) and initial conditions* (4.9) *satisfying the compatibility conditions* (4.13). *Let* $T_\star$ *satisfy*

$$T_\star > \frac{L}{|\overline{V} - \overline{\beta}|} + \frac{L}{\overline{V} + \overline{\beta}}.$$

Then, if $\| (a^0, v^0) \|_1$ *is sufficiently small, the corresponding solution is defined for all positive times, and, in fact,*

$$a(\cdot, t) \equiv 0 \quad \textit{and} \quad v(\cdot, t) \equiv 0 \quad \textit{for } t \geq T_\star.$$

Proof. In this case, we have that $\xi_-(L,t) \equiv 0$ or $g^L(\cdot) \equiv 0$ so that condition (4.14) of Theorem 2 is automatically satisfied, and, therefore, for small enough initial data, solutions are defined for all t and decay exponentially.

That the solutions in fact vanish for $t \geq T_\star$ follows from the following observations. Since $\xi_-(x,t)$ is constant along characteristics of negative slope, this function vanishes identically for all (x,t) lying above the characteristic of negative slope emanating from $(L,0)$. If that characteristic reaches $x = 0$ at time T_1, say, one can conclude from the boundary condition at $x = 0$ that $\xi_+(0,t) \equiv 0$ for $t \geq T_1$, and hence $\xi_+(x,t)$ vanishes identically above the characteristic of positive slope starting from $(0, T_1)$ which meets $x = L$ at time $T_1 + T_2$. For $t \geq T_1 + T_2$, both Riemann invariants and hence also $a(x,t)$ and $v(x,t)$ vanish identically. To estimate the "traverse times," we note that, for sufficiently small initial data, the characteristics lie arbitrarily close to those of the equilibrium given by the linear functions (4.8), and hence the traverse times T_1 and T_2 can be made arbitrarily close to

$$\frac{L}{|\overline{V} - \overline{\beta}|} \quad \text{and} \quad \frac{L}{\overline{V} + \overline{\beta}},$$

respectively, for sufficiently small initial data. The result follows. ∎

REMARK 7. *If the absorbing boundary condition is also imposed at* $x = 0$*, one can choose a smaller value of* $T_\star$ *satisfying*

$$T_\star > \max \left\{ \frac{L}{|\overline{V} - \overline{\beta}|}, \frac{L}{\overline{V} + \overline{\beta}} \right\}.$$

REMARK 8. *If one does not impose absorbing boundary conditions, which involve the somewhat complicated function* $P(a)$*, one can also expect rapid stabilization for boundary functions* f^0 *and* f^L *satisfying* $Df^0(0) = -p(0)$ *and* $Df^L(0) = p(0)$ *and, in particular, for the very simple linear conditions*

$$v(0,t) = -p(0)a(0,t) \quad \textit{and} \quad v(L,t) = p(0)a(L,t)$$

with $p(0) = \sqrt{g/[\overline{A}\,\overline{b}]}$. *In fact, one can prove that, with such a choice of boundary conditions, one can achieve an arbitrary exponential rate of decay for sufficiently small initial data.*

5. Stabilization and null controllability for a star configuration of canals. Now we consider a star configuration of canals and perturbations $\mathbf{A} = \overline{\mathbf{A}} + \mathbf{a}$ and $\mathbf{V} = \overline{\mathbf{V}} + \mathbf{v}$ of equilibrium states $\overline{\mathbf{A}}, \overline{\mathbf{V}}$ constructed at the end of the previous section. We assume that the equilibrium flow is subcritical on each canal so that this will continue to be the case for small perturbations. The flow in each canal is governed by (4.4) with all the quantities indexed by i and $x \in [0, L_i]$. It is convenient to suppose that, for each i, the parameter value $x = 0$ corresponds to that end of the ith canal which lies at the multiple node. It is then useful to parametrize all the canals with a parameter x over a common interval $[0, L]$, where L could, for example, be the average canal length. For the ith canal, this would entail a parameter change $x \mapsto L/L_i\, x$, and the system corresponding to that canal becomes

$$\partial_t \begin{pmatrix} a_i \\ v_i \end{pmatrix} + \frac{L}{L_i} \begin{pmatrix} \overline{V}_i + v_i & \overline{A}_i + a_i \\ g/b_i(\overline{A}_i + a_i) & \overline{V}_i + v_i \end{pmatrix} \partial_x \begin{pmatrix} a_i \\ v_i \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{5.1}$$

The corresponding eigenvalues of the matrix are now

$$\lambda^i_\pm(a_i, v_i) = \frac{L}{L_i}(\overline{V}_i + v_i \pm \beta_i(a_i)).$$

The Riemann invariants $\xi^i_\pm$ are unaffected by the parameter change and given by the indexed form of (4.6).

This gives us a hyperbolic system of $2n$ equations on $[0, L] \times [0, T]$. Such systems are discussed in the appendix, which uses a different indexing of the Riemann invariants, setting

$$\xi_i = \begin{cases} \xi^i_+ & \text{for } i = 1, \dots, n, \\ \xi^{i-n}_- & \text{for } i = n+1, \dots, 2n. \end{cases}$$

In our case, the equations are pairwise decoupled, and it is convenient to stay with the indexing $\xi^i_\pm$. Initial conditions are given by

$$\mathbf{a}(x, 0) = \mathbf{a}^0(x), \quad \mathbf{v}(x, 0) = \mathbf{v}^0(x). \tag{5.2}$$

At $x = L$, we can introduce decoupled boundary conditions acting independently on each canal:

$$v_i(L, t) = f^L_i(a_i(L, t)) \quad \text{or} \quad \xi^i_-(L, t) = g^L_i(\xi^i_+(L, t)). \tag{5.3}$$

The coupling between the variables occurs through the multiple node conditions which translate into a boundary condition at $x = 0$. Let

$$\begin{aligned} S_i(a_i, v_i) &\triangleq \frac{1}{2}(\overline{V}_i + v_i)^2 + g h_i(\overline{A}_i + a_i), \\ Q_i(a_i, v_i) &\triangleq (\overline{A}_i + a_i)(\overline{V}_i + v_i). \end{aligned}$$

Then we have the following set of n multiple node conditions in $2n$ variables holding at $(0, t)$:

$$\begin{cases} S_i(a_i, v_i) - S_n(a_n, v_n) = 0 \quad \text{for } i = 1, \dots, n-1, \\ \sum_{i=1}^n \epsilon_i Q_i(a_i, v_i) = 0. \end{cases} \tag{5.4}$$

To show that this system represents admissible boundary conditions at $x = 0$, we need to show that implicitly these equations determine

$$\boldsymbol{\xi}_+(0,t) \triangleq (\xi_+^1(0,t), \ldots, \xi_+^n(0,t))$$

as a function of

$$\boldsymbol{\xi}_-(0,t) = (\xi_-^1(0,t), \ldots, \xi_-^n(0,t)).$$

To do this, we have to show that the following Jacobian matrix is nonsingular:

$$J_+ \triangleq \begin{pmatrix} \partial_{\xi_+^1} S_1 & 0 & 0 & \ldots & 0 & -\partial_{\xi_+^n} S_n \\ 0 & \partial_{\xi_+^2} S_2 & 0 & \ldots & 0 & -\partial_{\xi_+^n} S_n \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \partial_{\xi_+^{n-1}} S_{n-1} & -\partial_{\xi_+^n} S_n \\ \partial_{\xi_+^1} Q_1 & \partial_{\xi_+^2} Q_2 & \partial_{\xi_+^3} Q_3 & \ldots & \partial_{\xi_+^{n-1}} Q_{n-1} & \partial_{\xi_+^n} Q_n \end{pmatrix},$$

where all the partial derivatives are evaluated at $(a, v) = (0, 0)$. We have, since $v_i = \xi_{i+} + \xi_{i-}$ and $P_i(a_i) = \xi_{i+} - \xi_{i-}$, that

$$\partial_{\xi_+^i} S_i = \partial_{a_i} S_i \partial_{\xi_+^i} a_i + \partial_{v_i} S_i \partial_{\xi_+^i} v_i = \frac{g}{b_i(\overline{A}_i + a_i)} \frac{1}{p_i(\overline{A}_i + a_i)} + (\overline{V}_i + v_i),$$

and so, noting that

$$\frac{g}{\overline{b}_i \overline{p}_i} = \sqrt{g\overline{A}_i/\overline{b}_i} = \overline{\beta}_i,$$

we get

$$\partial_{\xi_+^i} S_i(0,0) = \overline{V}_i + \frac{g}{\overline{b}_i \overline{p}_i} = \overline{V}_i + \overline{\beta}_i.$$

Similarly,

$$\partial_{\xi_+^i} Q_i(0,0) = \overline{A}_i + \frac{\overline{v}_i}{\overline{p}_i} = \frac{1}{\overline{p}_i}(\overline{V}_i + \overline{\beta}_i).$$

We note also that $\overline{V}_i + \overline{\beta}_i = (L_i/L)\overline{\lambda}_+^i$, where $\overline{\lambda}_+^i$ is the positive eigenvalue of the system associated with the ith canal after reparametrization. We introduce the diagonal matrices

$$\overline{\Lambda}_\pm \triangleq \operatorname{diag}\left(\frac{L_1}{L}\overline{\lambda}_\pm^1, \ldots, \frac{L_n}{L}\overline{\lambda}_\pm^n\right) = \operatorname{diag}(\overline{V}_1 \pm \overline{\beta}_1, \ldots, \overline{V}_n \pm \overline{\beta}_n),$$

where $\operatorname{diag}(\mu_1, \ldots, \mu_n)$ denotes the $n \times n$ diagonal matrix with specified diagonal elements μ_i. We also introduce the two matrices

$$G_\pm \triangleq \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 & -1 \\ 0 & 1 & 0 & \ldots & 0 & -1 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1 & -1 \\ \pm\frac{1}{\overline{p}_1} & \pm\frac{1}{\overline{p}_2} & \pm\frac{1}{\overline{p}_3} & \ldots & \pm\frac{1}{\overline{p}_{n-1}} & \pm\frac{1}{\overline{p}_n} \end{pmatrix}.$$

One now easily checks that the Jacobian matrix of the system (5.4) with respect to $\boldsymbol{\xi}_+$ is of the form $J_+ = G_+\overline{\Lambda}_+$. One can similarly evaluate the Jacobian of the system with respect to $\boldsymbol{\xi}_-$ to get $J_- = G_-\overline{\Lambda}_-$. By the implicit function theorem, the system (43) of node conditions can therefore be written as an admissible boundary condition at 0:

$$\boldsymbol{\xi}_+(0,t) = \mathbf{g}^0(\boldsymbol{\xi}_-(0,t)), \tag{5.5}$$

with

$$\nabla_{\boldsymbol{\xi}_-}\mathbf{g}^0(\mathbf{0}) = J_+^{-1}J_- = \overline{\Lambda}_+^{-1}[G_+^{-1}G_-]\overline{\Lambda}_-. \tag{5.6}$$

Now the following theorem concerning the stabilizability (and indeed the existence) of solutions for our star configuration of canals follows easily from Theorem 1.3 of [8].

THEOREM 4. *Consider the systems* (5.1) *with boundary conditions* (5.4) *(equivalent to* (5.5)*) and* (5.3) *holding at* $x = 0$ *and* $x = L$*, respectively, and initial conditions* (5.2) *with data satisfying the appropriate compatibility conditions. Suppose that*

$$\begin{aligned} &\| \nabla_{\boldsymbol{\xi}_-}\mathbf{g}^0(\mathbf{0}) \|_\infty \max_{1\le i\le n} |Dg_i^L(0)| \\ &\quad = \| \overline{\Lambda}_+^{-1}[G_+^{-1}G_-]\overline{\Lambda}_- \|_\infty \max_{1\le i\le n} \left| \frac{Df_i^L(0) - p_i(0)}{Df_i^L(0) + p_i(0)} \right| < 1. \end{aligned} \tag{5.7}$$

Then, for $\| (\mathbf{a}^0, \mathbf{v}^0) \|_1$ *sufficiently small, there exists a unique continuously differentiable solution to the problem which is defined for all positive* t *and satisfies*

$$\| (\mathbf{a}(\cdot,t), \mathbf{v}(\cdot,t)) \|_1 < Ce^{-\alpha t} \| (\mathbf{a}^0, \mathbf{v}^0) \|,$$

where C *and* α *are suitable positive constants.*

In order to apply the cited theorem, we recall the following definition of the "minimal characteristic number" of a square matrix used by Li in formulating the hypothesis of boundary damping:

$$\| A \|_{\min} = \inf \left\{ \| \gamma A \gamma^{-1} \|_\infty \mid \gamma \text{ is an invertible, square, diagonal matrix} \right\}.$$

The damping hypothesis required by Li is that $\| \theta \|_{\min} < 1$, where

$$\theta = \begin{pmatrix} 0 & \nabla_{\boldsymbol{\xi}_+}\mathbf{g}^L(\mathbf{0}) \\ \nabla_{\boldsymbol{\xi}_-}\mathbf{g}^0(\mathbf{0}) & 0 \end{pmatrix}$$

with $\nabla_{\boldsymbol{\xi}_-}\mathbf{g}^0(\mathbf{0})$ given by (5.6) and

$$\nabla_{\boldsymbol{\xi}_+}\mathbf{g}^L(\mathbf{0}) = \operatorname{diag}(\dots, Dg_i^L(0), \dots) = \operatorname{diag}\left(\dots, \frac{Df_i^L(0) - p_i(0)}{Df_i^L(0) + p_i(0)}, \dots \right).$$

Note that $\| A \|_{\min} = \| \gamma A \gamma^{-1} \|_{\min}$, that $\| A \|_{\min} \le \| A \|_\infty$, and that, if A is diagonal, $\| A \|_{\min} = \| A \|_\infty = \max_{1\le i\le n} |A_{ii}|$. Now consider partitioned matrices A of the form

$$A = \begin{pmatrix} 0 & B \\ C & 0 \end{pmatrix},$$

where both B and C are $n \times n$ matrices and B is diagonal. Using diagonal matrices of the form $\gamma = \mathrm{diag}(\lambda\gamma_1, \gamma_1)$, where γ_1 is a diagonal $n \times n$ invertible matrix and λ is an arbitrary positive scalar, it is not difficult to obtain the following estimate using the definition of $\|\cdot\|_{\min}$ and the previously stated properties:

$$\|A\|_{\min} \leq \|C\|_{\min} \max_{1\leq i\leq n} |B_{ii}|.$$

This gives estimate (5.7).

REMARK 9. *With some additional care, one can prove the following estimate, which can replace* (5.7) *as damping hypothesis:*

$$\|G_+^{-1}G_-\|_{\min} \max_{1\leq i\leq n} \frac{\overline{V}_i + \overline{\beta}_i}{|\overline{V}_i - \overline{\beta}_i|} \left| \frac{Df_i^L(0) - p_i(0)}{Df_i^L(0) + p_i(0)} \right| < 1.$$

REMARK 10. *It would be helpful to have a good estimate for* $\|G_+^{-1}G_-]\|_\infty$ *or for* $\|G_+^{-1}G_-]\|_{\min}$. *Since* $G_- = \mathrm{diag}(1, \ldots, 1, -1)G_+$, *one has*

$$G_+^{-1}G_- = G_+^{-1}\mathrm{diag}(1, \ldots, 1, -1)G_+$$

so that the eigenvalues of $G_+^{-1}G_-$ *are* 1 *with multiplicity* $n-1$ *and* -1 *with multiplicity* 1. *Unfortunately this does not imply anything about* $\|G_+^{-1}G_-\|_\infty$, *and what it may imply for* $\|G_+^{-1}G_-\|_{\min}$ *is not yet clear. However, in some situations (all canals and all data equal), the two norms coincide.*

Finally, it is also not difficult to see that the proof of Theorem 3 can be adapted to prove the following result on null controllability for the star system.

THEOREM 5. *Consider the systems* (5.1) *with boundary conditions* (5.4) *(equivalent to* (5.5)*) holding at* $x = 0$, *absorbing boundary conditions*

$$v_i(L,t) = P_i(a_i(L,t)) \quad \textit{for } i = 1, \ldots, n,$$

and holding at $x = L$ *and initial conditions* (5.2) *with data satisfying the appropriate compatibility conditions. Let* $T_\star$ *satisfy*

$$T_\star > \max_{1\leq i\leq n} \frac{L^2}{L_i(|\overline{V}_i - \overline{\beta}_i|)} + \max_{1\leq i\leq n} \frac{L^2}{L_i(\overline{V}_i + \overline{\beta}_i)}.$$

Then, for small enough initial data, the solution is defined for all positive times and, in fact,

$$\mathbf{a}(\cdot, t) \equiv \mathbf{0} \quad \textit{and} \quad \mathbf{v}(\cdot, t) \equiv \mathbf{0} \quad \textit{for } t \geq T_\star.$$

Acknowledgments. We note that an "open loop" controllability result for a *single* link being equivalent to the system governing the Riemann invariants discussed in section 4 has been given by Li, Rao, and Jin [7]. The authors thank the referees for their helpful suggestions.

REFERENCES

[1] M. CIRINÀ, *Nonlinear hyperbolic problems with solutions on preassigned sets*, Michigan Math J., 17 (1970), pp. 193–209.

[2] J. M. Coron, B. d'Ándréa-Novel, and G. Bastin, *A Lyapunov approach to control irrigation canals modeled by Saint-Venant equations*, in Proceedings of the 5th European Control Conference, Karlsruhe, Germany, 1999, CD-ROM, paper F1008-5.

[3] R. Courant and D. Hilbert, *Methods of Mathematical Physics, Volume* II, Interscience Publishers, New York, London, 1962.

[4] J. A. Cunge, F. M. Holly, and A. Verwey, *Practical Aspects of Computational River Hydraulics*, Pitman, Boston, 1980.

[5] J. M. Greenberg and T.-T. Li, *The effect of boundary damping for the quasilinear wave equation*, J. Differential Equations, 52 (1984), pp. 66–75.

[6] J. E. Lagnese, G. Leugering, and E. J. P. G. Schmidt, *Modeling, Analysis and Control of Dynamic Elastic Multi-Link Structures*, Birkhäuser, Boston, Basel, Berlin, 1994.

[7] T.-T. Li, B. P. Rao, and Y. Jin, *Solution C^1 semiglobale et contrôlabilité exacte frontière de systèmes hyperboliques quasilinéaires réductibles*, C. R. Acad. Sci. Paris Sér. I Math., 330 (2000), pp. 205–210.

[8] T.-T. Li, *Global Classical Solutions for Quasilinear Hyperbolic Systems*, RAM Res. Appl. Math. 32, Masson, Paris, John Wiley and Sons, Chichester, UK, 1994.

[9] M. E. Taylor, *Partial Differential Equations, Vol.* III, Appl. Math. Sci. 117, Springer-Verlag, New York, 1996.

[10] B. de Saint-Venant, *Théorie du mouvement non-permanent des eaux avec application aux crues des rivières et à l'íntroduction des marées dans leur lit*, Comptes Rendus Academie des Sciences, 73 (1871), pp. 148–154, 237–240.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 1, pp. 181–216

ON REACHABILITY UNDER UNCERTAINTY*

A. B. KURZHANSKI† AND P. VARAIYA‡

Abstract. The paper studies the problem of reachability for linear systems in the presence of uncertain (unknown but bounded) input disturbances that may also be interpreted as the action of an adversary in a game-theoretic setting.

It defines possible notions of reachability under uncertainty emphasizing the differences between reachability under open-loop and closed-loop control. Solution schemes for calculating reachability sets are then indicated. The situation when observations arrive at given isolated instances of time leads to problems of anticipative (maxmin) or nonanticipative (minmax) piecewise open-loop control with corrections and to the respective notions of reachability. As the number of corrections tends to infinity, one comes in both cases to reachability under nonanticipative feedback control. It is shown that the closed-loop reach sets under uncertainty may be found through a solution of the forward Hamilton–Jacobi–Bellman–Isaacs (HJBI) equation.

The basic relations are derived through the investigation of superpositions of value functions for appropriate sequential maxmin or minmax problems of control.

Key words. reachability, reach sets, differential inclusions, alternated integral, funnel equations, open-loop control, closed-loop control, dynamic programming, uncertainty, differential games, HJBI equation

AMS subject classifications. 34HO5, 34G25, 35F10, 49L, 49N70, 91A23

PII. S0363012999361093

Introduction. Recent developments in real-time automation have promoted new interest in the reachability problem—the computation of the set of states reachable by a controlled process through available controls. Being one of the basic problems of control theory, it was studied from the very beginning of investigations in this field (see [19]). The problem was usually studied in the absence of disturbances, under complete information on the system equations and the constraints on the control variables. It was shown, in particular, that the set of states reachable at a given time t under bounded controls is one and the same whether one uses open-loop or closed-loop (feedback) controls. It was also indicated that these "reachability sets" could be calculated as level sets for the (perhaps generalized) solutions to a "forward" Hamilton–Jacobi–Bellman equation [19], [20], [3], [16], [18].

However, in reality, the situation may be more complicated. Namely, if the system is subject to unknown but bounded disturbances, it may become necessary to compute the set of states reachable *despite the disturbances* or, if exact reachability is impossible, to find guaranteed errors for reachability.

These questions have implicitly been present in traditional studies on feedback control under uncertainty for continuous-time systems [11], [28], [4], [10], [13]. They have also appeared in studies on hybrid and other types of transition systems [1], [29], [22], [5].

This leads us to the topic of the present paper, which is the investigation of

*Received by the editors September 11, 1999; accepted for publication (in revised form) November 1, 2001; published electronically April 26, 2002. This research was supported by National Science Foundation grant ECS 9725148 and ONR grant N00014-98-1-0585.

http://www.siam.org/journals/sicon/41-1/36109.html

†Moscow State University. Current address: Department of Electrical Engineering and Computer Science, University of California at Berkeley, Berkeley, CA 94720 (kurzhans@eecs.berkeley.edu).

‡Department of Electrical Engineering and Computer Science, University of California at Berkeley, Berkeley, CA 94720 (varaiya@eecs.berkeley.edu).

reachability under uncertainty for continuous-time linear control systems subjected to unknown input disturbances, with prespecified geometric (hard) bounds on the controls and the unknowns. The paper indicates various notions of reachability, studies the properties of respective reach sets, and indicates routes for calculating them.

The first question here is to distinguish whether reachability under open-loop and closed-loop controls yields the same reach sets. Indeed, since closed-loop control is based on better information, namely, on the possibility of continuous on-line observations of the state space variable (with no knowledge of the disturbance), it must produce, generally speaking, a result which is at least "not worse," for example, than the one by an open-loop control which allows no such observations but only the knowledge of the initial state, with no knowledge of the disturbance. An open-loop control of the latter type is further referred to as "nonanticipative."

However, there are many other possibilities of introducing open-loop or piecewise open-loop controls, with or without the availability of some type of isolated on-line measurements of the state space variable, as well as with or without an a priori knowledge of the disturbance. Thus, in order to study the reachability problem in detail, we introduce a hierarchy of reachability problems formulated under an array of different "intermediate" information conditions. These are formulated in terms of some auxiliary extremal problems of the maxmin or minmax type.

Starting with open-loop controls, we first distinguish the case of *anticipative control* from *nonanticipative control.* The former, for example, is when a reachable set, from a given initial state x^0, at given time τ, is defined as the set $X_\mu^- = X^-(\tau, t_0, x^0, \mu)$ of such states x, that for any admissible disturbance given in advance, for the whole interval under consideration, there exists an admissible control that steers the system to a μ-neighborhood $\mathcal{B}_\mu(x) = \{z : (z-x, z-x) \leq \mu^2\}$. Here the respective auxiliary extremal problem is of the maxmin type (maximum in the disturbance and minimum in the control). On the other hand, for the *latter,* the disturbance is not known in advance. Then the reachability set from a given initial state is defined as the set $X_\mu^+ = X^+(\tau, t_0, x^0, \mu)$ of such states x whose μ-neighborhoods $\mathcal{B}_\mu(x)$ may be reached with the same admissible control *for all* admissible disturbances. Now the respective auxiliary problem is of the minmax type.

It is shown that always $X_\mu^+ \subseteq X_\mu^-$ and that the closed-loop reach set $\mathcal{X}_\mu = X(\tau, t_0, x^0, \mu)$ attained under *nonanticipative but feedback control* lies in between, namely,

$$X_\mu^+ \subseteq \mathcal{X}_\mu \subseteq X_\mu^-.$$

There are also some intermediate situations when the observations of the state space variable arrive at given N isolated instants of time. In that case, one has to deal with *reachability under possible corrections* of the control at these N time instants. Here again we distinguish between corrections implemented through anticipative control (when the future disturbance is known for each time interval in between the corrections) and nonanticipative control (when it is unknown). The respective extremal problems are of sequential maxmin and minmax types accordingly, and the controls are piecewise open-loop: at isolated time instants of correction, there arrives information on the state space variable, while in between these instants, the control is open-loop (either anticipative or not). Both cases produce respective sequences $X_{\mu,N}^- = X_N^-(\tau, t_0, x^0, \mu), X_{\mu,N}^+ = X_N^+(\tau, t_0, x^0, \mu)$ of "piecewise open-loop reach sets (OLRSs)." The relative positions of the reach sets in the hierarchical scheme are as

follows:

$$X_\mu^+ \subseteq X_{\mu,N}^+ \subseteq \mathcal{X}_\mu \subseteq X_{\mu,N}^- \subseteq X_\mu^-.$$

Finally, in the limit, as the number of corrections N tends to infinity, *both sequences* of reachability sets *converge to the closed-loop reach set.*[1]

The adopted scheme is based on constructing superpositions of value functions for open-loop control problems. In the limit these relations reflect the *principle of optimality* under set-membership uncertainty. This principle then allows one to describe the closed-loop reach set as a level set for the solution to the *forward HJBI (Hamilton–Jacobi–Bellman–Isaacs) equation*. The final results are then presented either in terms of value functions for this equation or in terms of set-valued relations.

Schemes of such type have been used in synthesizing solution strategies for differential games and related problems and were constructed in backward time [23], [12], [27], [28].

The topics of this paper were motivated by applied problems and also by the need for a theoretical basis for further algorithmic schemes.

1. Uncertain dynamics. Reachability under open-loop controls. In this section, we introduce the system under consideration and define two types of open-loop reachability sets. Namely, we discuss *reachability* under unknown but bounded disturbances for the system

$$\dot{x} = A(t)x + B(t)u + C(t)v(t), \tag{1}$$

with continuous matrix coefficients $A(t), B(t), C(t)$. Here $x \in \mathbb{R}^n$ is the *state* and $u \in \mathbb{R}^p$ is the *control* that may be selected either as an *open-loop control* OLC, a Lebesgue-measurable function of time t, restricted by the inclusion

$$u(t) \in \mathcal{P}(t), \tag{2}$$

or as a *closed-loop control* CLC, a *set-valued strategy*

$$u = \mathcal{U}(t,x) \subseteq \mathcal{P}(t). \tag{3}$$

Here $v \in \mathbb{R}^q$ is the unknown *input disturbance* with values

$$v(t) \in \mathcal{Q}(t), \tag{4}$$

and $\mathcal{P}(t), \mathcal{Q}(t)$ are set-valued continuous functions with convex compact values ($\mathcal{P} \in comp\mathbb{R}^p$, $\mathcal{Q} \in comp\mathbb{R}^q$).[2]

The class of OLCs $u(\cdot)$ bounded by inclusion (2) is denoted by $U_{\mathcal{O}}$, and the class of input disturbances $v(\cdot)$ bounded by (4) is denoted by $V_{\mathcal{O}}$. The strategies $\mathcal{U}$ are taken to be in $U_{\mathcal{C}}$—the class $U_{\mathcal{C}}$ of CLCs that are multivalued maps $\mathcal{U}(t,x)$ bounded by the inclusion (3), which guarantee the solutions to (1), $u = \mathcal{U}(t,x)$ (which now turns into a differential inclusion) for any Lebesgue-measurable function $v(\cdot)$.[3]

[1] As indicated in what follows, this is true when all of the sets involved are nonempty and when the problems satisfy some regularity conditions.

[2] Set-valued functions $\mathcal{P}(t)$ with values in $comp\mathbb{R}^p$ are defined to be continuous in $t \in [t_0, t_1]$ if the support functions $\rho(l|\mathcal{P}(t)) = \max\{(l,x)|x \in \mathcal{P}(t)\}$ are continuous in $t \in [t_0, t_1]$ uniformly in $\{l : (l,l) \leq 1\}$.

[3] For example, the class of set-valued functions with values in $comp\mathbb{R}^n$, upper semicontinuous in x and continuous in t.

We distinguish two types of OLRSs—the *maxmin* type and the *minmax* type. As we will see in the next section, the names *maxmin* and *minmax* assigned to these sets are due to the underlying optimization problems used for their calculation.

DEFINITION 1.1. *An OLRS of the* maxmin type *(from set $X^0 = X(t_0)$, at time $\tau \geq t_0$) is the set $X^-(\tau, t_0, X^0)$ of all vectors x such that for every disturbance $v(t) \in \mathcal{Q}(t)$, there exist an initial state $x^0 \in X^0$ and an OLC $u(t) \in \mathcal{P}(t)$ which steer the trajectory $x(t)$, $t_0 \leq t \leq \tau$, from state $x^0 = x(t_0)$ to state*

$$x(\tau) = x. \tag{5}$$

The set X^0 is assumed to be convex and compact ($X^0 \in comp\mathbb{R}^n$).

If $X^-(\tau, t_0, X^0)$ turns out to be empty, one may introduce the open-loop *μ-reachable set* $X^-(\tau, t_0, X^0, \mu)$ as in Definition 1.1 except that (5) is replaced by

$$x(\tau) \in \mathcal{B}_\mu(x).$$

Here

$$\mathcal{B}_\mu(x) = \{x : (x(\tau) - x, x(\tau) - x) \leq \mu^2\} = x + \mathcal{B}_\mu(0), \quad \mu \geq 0,$$

is the ball of radius μ with center x.

Thus the OLRS $X^-(\tau, t_0, X^0)$ of the maxmin type is the set of points $x \in \mathbb{R}^n$ that can be reached, for any disturbance $v(t) \in \mathcal{Q}(t)$ given in advance, for the whole interval $t_0 \leq t \leq \tau$, from some point $x(t_0) \in X^0$, through some open-loop control $u(\cdot) \in U_\mathcal{O}$.

The open-loop *μ-reach set* (μ-OLRS) $X^-(\tau, t_0, X^0, \mu)$ is the set of points $x \in \mathbb{R}^n$ whose μ-neighborhood $\mathcal{B}_\mu(x)$ may be reached, for any disturbance $v(t)$ given in advance, through some $x(t_0) \in X^0, u(\cdot) \in U_\mathcal{O}$.

By taking $\mu \geq 0$ large enough, we may assume $X^-(\tau, t_0, X^0, \mu) \neq \emptyset$.

Denote $x(t, t_0, x^0 | u(\cdot), v(\cdot))$ to be the unique trajectory corresponding to $x(t_0) = x^0$, control $u(\cdot)$, and disturbance $v(\cdot)$. Then

$$\cup\{x(t, t_0, x^0 | u(\cdot), v(\cdot)) | x(t_0) \in X^0,\ u(\cdot) \in U_\mathcal{O}\} = X(t, t_0, X^0 | \mathcal{P}(\cdot), v(\cdot))$$

is the *reach set* in the variable $u(\cdot) \in U_\mathcal{O}$ (at time t from set X^0) with fixed disturbance input $v(\cdot)$.

LEMMA 1.1.

$$X^-(\tau, t_0, X^0) = \cap\{X(t, t_0, X^0 | \mathcal{P}(\cdot), v(\cdot)) | v(\cdot) \in V_\mathcal{O}\}. \tag{6}$$

This formula follows from Definition 1.1. Recall the definition of the geometrical (Minkowski) difference $\mathcal{P} \dot{-} \mathcal{Q}$ of sets $\mathcal{P}, \mathcal{Q}$,

$$\mathcal{P} \dot{-} \mathcal{Q} = \{c : c + \mathcal{Q} \subseteq \mathcal{P}\}.$$

Then, directly from (1), one gets

$$X^-(\tau, t_0, X^0) = \left(S(t_0, \tau)X^0 + \int_{t_0}^{\tau} S(s, \tau)B(s)\mathcal{P}(s)ds\right) \dot{-} \int_{t_0}^{\tau} S(s, \tau)(-C(s)\mathcal{Q}(s))ds. \tag{7}$$

Here $S(s,t)$ stands for the matrix solution of the adjoint equation

$$\partial S(s,t)/\partial s = -S(s,t)A(t),\ S(t,t) = I.$$

In other words, the set

$$X^-(\tau, t_0, X^0) = X(t, t_0, X^0|\mathcal{P}(\cdot), \{0\}) \dot{-} X(t, t_0, 0|\{0\}, \mathcal{Q}(\cdot))$$

is the geometric difference of two "ordinary" reach sets, namely, the set $X(t, t_0, X^0|\mathcal{P}(\cdot), \{0\})$ taken from $X(t_0) = X^0$ and calculated in the variable u, with $v(t) \equiv 0$, and the set $X(t, t_0, 0|\{0\}, \mathcal{Q}(\cdot))$ taken from $x(t_0) = 0$ and calculated in the variable v, with $u(\cdot) \equiv 0$. This simple geometrical interpretation is, of course, due to the linearity of (1).

For the μ-reachable set, we have the following lemma.

LEMMA 1.2. *The set $X^-(\tau, t_0, X^0, \mu)$ may be expressed as*

$$\begin{aligned} X^-(\tau, t_0, X^0, \mu) &= \cap\{X(t, t_0, X^0|\mathcal{P}(\cdot), v(\cdot)) + \mathcal{B}_\mu(0)|v(\cdot) \in V_{\mathcal{O}}\} \\ &= (X(t, t_0, X^0|\mathcal{P}(\cdot), \{0\}) + \mathcal{B}_\mu(0)) \dot{-} X(t, t_0, 0|\{0\}, \mathcal{Q}(\cdot)) \end{aligned} \tag{8}$$

and also as

$$X^-(\tau, t_0, X^0, \mu_1) \subseteq X^-(\tau, t_0, X^0, \mu_2),\ \ \mu_1 \leq \mu_2.$$

Remark 1.1. Definition (8) of $X^-(\tau, t_0, X^0, \mu)$ may also be rewritten as

$$X^-(\tau, t_0, X^0, \mu) = \cap_v \cup_u \cup_{x^0} \{X(t, t_0, x^0|u(\cdot), v(\cdot)) + \mathcal{B}_\mu(0)|x^0 \in X^0, u(\cdot) \in U_{\mathcal{O}}, v(\cdot) \in V_{\mathcal{O}}\}.$$

We now define another class of OLRSs under uncertainty—the OLRS of the *min-max* type.

DEFINITION 1.2. *A μ-OLRS of the* minmax type *(from set $X^0 = X(t_0)$, at time $\tau \geq t_0$) is the set $X^+(\tau, t_0, X^0, \mu)$ of all x for each of which there exists a control $u(t) \in \mathcal{P}(t)$ that assigns to each $v(t) \in \mathcal{Q}(t)$ a vector $x^0 \in X^0$, such that the respective trajectory $x[t] = x(t, t_0, x^0|u(\cdot), v(\cdot))$ ends in $x[\tau] \in \mathcal{B}_\mu(x)$.*

Thus the μ-OLRS of minmax type consists of all x whose μ-neighborhood $\mathcal{B}_\mu(x)$ contains the states $x[\tau]$ generated by system (1) under some control $u(t) \in \mathcal{P}(t)$ and all $\{v(t) \in \mathcal{Q}(t), t_0 \leq t \leq \tau\}$ with $x^0 \in X^0$ selected depending on u, v.[4]

A reasoning similar to the above leads to the following lemma.

LEMMA 1.3. *The set $X^+(\tau, t_0, X^0, \mu)$ may be expressed as*

$$X^+(\tau, t_0, X^0, \mu) = \cup\{(X(\tau, t_0, X^0|u(\cdot), \{0\}) + \mathcal{B}_\mu(0)) \dot{-} X(t, t_0, 0|\{0\}, \mathcal{Q}(\cdot))|u(\cdot) \in \mathcal{U}_{\mathcal{O}}\} \tag{9}$$

and

$$X^+(\tau, t_0, X^0, \mu_1) \subseteq X^+(\tau, t_0, X^0, \mu_2), \quad \mu_1 \leq \mu_2.$$

Remark 1.2. Definition (9) of $X^+(\tau, t_0, X^0, \mu)$ may be rewritten as

$$\cup_u \cap_v \cup_{x^0} \{(x(t, t_0, x^0|u(\cdot), \{0\}) + \mathcal{B}_\mu(0)) - x(t, t_0, 0|\{0\}, v(\cdot))|x^0 \in X^0, u(\cdot) \in U_{\mathcal{O}}, v(\cdot) \in V_{\mathcal{O}}\}.$$

Direct calculation, based on the properties of set-valued operations, allows us to conclude the following.

[4] With $\mu = 0$ and X^0 single-valued, it usually turns out that $X^+_\mu = \emptyset$.

LEMMA 1.4. *When* $X^+(\tau, t_0, X^0, \mu), X^-(\tau, t_0, X^0, \mu)$ *are both nonempty for some* $\mu > 0$*, we have*

$$X^+(\tau, t_0, X^0, \mu) \subseteq X^-(\tau, t_0, X^0, \mu).$$

We shall now calculate the OLRSs defined above, using the techniques of convex analysis [25], [13], [16].

2. The calculation of OLRSs. Here we shall calculate the two basic types of OLRSs. The relations of this section will also serve as the basic elements for further constructions which will be produced as some superpositions of the relations of this section.

The calculations of this section and especially of later sections related to reachability under feedback control require a number of rather cumbersome calculations of geometrical (Minkowski) differences and their support functions. In order to simplify these calculations, we transform system (1) to a simpler form. Taking the transformation $z = S(t, t_0)x$, one gets

$$\dot{z} = B_1(t)u - C_1(t)v,$$

where $B_1(t) = S(t, t_0)B(t)$ and $C_1(t) = S(t, t_0)C(t)v$. (With this transformation, the terms in S will disappear from (7).)

Keeping the previous notations x, B, C for z, B_1, C_1, we thus come, without loss of generality, to the system

$$\dot{x} = B(t)u + C(t)v, \tag{10}$$

with the same constraints on u, v as before. For (10), consider the following two problems (where the condition $x(t_0) \in X^0$ is dropped).

Problem (I). Given a set X^0 and $x(t_0) \in \mathbb{R}^n$, find

$$V^-(\tau, x, \mu) = \max_v \min_u \min_{x(\tau)} d(x(t_0), X^0), \quad \tau \geq t_0,$$

under conditions $x(\tau) \in \mathcal{B}_\mu(x),\ u(\cdot) \in U_\mathcal{O},\ v(\cdot) \in V_\mathcal{O}$.

Problem (II). Given a set X^0 and $x(t_0) \in \mathbb{R}^n$, find

$$V^+(\tau, x, \mu) = \min_u \max_v \min_{x(\tau)} d(x(t_0), X^0), \quad \tau \geq t_0,$$

under conditions $x(\tau) \in \mathcal{B}_\mu(x), u(\cdot) \in U_\mathcal{O},\ v(\cdot) \in V_\mathcal{O}$.

Here

$$d^2(x, z) = (x - z, x - z),\ d(x, \mathcal{G}) = \min\{d(x, z) | z \in \mathcal{G})\},$$

and $\mathcal{G}$ is a closed set in $\mathbb{R}^n$. Thus

$$d(x, \mathcal{G}) = h_+(x, \mathcal{G}),$$

where $h_+(\mathcal{Q}, \mathcal{G})$ is the *Hausdorff semidistance* between compact sets $\mathcal{Q}, \mathcal{G}$, defined as

$$h_+(\mathcal{Q}, \mathcal{G}) \ = \ \max_x \min_z \{(x - z, x - z)^{1/2} | x \in \mathcal{Q}, z \in \mathcal{G}\}.$$

The *Hausdorff distance* is $h(\mathcal{Q}, \mathcal{M}) = \max\{h_+(\mathcal{Q}, \mathcal{G}), h_+(\mathcal{G}, \mathcal{Q})\}$.

In order to calculate the function V^- explicitly, we use the relations

$$x(t) = x(t_0) + \int_{t_0}^{t} (B(s)u(s) + C(s)v(s))ds \tag{11}$$

and (see [11], [16] for the next formula)

$$d(x, \mathcal{G}) = \max_l \{(l, x) - \rho(l|\mathcal{G})|(l, l) \leq 1\},^5$$

where

$$\rho(l|\mathcal{G}) = \sup\{(l, x)|x \in \mathcal{G}\}$$

is *the support function of* $\mathcal{G}$ [16]. (For compact $\mathcal{G}$, sup may be substituted by max.)

We thus need to calculate

$$V^-(\tau, x, \mu) = \max_v \min_u \min_{x(\tau)} \{d(x(t_0), X^0)|x(\tau) \in \mathcal{B}_\mu(x), u(\cdot) \in U_\mathcal{O}, v \in V_\mathcal{O}\},$$

which gives, after an application of the formula for $d(x, \mathcal{G})$ and of (11) and a further interchange of $\min_u, \min_{x(\tau)}$, and $\max_l$ (see [7]),

$$V^-(\tau, x, \mu) \tag{12}$$

$$= \max_l \left\{ (l, x) - \rho(l|X^0) - \mu(l, l)^{\frac{1}{2}} - \int_{t_0}^{\tau} (\rho(l|B(s)\mathcal{P}(s)) - \rho(-l|C(s)\mathcal{Q}(s))ds|(l, l) \leq 1 \right\}.$$

Due to (11), the last formula says simply that V^- is given by

$$V^-(\tau, x, \mu) = d(x, X^-(\tau, t_0, x^0, \mu)), \tag{13}$$

where

$$X^-(\tau, t_0, x^0, \mu) = \left(X^0 + \mathcal{B}_\mu(0) + \int_{t_0}^{\tau} B(t)\mathcal{P}(s)ds \right) \dot{-} \int_{t_0}^{\tau} (-C(s))\mathcal{Q}(s)ds. \tag{14}$$

It then follows that

$$X^-(\tau, t_0, X^0, \mu) = \{x : V^-(\tau, x, \mu) \leq 0\}, \tag{15}$$

and so (12) implies that $x \in X^-(\tau, t_0, X^0, \mu)$ iff

$$(l, x) \leq \rho(l|X^0) + \mu(l, l)^{\frac{1}{2}} + \int_{t_0}^{\tau} (\rho(l|B(t)\mathcal{P}(s)) - \rho(-l|C(t)\mathcal{Q}(s)))ds \quad \forall l \in \mathbb{R}^n.$$

(Of course, $\{x : V^-(\tau, x, \mu) \leq 0\} = \{x : V^-(\tau, x, \mu) = 0\}$. Even though $V^-(\tau, x, \mu)$ is nonnegative, we retain the notation $\{V^- \leq 0\}$ to suggest that for purposes of approximation it may be useful to consider $\{V^- \leq \epsilon\}$.) This gives, from the definitions of support function and geometrical difference,

[5] This formula is always interpreted as $d(x, \mathcal{G}) = \max\{0, \max_l\{(l, x) - \rho(l|\mathcal{G})|(l, l) \leq 1\}\}$.

$$\rho(l|X^-(\tau,t_0,X^0,\mu)) \tag{16}$$

$$= \rho\left(l\,\middle|\left(X^0+\mathcal{B}_\mu(0)+\int_{t_0}^{\tau}B(s)\mathcal{P}(s)ds\right)\dot{-}\int_{t_0}^{\tau}(-C(s)\mathcal{Q}(s))ds\right),$$

which, interpreted as integrals of multivalued functions, again results in (14).

THEOREM 2.1. *The set* $X^-(\tau,t_0,X^0,\mu)$ *is given by formula* (14) *and its support function* $\rho(l|X^-(\tau,t_0,X^0,\mu))$ *by* (16).

It is clear that if the difference

$$\int_{t_0}^{\tau}B(t)\mathcal{P}(s)ds\dot{-}\int_{t_0}^{\tau}C(s)(-\mathcal{Q}(s))ds\neq\emptyset,$$

then $X^-(\tau,t_0,x^0,0)\neq\emptyset$.

Note that function $V^-(\tau,x,\mu)$ may also be defined as the solution to

Problem (I*). Given X^0, find

$$V_*^-(\tau,x,\mu)=\max_v\min_u\min_{x(t_0)}\{d(x(\tau),\mathcal{B}_\mu(x))|x(t_0)\in X^0,u(\cdot)\in U_{\mathcal{O}},v(\cdot)\in V_{\mathcal{O}}\}.$$

Direct calculations then produce the formula

$$\{x:V_*^-(\tau,x,\mu)\leq 0\}=X^-(\tau,t_0,X^0,\mu), \tag{17}$$

which gives the same result as Problem (I).

Similarly, we may calculate

$$V^+(\tau,x,\mu)=\min_u\max_v\min_{x(\tau)}\{d(x(t_0),X^0)|x(\tau)\in\mathcal{B}_\mu(x),u(\cdot)\in U_{\mathcal{O}},v(\cdot)\in V_{\mathcal{O}}\}.$$

Taking into account the minmax theorem of [7] and the fact that

$$\max_l g(l)=\max_l(conc\ g)(l),\ (l,l)\leq 1,$$

we come to

$$V_+(\tau,x,\mu)=\max_l\left\{(l,x)-\int_{t_0}^{\tau}\rho(l|B(s)\mathcal{P}(s))ds+(conc(-h))(l)|(l,l)\leq 1\right\}, \tag{18}$$

$$h(l)=\rho(l|X^0)+\mu(l,l)^{\frac{1}{2}}-\int_{t_0}^{\tau}\rho(-l|C(s)\mathcal{Q}(s))ds.$$

Here $(conc\ h)(l)$ is the *closed concave hull* of $h(l)$. Note that

$$(conc\ h)(l)=-(conv(-h))(l),$$

where $(conv\ h)(l)=h^{**}(l)$ is the *closed convex hull* and also the Fenchel second conjugate $h^{**}(l)$ of $h(l)$ (see [25], [13] for the definitions).

Therefore,

$$V^+(\tau,x,\mu)=d(x,X^+(\tau,t_0,X^0,\mu)), \tag{19}$$

where

$$X^+(\tau, t_0, X^0, \mu) \tag{20}$$

$$= \left(\left(X^0 + \mathcal{B}_\mu(0) \right) \dot{-} \int_{t_0}^{\tau} (-C(s)) \mathcal{Q}(s) ds \right) + \int_{t_o}^{\tau} B(s) \mathcal{P}(s) ds.$$

It then follows that

$$X^+(\tau, t_0, X^0, \mu) = \{x : V^+(\tau, x, \mu) \leq 0\}. \tag{21}$$

Similarly, (18) implies that $V^+(\tau, x, \mu) \leq 0$ iff

$$(l, x) \leq (l, x^0) + \int_{t_0}^{\tau} (\rho(l|B(s)\mathcal{P}(s)) ds - (conv\ h)(l)\ \forall l \in \mathbb{R}^n$$

so that the support function

$$\rho(l|X^+(\tau, t_0, X^0, \mu)) \tag{22}$$

$$= \rho\left(l \left| \int_{t_0}^{\tau} B(s)\mathcal{P}(s) ds \right. \right) + \rho\left(l \left| \left(X^0 + \mathcal{B}_\mu(0) \right) \dot{-} \int_{t_0}^{\tau} (-C(s)) \mathcal{Q}(s) ds \right. \right).$$

THEOREM 2.2. *The set $X^+(\tau, t_0, X^0, \mu)$ is given by* (20) *and its support function $\rho(l|X^+(\tau, t_0, X^0, \mu))$ by* (22).

It can be seen from (22) that $X^+(\tau, t_0, x^0, 0)$ may be empty. At the same time, in order that $X^+(\tau, t_0, x^0, \mu) \neq \emptyset$, it is sufficient that

$$\mathcal{B}_\mu(0) \dot{-} \int_{t_0}^{\tau} (-C(s)) \mathcal{Q}(s) ds \neq \emptyset,$$

which holds for $\mu > 0$ sufficiently large.

It is worth mentioning that a minmax OLRS may be also be specified through an alternative definition.

DEFINITION 2.1. *A μ-OLRS of the* minmax type *(from set X^0, at time $\tau \geq t_0$) is the union*

$$X^+(\tau, t_0, X^0, \mu) = \cup \{X^+(\tau, t_0, x^0, \mu) | x^0 \in X^0\}, \tag{23}$$

where

$$X^+(\tau, t_0, x^0, \mu) = \{x : X(\tau, t_0, x^0 | u(\cdot), \mathcal{Q}(\cdot)) \subseteq \mathcal{B}_\mu(x)\}$$

for some $u(\cdot) \in U_\mathcal{P}$ with $\mu \geq 0$ given and each set $X^+(\tau, t_0, x^0, \mu) \neq \emptyset$.

This leads to the following problem.

Problem (II*). Given set X^0, and vector $x \in \mathbb{R}^n$, find

$$V_*^+(\tau, x, \mu) = \min_u \max_v \min_{x(\tau)} d(x(\tau), \mathcal{B}_\mu(x)), \quad \tau \geq t_0,$$

under conditions $x(t_0) \in X^0, u(\cdot) \in U_\mathcal{O},\ v(\cdot) \in V_\mathcal{O}$.

Direct calculations here lead to the formula

$$X^+(\tau, t_0, x^0, \mu) = \{x : V_*^+(\tau, x, \mu) \leq 0\},$$

which is the same result as that of Problem (II).

The equivalence of Problems (II) and (II*) means that Definitions 1.2 and 2.1 both lead to the same set $X^+(\tau, t_0, x^0, \mu)$. As we shall see, this is not so for the problem of reachability with corrections. A similar observation holds for Problems (I) and (I*).

Remark 2.1. For the case in which $X^0 = \{x^0\}$ is a singleton, one should recognize the following. The OLRS of the *maxmin* type is the set of points reachable at time τ from a given point x^0 for any disturbance $v(\cdot) \in V_{\mathcal{O}}$, provided function $v(t)$, $t_0 \le t \le \tau$, is communicated to the controller *in advance, before the selection of control* $u(t)$. As mentioned above, the control $u(\cdot)$ is then selected through an *anticipative control* procedure.

On the other hand, for the construction of the μ-reach set of the *minmax* type, there is *no information provided in advance on the realization of* $v(\cdot)$, which becomes known only *after* the selection of u. Indeed, given point $x(t_0) = x^0$, one has to select the control $u(t)$ for the whole time interval $t_0 \le t \le \tau$, whatever the unknown $v(t)$ over the same interval is. The control $u(\cdot)$ is then selected through a *nonanticipative control* procedure. Such a definition allows us to specify an OLRS as consisting of points x, each of which is complemented by a neighborhood $\mathcal{B}_\mu(x)$, so that

$$X(\tau, t_0, x^0 | u(\cdot), \mathcal{Q}(\cdot)) \subseteq \mathcal{B}_\mu(x)$$

for a certain control $u(\cdot) \in U_{\mathcal{O}}$. This requires $\mu > 0$ to be sufficiently large.

As a first step toward reachability under feedback, we consider *piecewise open-loop controls* with the possibility of corrections at fixed instants of time.

3. Piecewise open-loop controls: Reachability with corrections. Here we define and calculate reachability sets under a finite number of corrections. This is done either through the solution of problems of sequential maxmin and minmax or through operations on set-valued integrals.

Taking a given instant of time $t^* \in [t_0, t_1] = T$ that divides the interval T in two, namely,

$$T_1 = [t_0, t_0 + \sigma), \;\; T_2 = [t_0 + \sigma, t_1], \; \sigma = t^* - t_0,$$

consider the following *sequential maxmin* problem.

Problem (I_1). Given set X^0, $x \in \mathbb{R}^n$, and numbers $\mu_1 \ge 0, \mu_2 \ge 0$, find

$$V_1^-(t_0 + \sigma, x, \mu_1)$$

$$= \max_v \min_u \min_{x(t_0+\sigma)} \{d(x(t_0), X^0) | x(t_0 + \sigma) \in \mathcal{B}_{\mu_1}(x); u(t) \in \mathcal{P}(t), v(t) \in \mathcal{Q}(t), t \in T_1\},$$

and then find

$$V_1^-(\tau, x, \{\mu_1, \mu_2\}) \tag{24}$$

$$= \max_v \min_u \min_{x(\tau)} \{V_1^-(t_0+\sigma, x(t_0+\sigma), \mu_1) | x(\tau) \in \mathcal{B}_{\mu_2}(x), u(t) \in \mathcal{P}(t), v(t) \in \mathcal{Q}(t), t \in T_2\}.$$

The latter is a problem on finding a sequential maxmin with one "point of correction" $t = t^*$. Using the notation $\mu[1,2] = \{\mu_1, \mu_2\}$, denote

$$X_1^-(\tau, t_0, \mu[1,2]) = \{x : V_1^-(\tau, x, \mu[1,2]) \le 0\}.$$

Let us find $X_1^-(\tau, t_0, \mu[1,2]), V_1^-(\tau, x, \mu[1,2])$, using the technique of convex analysis. According to section 2 (see (11)), we have

$$V_1^-(t_0+\sigma, x(t_0+\sigma), \mu_1)$$

$$= \max\Bigg\{(l, x(t_0+\sigma)) - \mu_1(l,l)^{\frac{1}{2}} - \rho(l|X^0)$$

$$-\int_{t_0}^{t_0+\sigma}(\rho(l|B(s)\mathcal{P}(s))) - \rho(-l|C(s)\mathcal{Q}(s))ds|(l,l) \le 1\Bigg\}$$

$$= d(x(t_0+\sigma), Z_1^-(t_0+\sigma, t_0, \mu_1)),$$

where

$$Z_1^-(t_0+\sigma, t_0, \mu_1) = \left(X^0 + \mathcal{B}_{\mu_1}(0) + \int_{t_0}^{t_0+\sigma} B(s)\mathcal{P}(s)ds\right) \dot{-} \int_{t_0}^{t_0+\sigma}(-C(s))\mathcal{Q}(s))ds.$$

Substituting this in (24), we have

$$V_1^-(\tau, x, \mu[1,2])$$

$$= \max_v \min_u \min_{x(\tau)} \max_l \Bigg\{(l, x(\tau)) - \int_{t_0+\sigma}^{\tau}(l, u(s)+v(s))ds - \rho(l|Z_1^-(t_0+\sigma, t_0, \mu_1))|$$

$$|(l,l) \le 1, x(\tau) \in \mathcal{B}_{\mu_2}(x), u(t) \in \mathcal{P}(t), v(t) \in \mathcal{Q}(t), t \in T_2\Bigg\}.$$

Continuing the calculation, we come to

$$V_1^-(\tau, x, \mu[1,2]) \tag{25}$$

$$= \max_l \Bigg\{(l,x) - \mu_2(l,l)^{\frac{1}{2}} - \int_{t_0+\sigma}^{\tau}(\rho(l|B(s)\mathcal{P}(s)))ds + \int_{t_0+\sigma}^{\tau}(\rho(-l|C(s)\mathcal{Q}(s)))ds$$

$$-(conv\ h_1)(l)\Bigg\},$$

where

$$h_1(l) = \rho(l|X^0) + \mu_1(l,l)^{\frac{1}{2}} + \int_{t_0}^{t_0+\sigma}(\rho(l|B(s)\mathcal{P}(s)) - \rho(-l|C(s)\mathcal{Q}(s)))ds.$$

So $(conv\ h_1)(l)$ is the support function of the set

$$\left(X^0 + \mathcal{B}_{\mu_1}(0) + \int_{t_0}^{t_0+\sigma} B(s)\mathcal{P}(s)ds\right) \dot{-} \int_{t_0}^{t_0+\sigma}(-C(s)\mathcal{Q}(s))ds.$$

Together with (25), this allows us, as in section 2, to express $V_1^-(\tau, x, \mu[1,2])$ as

$$V_1^-(\tau, x, \mu[1,2]) = d(x, X_1^-(\tau, t_0, x^0, \mu[1,2])),$$

where

$$X_1^-(\tau, t_0, x^0, \mu[1,2]) \tag{26}$$

$$= \left(\left(\left(X^0 + \mathcal{B}_{\mu_1}(0) + \int_{t_0}^{t_0+\sigma} B(s)\mathcal{P}(s)ds\right) \dot{-} \int_{t_0}^{t_0+\sigma} (-C(s))\mathcal{Q}(s)ds\right)\right.$$

$$\left. + \mathcal{B}_{\mu_2}(0) + \int_{t_0+\sigma}^{\tau} B(s)\mathcal{P}(s)ds\right) \dot{-} \int_{t_0+\sigma}^{\tau} (-C(s))\mathcal{Q}(s)ds.$$

Formula (26) shows that $X_1^-(\tau, t_0, x^0, \mu[1,2])$ (defined as the level set of V_1^-) is also the reach set with one correction. In particular, $X_1^-(\tau, t_0, x^0, 0)$ consists of all states x that may be reached *for any* function $v(\cdot) \in V_{\mathcal{P}}$, whose values are communicated in two stages, through two consecutive selections of some open-loop control $u(t)$ according to the following scheme.

Stage 1. *Given at time* t_0 are the initial state x^0 and the function $v(t)$ for $t \in T_1$; select at time t_0 the control $u(t)$ for $t \in T_1$.

Then at the *instant of correction* $t_* = t_0 + \sigma$ comes additional information for stage 2.

Stage 2. *Given at time* t^* are the state $x(t^*)$ and the function $v(t)$ for $t \in T_2$; select at time $t = t^*$ the control $u(t)$ for $t \in T_2$.

This proves Theorem 3.1.

THEOREM 3.1. *The set*

$$X_1^-(\tau, t_0, x^0, \mu[1,2]) = \{x : V_1^-(\tau, x, \mu[1,2]) \le 0\}$$

is the maxmin OLRS with one correction at instant $t_0 + \sigma$ *and is given by formula* (26).

We refer to $X_1^-(\tau, t_0, x^0, \mu[1,2])$ as *the maxmin OLRS with one correction* at instant $\tau_1 = t_0 + \sigma$.

The two-stage scheme may be further propagated to the class of piecewise open-loop controls with k corrections. Taking the interval $T = [t_0, \tau]$, introduce a partition

$$\Sigma_k = \{t_0 = \tau_0, \tau_1, \dots, \tau_k, \tau = \tau_{k+1}\},\ \tau_i - \tau_{i-1} = \sigma_i \ge 0, \quad i = 1, \dots, k+1,$$

so that the interval T is now divided into $k+1$ parts

$$T_1 = [t_0, \tau_1),\ T_2 = [\tau_1, \tau_2), \dots, T_{k+1} = [t_1 - \tau_k, t_1],$$

where

$$\tau_i = t_0 + \sum_{j=1}^{i} \sigma_j, \quad i = 1, \dots, k,$$

are *the points of correction.*

Consider also a nondecreasing continuous function $\mu(t) \geq 0,\ \mu(t_0) = 0$ denoting

$$\mu_1 = \mu(\tau_1) - \mu(\tau_0),\ \mu_i = \mu(\tau_i) - \mu(\tau_{i-1}),\ \ i = 1, \ldots, k+1;\ \ \mu = \mu(\tau_{k+1}) - \mu(\tau_0)$$

and also

$$\mu[1, i] = \{\mu_1, \ldots, \mu_i\}, \quad k \geq i > 1.$$

Problem (I_k). Solve the following consecutive optimization problems.
Find

$$V_k^-(\tau_1, x, \mu_1)$$

$$= \max_v \min_u \min_{x(\tau_1)} \{d(x(t_0), X^0) | x(\tau_1) \in \mathcal{B}_{\mu_1}(x), u(t) \in \mathcal{P}(t), v(t) \in \mathcal{Q}(t), t \in T_1\};$$

then find

$$V_k^-(\tau_2, x, \mu[1, 2])$$

$$= \max_v \min_u \min_{x(\tau_2)} \{V_k^-(\tau_1, x(t_0 + \sigma_1), \mu_1) | x(\tau_2) \in \mathcal{B}_{\mu_2}(x), u(t) \in \mathcal{P}(t), v(t) \in \mathcal{Q}(t), t \in T_2\};$$

then, consecutively, for $i = 3, \ldots, k$, find

$$V_k^-(\tau_i, x, \mu[1, i])$$

$$= \max_v \min_u \min_{x(\tau_i)} \{V_k^-(\tau_{i-1}, x(\tau_{i-1}), \mu[1, i-1]) | x(\tau_i) \in \mathcal{B}_{\mu_i}(x), u(t) \in \mathcal{P}(t),$$

$$v(t) \in \mathcal{Q}(t), t \in T_i\},$$

and, finally,

$$V_k^-(\tau, x, \mu[1, k+1])$$

$$= \max_v \min_u \min_{x(\tau)} \{V_k^-(\tau_k, x(\tau_k), \mu[1, \ldots, k]) | x(\tau) \in \mathcal{B}_{\mu_{k+1}}(x), u(t) \in \mathcal{P}(t),$$

$$v(t) \in \mathcal{Q}(t), t \in T_{k+1}\}.$$

Direct calculation gives

$$V_k^-(\tau_1, x, \mu_1) = d(x, X_k(\tau_1, t_0, X^0, \mu_1)), \tag{27}$$

with

$$X_k(\tau_1, t_0, X^0, \mu_1) = \left(X^0 + \mathcal{B}_{\mu_1}(0) + \int_{t_0}^{\tau_1} B(s)\mathcal{P}(s)ds \right) \dot{-} \int_{t_0}^{\tau_1} (-C(s))\mathcal{Q}(s)ds;$$

then

$$V_k^-(\tau_2, x, \mu[1, 2]) = d(x, X_k(\tau_2, t_0, X^0, \mu[1, 2])) = d(x, X_k^-(\tau_2, \tau_1, X_k^-(\tau_1, t_0, X^0, \mu_1), \mu_2)),$$

with

$$X_k(\tau_2, t_0, X^0, \mu[1,2])$$

$$= \left(\left(\left(X^0 + \mathcal{B}_{\mu_1}(0) + \int_{t_0}^{\tau_1} B(s)\mathcal{P}(s)ds\right) \dot{-} \int_{t_0}^{\tau_1} (-C(s))\mathcal{Q}(s)ds\right) + \mathcal{B}_{\mu_2}(0)\right.$$

$$\left. + \int_{\tau_1}^{\tau_2} B(s)\mathcal{P}(s)ds\right) \dot{-} \int_{\tau_1}^{\tau_2} (-C(s))\mathcal{Q}(s)ds;$$

then, consecutively,

$$\begin{aligned} V_k^-(\tau_i, x, \mu[1,i]) &= d(x, X_k^-(\tau_i, t_0, X^0, \mu[1,i])) \\ &= d(x, X_k^-(\tau_i, \tau_{i-1}, X_k^-(\tau_{i-1}, t_0, X^0, \mu[1,i-1]), \mu_i)), \end{aligned}$$

with

$$X_k^-(\tau_i, t_0, X^0, \mu[1,i]) = \left(\cdots\left(X^0 + \mathcal{B}_{\mu_1}(0) + \int_{t_0}^{\tau_1} B(s)\mathcal{P}(s)ds\right) \dot{-} \int_{t_0}^{\tau_1} (-C(s))\mathcal{Q}(s)ds\right)$$

$$+ \cdots + \mathcal{B}_{\mu_i}(0) + \int_{\tau_{i-1}}^{\tau_i} B(s)\mathcal{P}(s)ds\Big) \dot{-} \int_{\tau_{i-1}}^{\tau_i} (-C(s))\mathcal{Q}(s)ds;$$

and, finally,

$$V_k^-(\tau, x, \mu[1,k+1]) = d(x, X_k^-(\tau, t_0, X^0, \mu[1,k+1])), \tag{28}$$

where

$$X_k^-(\tau, t_0, x^0, \mu[1,k+1]) = \left(\cdots\left(X^0 + \mathcal{B}_{\mu_1}(0) + \int_{t_0}^{\tau_1} B(s)\mathcal{P}(s)ds\right)\right. \tag{29}$$

$$\left. \dot{-} \int_{t_0}^{\tau_1} (-C(s))\mathcal{Q}(s))ds\right) + \cdots$$

$$+\mathcal{B}_{\mu_i}(0) + \int_{\tau_i - 1}^{\tau_i} B(s)\mathcal{P}(s)ds\Big) \dot{-} \int_{\tau_i - 1}^{\tau_i} (-C(s))\mathcal{Q}(s)ds\Big) + \cdots$$

$$+\mathcal{B}_{\mu_{k+1}}(0) + \int_{\tau_k}^{\tau} B(s)\mathcal{P}(s)ds\Big) \dot{-} \int_{\tau_k}^{\tau} (-C(s))\mathcal{Q}(s)ds.$$

We refer to $X_k^-(\tau, t_0, \mu[1,k+1])$ as the *maxmin OLRS with k corrections* at points $\tau_i, i = [1 \dots k]$.

THEOREM 3.2. *The set*

$$X_k^-(\tau, t_0, X^0, \mu[1,k+1]) = \{x : V_k^-(\tau, t_0, x, \mu[1,k+1]) \le 0\} \tag{30}$$

is given by formula (29).

We denote

$$V_0^-(\tau, x, \mu_1) = V^-(\tau, x, \mu_1)$$

and also introduce additional notation for the functions $V_i^-(\tau, x, \mu[1, i+1])$. Denote

$$V_i^-(\tau, x, \mu[1, i+1]) = V_i^-(\tau, x, \mu[1, i+1])|V_i^-(t_0, \cdot, 0),$$

emphasizing the dependence of $V_i^-(\tau, x, \mu[1, i+1])$ on the initial condition $V_i^-(t_0, \cdot, 0)$.

We further assume that $V(t_0, x, 0) = d(x, X^0)$ and also take $d(x, X^0) = V_i^-(t_0, x, 0)$ for all i.

Note that the number of nodes τ_j in any partition Σ_k is $k+2$ as $j = 0, \ldots, k+1$. The partition applied to a function V_k is precisely Σ_k. Consequently, the increment

$$\mu = \mu(\tau) - \mu(t_0) = \sum_{j=1}^{k+1} \mu_j$$

is presented as a sum of $k+1$ increments $\mu_j \geq 0$ once it is applied to a function V_k with index k.

A sequence of partitions Σ_k is *monotone* in k if for every $k_1 < k_2$ partition Σ_{k_2} contains all the nodes τ_j of partition Σ_{k_1}.

THEOREM 3.3. *Given are a monotone sequence of partitions Σ_k, $k = 1, 2, \ldots, N, \ldots$ and a continuous nondecreasing function $\mu(t) \geq 0, \mu(t_0) = 0$ that generates for any partition Σ_k a sequence of numbers $\mu_j = \mu(\tau_j) - \mu(\tau_{j-1}), j = 1, \ldots, k+1$.*

Given also are a sequence of value functions $V_k^-(\tau_i, t_0, \mu[1, i])$, each of which is formed by the partition Σ_k, and a sequence $\mu_j, j = 1, \ldots, k+1$ (k is the index of V_k^-).

Then the following relations are true.

(i) *For any fixed τ, x, one has*

$$V_0^-(\tau, x, \mu_1) \leq \cdots \leq V_i^-(\tau, x, \mu[1, i+1]) \leq V_{i+1}^-(\tau, x, \mu[1, i+2]) \leq \cdots \tag{31}$$

$$\cdots \leq V_k^-(\tau, x, \mu[1, \cdots, k+1]).$$

(ii) *For any fixed τ, x, and index $i \in [1, k]$, one has*

$$V_i^-(\tau, x, \mu[1, i+1]) \leq V_i^-(\tau, x, \mu^*[1, i+1]), \tag{32}$$

provided $\mu_j \leq \mu_j^*, j = 1, \ldots, i+1$.

(iii) *The following inclusions are true for $i \in [1, k]$:*

$$X_{i-1}^-(\tau, t_0, X^0, \mu[1, i]) \supseteq X_i^-(\tau, t_0, X^0, \mu[1, i+1]), \tag{33}$$

where the sets X_i^- are defined by (30).

The proofs are based on the following properties of the geometrical (Minkowski) sums and differences of sets $\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3$:

$$(\mathcal{P}_1 + \mathcal{P}_2) \dot{-} \mathcal{P}_3 \supseteq \mathcal{P}_1 + (\mathcal{P}_2 \dot{-} \mathcal{P}_3), \ \mathcal{P}_1 \dot{-} (\mathcal{P}_2 + \mathcal{P}_3) = (\mathcal{P}_1 \dot{-} \mathcal{P}_2) \dot{-} \mathcal{P}_3,$$

and the fact that, in general, a maxmin does not exceed a minmax. Direct calculations indicate that the following superpositions will also be true.

LEMMA 3.1. *The functions* V_k^- *satisfy the property*

$$V_k^-(\tau_i, x, \mu[1,i]\ |V_k^-(t_0,\cdot,0)) = V_k^-(\tau_i, x, \mu[j+1,i]\ |V_k^-(\tau_j,\cdot,\mu[1,j]\ |V_k^-(t_0,\cdot,0))), \tag{34}$$

provided $k+1 \geq i \geq j \geq 1$.

This follows from Theorem 3.2 and the definitions of the respective functions V_i^-.

Remark 3.1. Formula (34) reflects a *semigroup* property but *only for the selected points of correction* $\tau_i,\ i = 1,\dots,k$.

The reasoning above indicates, for example, that $X_k^-(\tau, t_0, x^0, 0)$ is the set of states that may be reached *for any* function $v(\cdot) \in V_{\mathcal{O}}$, whose values are communicated in advance in k stages, through $k+1$ consecutive selections of some open-loop control $u(t)$ according to the following scheme.

Stage 1. *Given at time* t_0 are the initial state x^0 and the function $v(t)$ for $t \in T_1$; select at time t_0 control $u(t)$, for $t \in T_1$.

Then at *instant of correction* τ_j comes additional information for stage $(j+1)$.

Stage j $(j = 2,\dots,k)$. *Given at time* τ_j are the state $x(\tau_j)$ and the function $v(t)$ for $t \in T_{j+1}$; select at time τ the control $u(t)$ for $t \in T_{j+1}$.

Remark 3.2. There is a case when all the functions $V_k^-(\tau, x, 0)$ taken for all the integers $k \geq 0$ coincide. This is when system (10) satisfies the so-called *matching conditions*:

$$B(t)\mathcal{P}(t) \equiv \alpha(t)C(t)\mathcal{Q}(t), \quad \alpha(t) \in [0,1), \quad t \in [t_0, \tau].$$

We now pass to the problem of *sequential minmax*, with one correction at instant $t_0 + \sigma = t^*$, using the notation for Problem (I_1).

Problem (II_1). Given set X^0, vector $x \in \mathbb{R}^n$, and numbers $\mu_1, \mu_2 \geq 0$, find

$$V_1^+(t_0+\sigma, x, \mu_1)$$

$$= \min_u \max_v \min_{x(t_0+\sigma)} \{d(x(t_0), X^0) | x(t_0+\sigma) \in \mathcal{B}_{\mu_1}(x); u(t) \in \mathcal{P}(t), v(t) \in \mathcal{Q}(t), t \in T_1\};$$

then find

$$V_1^+(\tau, x, \mu[1,2]) \tag{35}$$

$$= \min_u \max_v \min_{x(\tau)} \{V_1^+(t_0+\sigma, x(t_0+\sigma), \mu_1) | x(\tau) \in \mathcal{B}_{\mu_2}(x), u(t) \in \mathcal{P}(t), v(t) \in \mathcal{Q}(t), t \in T_2\}.$$

The latter is a problem of finding a sequential minmax with one point of correction $t = t^*$.

Denoting

$$X_1^+(\tau, t_0, X^0, \mu[1,2]) = \{x : V_1^+(\tau, x, \mu[1,2]) \leq 0\},$$

let us find $X_1^+(\tau, t_0, X^0, \mu[1,2]), V_1^+(\tau, x, \mu[1,2])$ using the techniques of convex analysis (as above, with obvious changes).

This gives

$$V_1^+(t_0+\sigma, x, 0) = d(x, Z_1^+(t+\sigma, t_0, \mu_1)),$$

where

$$Z_1^+(t+\sigma,t_0,0)=\left((X^0+\mathcal{B}_{\mu_1}(0))\dot{-}\int_{t_0}^{t_0+\sigma}(-C(s))\mathcal{Q}(s)ds\right)+\int_{t_0}^{t_0+\sigma}B(s)\mathcal{P}(s)ds.$$

Continuing the calculations, we have

$$V_1^+(\tau,x,\mu)$$

$$=\min_u\max_v\min_{x(\tau)}\{d(x(t+\sigma),Z_1^+(t+\sigma,t_0,0))|x(\tau)\in\mathcal{B}_{\mu_2}(x),u(t)\in\mathcal{P}(t),v(t)\in\mathcal{Q}(t),t\in T_2\}$$

$$=d(x,X_1^+(\tau,t_0,X^0,\mu[1,2])),$$

where

$$X_1^+(\tau,t_0,X^0,\mu[1,2]) \tag{36}$$

$$=\left(\left(Z_1^+(t+\sigma,t_0,0)+\mathcal{B}_{\mu_2}(0)\right)\dot{-}\int_{t_0+\sigma}^{\tau}(-C(s))Q(s)ds\right)+\int_{t_0+\sigma}^{\tau}B(s)\mathcal{P}(s)ds.$$

This proves Theorem 3.4.

THEOREM 3.4. *The set*

$$X_1^+(\tau,t_0,X^0,\mu[1,2])=\{x:V_1^+(\tau,x,\mu[1,2])\le 0\}$$

is the minmax OLRS with one correction at instant $t=t_0+\sigma$*, given by formula* (36).

Here the problem is again solved in two stages, according to the following scheme.

Stage 1. *Given at time* t_0 are set X^0 and $x\in\mathbb{R}^n$. Select control $u(t)$ (*one and the same for all* v) and for each $v(t),t\in T_1$, assign a vector $x(t_0)\in X^0$ that jointly with u,v produces $x(\tau)\in\mathcal{B}_{\mu_1}(0)$.

Then at instant of correction $t^*=t_0+\sigma$ comes additional information for Stage 2.

Stage 2. Given at time t^* are $x(t^*)$ and vector $x\in\mathbb{R}^n$. Select control $u(t),t\in T_2$ (one and the same for all v), and for each $v(t),t\in T_2$, assign a vector $x(t+\sigma)\in Z^+(t_0+\sigma,t_0,\mu_1)$ that jointly with u,v steers the system to state $x(\tau)\in\mathcal{B}_{\mu_2}(x)$.

We now propagate this *minmax* procedure to a sequential minmax problem in the class of piecewise open-loop controls with k corrections, using the notations of Problem (I_k).

Problem (II_k). Solve the following consecutive optimization problems.

Find

$$V_k^+(\tau_1,x,\mu_1)$$

$$=\min_u\max_v\min_{x(\tau_1)}\{d(x(t_0),X^0)|x(\tau_1)\in\mathcal{B}_{\mu_1}(x);u(t)\in\mathcal{P}(t),v(t)\in\mathcal{Q}(t),t\in T_1\};$$

then, consecutively, for $i=2,\dots,k$, find

$$V_k^-(\tau_i,x,\mu[1,i])$$

$$= \min_u \max_v \min_{x(\tau_{i-1})} \{V_k^+(\tau_{i-1}, x(\tau_{i-1}), \mu[1, i-1]) \,|x(\tau_i) \in \mathcal{B}_{\mu_i}(x), \\ u(t) \in \mathcal{P}(t), v(t) \in \mathcal{Q}(t), t \in T_i\};$$

and, finally,

$$V_k^+(\tau, x, \mu[1, k+1])$$

$$= \max_v \min_u \{V_k^+(\tau_k, x(\tau_k), \mu[1, k+1]) \,|x(\tau) \in \mathcal{B}_{\mu_{k+1}}(x), u(t) \in \mathcal{P}(t), v(t) \in \mathcal{Q}(t), t \in T_{k+1}\}.$$

This time, direct calculation gives

$$V_k^+(\tau, x, \mu) = d(x, X_k^+(\tau, t_0, X^0, \mu)), \tag{37}$$

where

$$X_k^+(\tau, t_0, x^0, \mu) \tag{38}$$

$$= \Bigg(\Bigg(\cdots \Bigg((X^0 + \mathcal{B}_{\mu_1}(0)) \dot{-} \int_{t_0}^{\tau_1} (-C(s)) \mathcal{Q}(s) ds \Bigg) + \int_{t_0}^{\tau_1} B(s) \mathcal{P}(s) ds + \mathcal{B}_{\mu_2}(0) \Bigg) \dot{-} \cdots$$

$$\dot{-} \int_{\tau_{i-1}}^{\tau_i} (-C(s)) \mathcal{Q}(s) ds \Bigg) + \int_{\tau_{i-1}}^{\tau_i} B(s) \mathcal{P}(s) ds + \mathcal{B}_{\mu_{i+1}}(0) \Bigg) \dot{-} \cdots$$

$$\dot{-} \int_{\tau_k}^{\tau} (-C(s)) \mathcal{Q}(s) ds \Bigg) + \int_{\tau_k}^{\tau} B(s) \mathcal{P}(s) ds \Bigg).$$

We refer to $X_k^+(\tau, t_0, x^0, \mu)$ as the *maxmin OLRS with k corrections* at points τ_k.

THEOREM 3.5. *The set*

$$X_k^+(\tau, t_0, x^0, \mu[1, k+1]) = \{x : V_k^+(\tau, x, \mu[1, k+1]) \le 0\}$$

is then the minmax OLRS with one correction and is given by formula (38).

Denote

$$V^+(\tau, x, \mu) = V_0^+(\tau, x, \mu);\ V_i^+(\tau, x, \mu[1, i+1]) = V_i^+(\tau, x, \mu[1, i+1] | V_i^+(t_0, \cdot, 0)),$$

assuming $V_0^+(t_0, x, \mu) = d(x, X^0)$ and, further, taking $V_i^+(t_0, x, 0) = d(x, X^0)$ for all i. Under the assumptions and notation of Theorem 3.3, the last results may be summarized in the following theorem.

THEOREM 3.6. (i) *For any fixed values* τ, x *one has*

$$V_0^+(\tau, x, \mu_1) \ge \cdots \ge V_i^+(\tau, x, \mu[1, i+1]) \ge V_{i+1}^+(\tau, x, \mu[1, i+2]) \\ \ge \cdots \ge V_k^+(\tau, x, \mu[1, k+1]). \tag{39}$$

(ii) *For any fixed* τ, x, *and index* $i \in [1, k]$, *one has*

$$V_i^+(\tau, x, \mu_1) \le V_i^+(\tau, x, \mu_2), \tag{40}$$

provided $\mu_1 \le \mu_2$.

(iii) *The following inclusions are true for* $i \in [1, \dots, k], \mu \geq 0$*:*

$$X^+_{i-1}(\tau, t_0, X^0, \mu[i-1]) \subseteq X^+_i(\tau, t_0, X^0, \mu[1, i]). \tag{41}$$

(iv) *The following superpositions will also be true:*

$$V^+_k(\tau_i, x, \mu[1, i]|V^+_k(t_0, \cdot, 0)) = V^+_k(\tau_i, x, \mu[j+1, i]|V^+_k(\tau_j, \cdot, \mu[1, j]|V^+_k(t_0, \cdot, 0))),$$

provided $k + 1 \geq i \geq j \geq 1$.

In this section, we have considered problems with a finite number of possible corrections and additional information coming at fixed instants of time, having presented a hierarchy of piecewise OLRSs of the anticipative (maxmin) or of the nonanticipative type. These were presented as level sets for value functions which are superpositions of "one-stage" value functions calculated in section 2. A semigroup-type property (34) for these value functions was indicated which is true only for the points of correction Remark 3.1. In the continuous case, however, we shall need this property to be true *for any points*. Then it would be possible to formulate the principle of optimality under uncertainty for our class of problems.

We shall therefore investigate some limit transitions with a number of corrections tending to infinity. This will allow a further possibility of continuous corrections of the control under unknown disturbances.

4. The alternated integrals and the value functions. We observed above that the OLRSs of both types (maxmin and minmax) are described as the level sets of some value functions, namely,[6]

$$\begin{aligned} X^-_k(\tau, t_0, X^0, \mu(\cdot)) &= \{x : V^-_k(\tau, x, \mu(\cdot)) \leq 0\},\ X^+_k(\tau, t_0, X^0, \mu(\cdot)) \\ &= \{x : V^+_k(\tau, x, \mu(\cdot)) \leq 0\}. \end{aligned}$$

We now propagate this approach, based on using value functions, to systems with continuous measurements of the state to allow continuous corrections of the control.

First, note that inequality

$$V^-_i(\tau, x, \mu(\cdot)) \leq V^-_i(\tau, x, 0) - \mu$$

is always true with equality attained, for example, under the following assumption.

Assumption 4.1. There exists a scalar function $\epsilon(t) > 0$ such that

$$B(t)\mathcal{P}(t) \dot{-} (C(t)\mathcal{Q}(t) + \epsilon(t)\mathcal{B}_1(0)) \neq \emptyset$$

for all $t \in [t_0, \tau]$.

In order to simplify further explanations, we shall further deal in this section with the case when $\mu = 0$, omitting the last symbol 0 in the notation for V^-, V^+.[7]

Now note that Lemmas 3.1 and 3.2 indicate that each of the functions

$$V^-_k(\tau, x, 0\,|V^-_k(t_0, \cdot, 0)),\ V^+_k(\tau, x, 0\,|V^+_k(t_0, \cdot, 0))$$

[6]Here, without abuse of notation for V^-_k, X^-_k, V^+, X^+_k, we shall use the symbol $\mu(\cdot)$ rather than the earlier $\mu[1, k+1]$, emphasizing the function $\mu(t)$, $\mu(\tau) - \mu(t_0) = \mu$ used in the respective constructions.

[7]The case when $\mu(\cdot) \neq 0$ would add to the length of the expressions but not to the essence of the scheme. This case could be treated similarly, with obvious complements.

may be determined through a sequential procedure

$$V_k^-(\tau, x\,|V_k^-(t_0,\cdot)) = V_k^-(\tau, x\,|V_k^-(\tau-\sigma_k,\cdot\,|\,\ldots\,|V_k^-(t_0+\sigma_1,\cdot\,|V_k^-(t_0,\cdot)\ldots) \tag{42}$$

for V_k^- and a similar one for V_k^+. How could one express this procedure in terms of set-valued representations?

For a given partition Σ_k, we have $(j \le i)$

$$\left\{x : V_k^-\left(t_0+\sum_{l=1}^{i}\sigma_l, x\,|V_k^-\left(t_0+\sum_{l=1}^{j}\sigma_l,\cdot\right)\right)\le 0\right\} =$$

$$= X_k^-\left(t_0+\sum_{l=1}^{i}\sigma_l, t_0+\sum_{l=1}^{j}\sigma_l, \mathcal{M}_j^-\right),$$

where $\mathcal{M}_j^- = \{x : V_k^-(t_0+\sum_{l=1}^{j}\sigma_l,\cdot)\le 0\}$. Then, in view of the previous relations (see (27)–(29)), we may formulate a set-valued analogy of Lemma 3.1.

LEMMA 4.1. *The following relations are true:*

$$X_k^-(\tau, t_0, X^0) \tag{43}$$

$$= X_k^-(\tau, \tau-\sigma_{k+1}, X_k^-(\tau-\sigma_{k+1}, \tau-\sigma_{k+1}-\sigma_k, \ldots X_k^-(t_0+\sigma_2, t_0+\sigma_1, X_k^-(t_0+\sigma_1, t_0, X^0)\ldots).$$

In terms of set-valued integrals, (43) *is precisely the equivalent of* (29).

Moreover,

$$V_k^-\left(t_0+\sum_{l=1}^{i}\sigma_l, x|V_k^-\left(t_0+\sum_{l=1}^{j}\sigma_l,\cdot\right)\right) \tag{44}$$

$$= \max_v\min_u\ldots\max_v\min_u\left\{d\left(x\left(t_0+\sum_{l=1}^{j}\sigma_l\right), \mathcal{M}_j^-\right)|x\left(t_0+\sum_{l=1}^{i}\sigma_l\right) = x;\right.$$

$$\left. u(t)\in\mathcal{P}(t), v(t)\in\mathcal{Q}(t), t\in T_j;\ldots;u(t)\in\mathcal{P}(t), v(t)\in\mathcal{Q}(t), t\in T_i\right\}.$$

Similarly, for the sequential minmax, we have

$$V_k^+(\tau, x|V_k^+(t_0,\cdot)) = V_k^+(\tau, x|V_k^+(\tau-\sigma_{k+1},\cdot|\ldots|V_k^+(t_0+\sigma_1,\cdot|V_k^+(t_0,\cdot)\ldots). \tag{45}$$

Using notation identical to (42) and (43), but with minus changed to plus in the symbols for V_k^-, X_k^-, we have Lemma 4.2.

LEMMA 4.2. *The following relations are true:*

$$X_k^+(\tau, t_0, X^0) = X_k^+(\tau, \tau-\sigma_{k+1}, X_k^+(\tau-\sigma_{k+1}, \tau-\sigma_{k+1}-\sigma_k, \ldots, \tag{46}$$

$$X_k^+(t_0+\sigma_2, t_0+\sigma_1, X_k^+(t_0+\sigma_1, t_0, X^0)\ldots).$$

In terms of set-valued integrals, formula (46) *is precisely the equivalent of* (38), *provided* $\mu(t) \equiv 0$.[8]

Moreover,

$$V_k^+\left(t_0+\sum_{l=1}^{i}\sigma_l, x|V_k^+\left(t_0+\sum_{l=1}^{j}\sigma_l,\cdot\right)\right) \tag{47}$$

$$= \min_u \max_v \ldots \min_u \max_v \left\{ d\left(x\left(t_0+\sum_{l=1}^{j}\sigma_l\right), \mathcal{M}_j^+ \middle| x\left(t_0+\sum_{l=1}^{i}\sigma_l\right) = x;\right.\right.$$

$$\left.\left. u(t)\in\mathcal{P}(t), v(t)\in\mathcal{Q}(t), t\in T_j, ;\ldots; u(t)\in\mathcal{P}(t), v(t)\in\mathcal{Q}(t), t\in T_i\right\},\right.$$

where $\mathcal{M}_j^+ = \{x : V_k^+(t_0+\sum_{l=1}^{j}\sigma_l,\cdot)\le 0\}$.

It is important to emphasize that until now all the relations were derived *for a fixed partition*

$$\Sigma_k = \{t_0=\tau_0,\tau_1,\ldots,\tau_k,\tau=\tau_{k+1}\};\ \tau_i-\tau_{i-1}=\sigma_i;\ i=1,\ldots,k+1.$$

What would happen, however, if k increases to infinity with

$$\max\{\sigma_i\ :\ i=1,\ldots,k+1\}\to 0,\ k\to\infty, \sum_{i=1}^{k+1}\sigma_i = \tau - t_0, \tag{48}$$

and would the result depend on the type of partition?

Our further discussion will require an important *nondegeneracy assumption*.

Assumption 4.2. There exist continuous vector functions $\beta_1(t),\ \beta_2(t)\in\mathbb{R}^n, t\in[t_0,t_1]$, and a number $\epsilon>0$ such that

$$\text{(a)}\quad \beta_1(\tau_j)+\epsilon\mathcal{B}(0)\subseteq X_j^-(\tau_j,t_0,X^0) \tag{49}$$

for all of the sets

$$X_j^-(\tau_j,t_0,X^0)$$

$$= X_j^-(\tau_j,\tau_{j-1},X_j^-(\tau_{j-1},\tau_{j-2},\ldots,X_j^-(\tau_1,t_0,X^0)\ldots)$$

and

$$\text{(b)}\quad \beta_2(\tau_j)+\epsilon\mathcal{B}(0)\subseteq X_j^+(\tau_j,t_0,X^0) \tag{50}$$

for all of the sets

[8] Also note that, under Assumption 4.1, with X^0 single-valued, one may treat the sets $X_k^+(\tau,x,0)$ as the Hausdorff limits $X_k^+(\tau,x,0)=\lim_{\mu\to+0}X_k^+(\tau,x,\mu)$.

$$X_j^+(\tau_j, t_0, X^0) \tag{51}$$

$$= X_j^+(\tau_j, \tau_{j-1}, X_j^+(\tau_{j-1}, \tau_{j-2}, \ldots, X_j^+(\tau_1, t_0, X^0)\ldots)$$

with $j = 1, \ldots, k+1$, whatever the partition Σ_k is.

This last assumption is further taken to be true *without further notice.*[9]

Observing that (29) and (38) have the form of certain set-valued integral sums ("the alternated sums"), we introduce the additional notation

$$X_k^-(\tau, t_0, X^0) = \mathcal{I}^-(\tau, t_0, X^0, \Sigma_k); \;\; X_k^+(\tau, t_0, X^0) = \mathcal{I}^+(\tau, t_0, X^0, \Sigma_k).$$

Let us now proceed with the limit operation. Take a monotone sequence of partitions Σ_k, $k \to \infty$. Due to inclusions (33) and the boundedness of the sequence $X_k^-(\tau, t_0, X^0)$ from below by any of the sets $X_i^+(\tau, t_0, X^0)$, $i \le k$, the sequence $\mathcal{I}^-(\tau, t_0, X^0, \Sigma_k)$ has a set-valued limit. Similarly, the inclusions (40) and the boundedness of the sequence $X_k^+(\tau, t_0, X^0, \Sigma_k)$ from above ensure that it also has a set-valued limit. A more detailed investigation of this scheme along the lines of [23] would indicate that, under Assumption 4.2 (a) and (b), these set-valued limits *do not depend on the type of partition* Σ_k. This leads to Theorem 4.1.

THEOREM 4.1. *There exist Hausdorff limits* $\mathcal{I}^-(\tau, t_0, X^0) = X^-(\tau, t_0, X^0)$, $\mathcal{I}^+(\tau, t_0, X^0) = X^+(\tau, t_0, X^0)$*:*

$$\lim h(\mathcal{I}^-(\tau, t_0, X^0, \Sigma_k), \mathcal{I}^-(\tau, t_0, X^0)) = 0,$$

$$\lim h(\mathcal{I}^+(\tau, t_0, X^0, \Sigma_k), \mathcal{I}^+(\tau, t_0, X^0)) = 0,$$

with

$$\max\{\sigma_i \;:\; i = 1, \ldots, k+1\} \to 0, \; k \to \infty, \sum_{i=1}^{k+1} \sigma_i = \tau - t_0.$$

These limits do not depend on the type of partition Σ_k.

Moreover,

$$\mathcal{I}^-(\tau, t_0, X^0) = \mathcal{I}^+(\tau, t_0, X^0) = \mathcal{I}(\tau, t_0, X^0) \tag{52}$$

so that

$$X_k^-(\tau, t_0, X^0) = X_k^+(\tau, t_0, X^0) = X(\tau, t_0, X^0).$$

We refer to $\mathcal{I}(\tau, t_0, X^0) = X(\tau, t_0, X^0)$ as the *alternated reach set.*[10]

The proofs of the convergence of the alternated integral sums to their Hausdorff limits and of the equalities (52) are not given here. They follow the lines of those given in detail in [15] for problems on sequential maxmin and minmax considered in backward time (see also [23], [9], [14]).

Let us now study the behavior of the function $V_i^-(\tau, x | V_i^-(t_0, \cdot))$ under condition (48). According to (38) and (31), the sequence $V_i^-(\tau, x)$ is increasing in i with $i \to \infty$.

[9] If at some stage this assumption is not fulfilled, it may be applied to sets of type $X_j^-(\tau_j, t_0, X^0, \mu(\cdot)), X_j^+(\tau_j, t_0, X^0, \mu(\cdot))$ with $\mu(\cdot)$ sufficiently large.

[10] A maxmin construction of the indicated type had been introduced in detail in [23], where it was constructed in backward time, becoming known as the *alternated integral of Pontryagin.*

This sequence is pointwise bounded in x by any solution of Problem (II_k) and therefore has a pointwise limit. Due to (29), Theorem 4.1, and the continuity of the distance function $d(x,\mathcal{M})$ in x, $\mathcal{M}\in conv\mathbb{R}^n$, we have, with $k\to\infty$,

$$\lim d(x,\mathcal{I}_k^-)=d(x,\lim\mathcal{I}_k^-)=d(x,\mathcal{I}^-),$$

and therefore we may conclude that

$$V_k^-(\tau,x)\to d(x,\mathcal{I}^-(\tau,t_0,X^0))=\mathcal{V}^-(\tau,x)$$

under condition (48). This yields Theorem 4.2.

THEOREM 4.2. *Under condition* (48)*, there exists a pointwise limit*

$$\lim_{k\to\infty}V_k^-(\tau,x)=\mathcal{V}^-(t,x)=d(x,X^-(\tau,t_0,X^0)),\tag{53}$$

where $X^-(\tau,t_0,X^0)=\mathcal{I}^-(\tau,t_0,X^0)$*. This limit does not depend on the type of partition* Σ_k*.*

The alternated integral is the level set of the function $\mathcal{V}^-(\tau,x)$*,*

$$\mathcal{I}^-(\tau,t_0,X^0)=\{x:\mathcal{V}^-(\tau,x)\le 0\}.$$

Since $\mathcal{V}^-(t,x)$ does not depend on the partition Σ_k and due to the properties of minmax, we also come to the following conclusion.

THEOREM 4.3. *The function* $\mathcal{V}^-(\tau,x)$ *satisfies the semigroup property*

$$\mathcal{V}^-(\tau,x|\mathcal{V}^-(t_0,\cdot))=\mathcal{V}^-(\tau,x|\mathcal{V}^-(t,\cdot|\mathcal{V}^-(t_0,\cdot)))\tag{54}$$

for $t\in[t_0,\tau]$*. The following inequality is true:*

$$\mathcal{V}^-(t,x)\ge\left\{\max_v\min_u\mathcal{V}^-(t-\sigma,x(t-\sigma))|x(t)=x\right\},\quad\sigma>0.\tag{55}$$

Similarly, for the decreasing sequence of functions $V_k^+(\tau,x)$, we have Theorem 4.4.

THEOREM 4.4. (i) *Under condition* (48) *there exists a pointwise limit*

$$\lim_{k\to\infty}V_k^+(\tau,x)=\mathcal{V}^+(t,x)=d(x,X^+(\tau,t_0,X^0)),\tag{56}$$

where $X^+(\tau,t_0,X^0)=\mathcal{I}^+(\tau,t_0,X^0)$*. This limit does not depend on the type of partition* Σ_k*.*

(ii) *The alternated integral is the level set of the function* $\mathcal{V}^+(\tau,x)$*,*

$$\mathcal{I}^+(\tau,t_0,X^0)=\{x:\mathcal{V}^+(\tau,x)\le 0\}.$$

(iii) *The function* $\mathcal{V}^+(\tau,x)$ *satisfies the semigroup property*

$$\mathcal{V}^+(\tau,x|\mathcal{V}^+(t_0,\cdot))=\mathcal{V}^+(\tau,x|\mathcal{V}^+(t,\cdot|\mathcal{V}^+(t_0,\cdot)))\tag{57}$$

for $t\in[t_0,\tau]$*.*

(iv) *The following inequality is true:*

$$\mathcal{V}^+(t,x)\le\left\{\min_u\max_v\mathcal{V}^+(t-\sigma,x(t-\sigma))|x(t)=x\right\},\quad\sigma>0.\tag{58}$$

A consequence of (52) is the basic assertion Theorem 4.5.

THEOREM 4.5. *With the initial condition* $\mathcal{V}^-(t_0,x) = \mathcal{V}^+(t_0,x) = d(x,X^0) = \mathcal{V}(t_0,x)$, *the following equality is true:*

$$\mathcal{V}^+(\tau,x|\mathcal{V}^+(t_0,\cdot)) = \mathcal{V}^-(\tau,x|\mathcal{V}^-(t_0,\cdot)) = \mathcal{V}(\tau,x|\mathcal{V}(t_0,\cdot)) = d(x,X(\tau,t_0,X^0)). \tag{59}$$

The function $\mathcal{V}(\tau,x)$ *satisfies the semigroup property*

$$\mathcal{V}(\tau,x|\mathcal{V}(t_0,\cdot)) = \mathcal{V}(\tau,x|\mathcal{V}(t,\cdot|\mathcal{V}(t_0,\cdot))). \tag{60}$$

The last relation follows from (59), (54), (57).

Thus, under the nondegeneracy Assumption 4.2, the two forward alternated integrals $\mathcal{I}^+,\mathcal{I}^-$ coincide, and so do the value functions $\mathcal{V}^-,\mathcal{V}^+$.

Relations (55), (58), and (59) allow us to construct a partial differential equation for the function $\mathcal{V}(t,x)$—the so-called HJBI *equation*.

We now investigate the existence of the total derivative $d\mathcal{V}(t,x)/dt$ along the trajectories of system (10). Due to (59) and (13), we have

$$\mathcal{V}(t,x) = d(x,X(t,t_0,X^0)) = \max\{(l,x) - \rho(l|X(t,t_0,X^0))|(l,l) \le 1\}.$$

Observing that for $d(x,X(t,t_0,X^0)) > 0$ the maximizer $l^0(t,x)$ of (61) is unique and taking $l^0(t,x) = 0$ if $d(x,X(t,t_0,X^0)) = 0$, we may apply the rules for differentiating a "maximum"-type function [6] to get

$$d\mathcal{V}(t,x)/dt = \partial\mathcal{V}/\partial t + (\partial\mathcal{V}/\partial x,\dot{x}) = (l^0,\dot{x}) - \partial\rho(l^0|X(t,t_0,X^0))/\partial t.$$

Direct calculations indicate that the respective partials exist and are continuous in the domain $\mathcal{D}\cup\mathrm{int}\mathcal{D}_0$, where $\mathcal{D} = \{x : d(x,X(t,t_0,X^0)) > 0\}, \mathcal{D}_0 = \{x : d(x,X(t,t_0,X^0)) = 0\}$ and $\mathrm{int}\mathcal{D}_0$ stands for the interior of the respective set.

To find the value of the total derivative, take inequalities (58) and (55), which may be rewritten as

$$0 \le \min_u \max_v \{V^+(t-\sigma,x(t-\sigma)) - V^+(t,x)|x(t) = x\} \tag{61}$$

and

$$0 \ge \max_v \min_u \{V^-(t-\sigma,x(t-\sigma)) - V^-(t,x)|x(t) = x\}. \tag{62}$$

Dividing both relations by $\sigma > 0$ and passing to the limit with $\sigma \to 0$, we get

$$\max_u \min_v d\mathcal{V}^+(t,x)/dt \le 0,\ \min_v \max_u d\mathcal{V}^-(t,x)/dt \ge 0. \tag{63}$$

Since in Theorem 4.5 we had $\mathcal{V}^+(t,x) = \mathcal{V}^-(t,x) = \mathcal{V}(t,x)$, for the linear system (10) we have

$$\max_u \min_v d\mathcal{V}(t,x)/dt = \min_v \max_u d\mathcal{V}(t,x)/dt, \quad u\in\mathcal{P}(t), \quad v\in\mathcal{Q}(t),$$

which results in the next proposition.

THEOREM 4.6. *In the domain* $\mathcal{D}\cup\mathrm{int}\mathcal{D}_0$, *the value function* $\mathcal{V}(t,x)$ *satisfies the "forward" equation*

$$\partial\mathcal{V}/\partial t + \max_u \min_v (\partial\mathcal{V}/\partial x, B(t)u + C(t)v) = 0 \tag{64}$$

over $u \in \mathcal{P}(t), v \in \mathcal{Q}(t)$ *with boundary condition*

$$\mathcal{V}(t_0, x) = d(x, X^0). \tag{65}$$

Equation (63) may be rewritten as

$$\partial\mathcal{V}/\partial t + \rho(\partial\mathcal{V}/\partial x|B(t)\mathcal{P}(t)) - \rho(\partial\mathcal{V}/\partial x|(-C(t))\mathcal{Q}(t)) = 0. \tag{66}$$

The last theorem indicates that the HJBI equation (63) is satisfied everywhere in the open domain $\mathcal{D} \cup \mathrm{int}\mathcal{D}_0$. However, the continuity of the partials $\partial\mathcal{V}/\partial x, \partial\mathcal{V}/\partial t$ on the boundary of the domains $\mathcal{D}, \mathcal{D}_0$ was not investigated and in fact may not hold. But it is not difficult to check that with boundary condition (65) the function $\mathcal{V}(t, x)$ will be a *minmax solution* to (66) in the sense of [26], which is equivalent to the statement that $\mathcal{V}(t, x)$ is a *viscosity solution* (see [3], [21]) to (66), (67). This particularly follows from the fact that function $\mathcal{V}(t, x)$ is convex in x, being a pointwise limit of convex in x functions $\mathcal{V}_k^-(t, x)$ (see [8]).

Let us note here that the problem under discussion may be treated not only as above but also *within the notion of classical solutions* to (66) and (65). Indeed, although all the results above were proved for the criterion $d(x(t_0, X^0))$ in the respective problems, the following assertion is also true.

Assertion 4.1. *Theorems* 3.1–3.6 *and* 4.1–4.6 *are all true with the criterion* $d(x(t_0, X^0))$ *in the respective problems substituted by* $d^2(x(t_0, X^0))$.

This assertion follows from direct calculations, as in paper [14], with formula (11) substituted by

$$d^2(x, \mathcal{G}) = \max\{(l, x) - \rho(l|\mathcal{G}) - (1/4)(l, l)^{1/2}\}.$$

The respective value function similar to $\mathcal{V}(t, x)$, denoted further as $\mathcal{V}_1(t, x)$, will now be a solution to (66) with boundary condition

$$\mathcal{V}_1(t_0, x) = d^2(x, X^0). \tag{67}$$

Moreover, $\mathcal{V}_1(t, x)$, together with its first partials, turns out to be continuous in $t, x \in \mathcal{D} \cup \mathcal{D}_0$. Thus we come to the following theorem.

Theorem 4.7. *The function* $\mathcal{V}_1(t, x)$*—a classical solution to* (66), (67)*—satisfies the relations*

$$\{x : \mathcal{V}_1(t, x) \le 0\} = X(t, t_0, X^0) = \mathcal{I}(t, t_0, X^0). \tag{68}$$

We have constructed the set $X(t, t_0, X^0)$ as the limit of the OLRS and the level set of function $\mathcal{V}(t, x)$, (or function $\mathcal{V}_1(t, x)$)—the sequential maxmin or minmax of function $d(t, X^0)$ (or function $d^2(t, X^0)$) under restriction $x(t) = x$. It remains to show that $X(t, t_0, X^0)$ is precisely the set of points that may be reached from X^0 with a certain feedback control strategy $\mathcal{U}(t, x)$, whatever the function $v(t)$ is.

Prior to the next section, we wish to note the following. Function $\mathcal{V}(t, x) = \mathcal{V}(t, x|\mathcal{V}(t_0, \cdot))$ may be interpreted as the value function for the following problem.

Problem (IV). Find the *value function*

$$\mathcal{V}(\tau, x) = \min_{\mathcal{U}} \max_{x(\cdot)} \{d(x(t_0), X^0)|\ x(\tau) = x), \mathcal{U} \in U_{\mathcal{C}}, x(\cdot) \in \mathcal{X}_{\mathcal{U}}(\cdot)\},$$

where $\mathcal{U} = \mathcal{U}(t, x) \in U_{\mathcal{C}}$ is a CLC (see section 1) and $\mathcal{X}_{\mathcal{U}}(\cdot)$ is the set of all solutions to the differential inclusion

$$\dot{x} \in B(t)\mathcal{U}(t, x) + C(t)\mathcal{Q}(t), \quad x(\tau) = x, \tag{69}$$

generated by $x(\tau) = x$, $\mathcal{U}(t,x)$, and taken within the interval $t \in [t_0, \tau]$.

Its level set

$$\mathcal{X}_\mu(\tau) = \{x : \mathcal{V}(\tau, x) \leq \mu\}$$

is precisely the closed-loop reach set. It is the set of such points $x \in \mathbb{R}^n$ for which there exists a strategy $\mathcal{U} \in U_\mathcal{C}$ which for any solution $x(t)$ of (69), $x(\tau) = x, t \in [t_0, \tau]$, ensures the inequality $d(x(t_0), X^0) \leq \mu$. Due to the structure of (69) ($A(t) \equiv 0$), this is equivalent to the following *definition of closed-loop reachability sets.*

DEFINITION 4.1. *A closed-loop reachability set $\mathcal{X}_\mu(\tau)$ is the set of such points $x \in \mathbb{R}^n$ for each of which there exists a strategy $\mathcal{U} \in U_\mathcal{C}$ that for every $v(\cdot) \in V_\mathcal{O}$ assigns a point $x^0 \in X^0$, such that every solution $x[t]$ of the differential inclusion*

$$\dot{x} \in B(t)\mathcal{U}(t,x) + C(t)v(t),\ x(t_0) = x^0, \quad t_0 \leq t \leq \tau,$$

satisfies the inequality $d(x[\tau], x) \leq \mu$.

Once the principle of optimality (60) is true, it may also be used directly to derive (64)—the HJBI equation for the function $\mathcal{V}(t,x)$. Therefore, set $\mathcal{X}_\mu(\tau)$ (if nonempty) will be nothing else than the set $X(\tau, t_0, X^0)$ defined earlier as the limit of OLRSs.

5. Closed-loop reachability under uncertainty. We shall now show that each point of $X(t, t_0, X^0)$ may be reached from X^0 with a certain feedback control strategy $\mathcal{U}(t,x)$, whatever the function $v(t)$ is.

In order to do this, we shall need the notion of the *solvability set* (or, in other terms, "the backward reachability set"—see [12], [27], [16])—a set similar to $X(t, t_0, X^0)$ but constructed in backward time. We first recall from [14] some properties of these sets. Consider the following problem.

Problem (V). Find the *value function*

$$\mathcal{V}_*(t,x) \ = \ \min_{\mathcal{U}} \max_{x(\cdot)} \{d^2(x[t_1], \mathcal{M}) | \mathcal{U} \in U_\mathcal{C}, x(\cdot) \in \mathcal{X}_\mathcal{U}\}, \tag{70}$$

where $\mathcal{M}$ is a given convex compact set ($\mathcal{M} \in conv\mathbb{R}^n$) and $\mathcal{X}_\mathcal{U}$ is the variety of all trajectories $x(\cdot)$ of the differential inclusion (69), $x(\tau) = x,\ t \in [t_0, t_1]$, generated by a given strategy $\mathcal{U} \in U_\mathcal{C}$.

The formal HJBI equation for the value $\mathcal{V}_*(t,x)$ is

$$\frac{\partial \mathcal{V}_*}{\partial t} \ + \ \min_u \max_v \left(\frac{\partial \mathcal{V}_*}{\partial x}, B(t)u + C(t)v \right) = 0, \quad u \in \mathcal{P}(t), \quad v \in \mathcal{Q}(t), \tag{71}$$

with boundary condition

$$\mathcal{V}_*(t_1, x) \ = \ d^2(x, \mathcal{M}). \tag{72}$$

Equation (71) may be rewritten as

$$\frac{\partial \mathcal{V}_*}{\partial t} \ - \ \rho\left(\frac{\partial \mathcal{V}_*}{\partial x} | - B(t)\mathcal{P}(t) \right) + \rho\left(\frac{\partial \mathcal{V}_*}{\partial x} | C(t)\mathcal{Q}(t) \right) = 0. \tag{73}$$

An important feature is that function $\mathcal{V}_*(t,x)$ may be interpreted as a sequential maxmin similar to the one in section 3. Namely, taking the interval $\tau \leq t \leq t_1$, introduce a partition $\Sigma_k = \{\tau = \tau_0, \tau_1, \ldots, \tau_k, \tau_{k+1} = t_1\}$, $h_1 = \tau_{k+1} - \tau_k, \ldots, h_{i+1} =$

$\tau_{k+i+1} - \tau_{k+i}, \dots, h_{k+1} = \tau_1 - \tau_0$, similar to that of section 3. For the given partition, consider the recurrence relations

$$V^-_{*k}(t_1 - h_1, x)$$

$$= \left\{ \max_v \min_u d^2(x(t_1), \mathcal{M}) |\ t_1 - h_1 \le t \le t_1\ , x(t_1 - h_1) = x \right\},$$

$$V^-_{*k}(t_1 - h_1 - h_2, x)$$

$$= \left\{ \max_v \min_u V^-_{*k}(t_1 - h_1, x(t_1 - h_1)) |\ t_1 - h_1 - h_2 \le t \le t_1 - h_1\ , x(t_1 - h_1 - h_2) = x \right\},$$

$$V^-_{*k}(\tau, x) = \left\{ \max_v \min_u V^-_{*k}(\tau + h_{k+1}, x(\tau + h_{k+1})) |\ \tau \le t \le \tau + h_{k+1},\ x(\tau) = x \right\},$$

where $v(t) \in \mathcal{Q}(t),\ u(t) \in \mathcal{P}(t)$ almost everywhere in the respective intervals.

LEMMA 5.1 (see [14]). *With*

$$\max\{h_i : i = 1, \dots, k+1\} \to 0,\ k \to \infty,\ \Sigma_{i=1}^{k+1} h_i = t_1 - \tau, \tag{74}$$

there exists a pointwise limit

$$\mathcal{V}^-_*(\tau, x) = \lim_{k \to \infty} V^-_{*k}(\tau, x)$$

that does not depend upon the type of partition Σ_k.

The function $\mathcal{V}^-_*(\tau, x)$ *coincides with* $\mathcal{V}_*(\tau, x)$.

We shall refer to $V^-_*(\tau, x) = \mathcal{V}_*(\tau, x)$ as the *sequential maxmin*. This function enjoys properties similar to those of its "forward time" counterpart, the function $\mathcal{V}^-(\tau, x)$ of section 3. A similar construction is possible for a "backward" version of the sequential minmax.

The level set

$$\mathcal{W}(t, t_1, \mathcal{M}) = \{x : \mathcal{V}_*(t, x) \le 0\}$$

is referred to as the *closed-loop solvability set* CLSS at time $\tau = t$, from set $\mathcal{M}$. It may be presented as an alternated integral of Pontryagin—the Hausdorff limit of the sequence

$$\mathcal{I}_*(t, t_1, \mathcal{M}, \Sigma_k) \tag{75}$$

$$= \left(\dots \left(\mathcal{M} + \int_{t_1 - \sigma_1}^{t_1} B(\tau)\mathcal{P}(\tau) d\tau \right) \dot{-} \int_{t_1 - \sigma_1}^{t_1} C(\tau)\mathcal{Q}(\tau) d\tau \right) \dots$$

$$\dot{-} \int_t^{t + \sigma_k} C(\tau)\mathcal{Q}(\tau) d\tau,$$

under conditions (74). Also presumed is a nondegeneracy assumption similar to Assumption 4.2.

Assumption 5.1. For a given set $\mathcal{M} \in conv\mathbb{R}^n$, there exists a continuous function $\beta_3(t) \in \mathbb{R}^n$ and a number $\epsilon > 0$ such that

$$\beta_3(\tau_i) + \epsilon\mathcal{B}(0) \subseteq \mathcal{I}_*(\tau_i, t_1, \mathcal{M}, \Sigma_k) \tag{76}$$

for any $i = 1, \ldots, k+1$, whatever the partition Σ_k is.

This assumption is presumed in the next lemma.

LEMMA 5.2. *Under condition* (76) *there exists a Hausdorff limit* $\mathcal{I}_*(t, t_1, \mathcal{M})$*:*

$$\lim h(\mathcal{I}_*(t, t_1, \mathcal{M}, \Sigma_k)\ ,\ \mathcal{I}_*(t, t_1, \mathcal{M})) \ =\ 0.$$

This limit does not depend on the type of partition Σ_k *and coincides with the CLSS,* $\mathcal{W}(t, t_1, \mathcal{M})$*:*

$$\mathcal{I}_*(t, t_1, \mathcal{M}) = \mathcal{W}(t, t_1, \mathcal{M}). \tag{77}$$

From the theory of control under uncertainty and differential games, it is known that if a certain point $x^* \in \mathcal{W}(t, t_1, \mathcal{M})$, there exists a feedback strategy $\mathcal{U}(t, x) \in U_{\mathcal{C}}$ that steers system (10) from position $\{t, x^*\}$ $(x(t) = x^*)$ to set $\mathcal{M}$ *whatever the unknown disturbance* $v(\cdot)$ is [12], [29], [16]. Therefore, assuming $X(t_1, t, x^*) \neq \emptyset$ (which is true under further assumptions), we just have to prove the inclusion

$$x^* \in \mathcal{W}(t, t_1, X(t_1, t, x^*))$$

or, in view of the properties of $\mathcal{V}(t, x), \mathcal{V}_*(t, x)$, that

$$x^* \in \{x : \mathcal{V}_*(t, x|t_1, \mathcal{V}(t_1, x)) \leq 0\}.$$

Here $\mathcal{V}_*(t, x|t_1, \mathcal{V}(t_1, x)) = \mathcal{V}_*(t, x)$ is the solution to (71) with boundary condition

$$\mathcal{V}_*(t_1, x) = \mathcal{V}(t_1, x). \tag{78}$$

(Recall that $\mathcal{V}(t_1, x) = d^2(x, X(t_1, t, X^*))$.)

Due to the definition of the geometrical difference and of the integral $\mathcal{I}_-(t_1, t, x^*)$, one may check that

$$\mathcal{I}_-(t_1, t, x^*) = \mathcal{I}_-(t_1, t, 0) + x^*.$$

We thus have to prove the inclusion

$$0 \in \mathcal{I}_*(t, t_1, \mathcal{I}_-(t_1, t, 0)).$$

Under Assumptions 4.2 (a) taken for $X^0 = \{0\}$ and 5.1 for $\mathcal{M} = \{0\}$, or under Assumption 4.1, it is possible to observe, through direct calculation, using the properties of integrals $\mathcal{I}_*, \mathcal{I}_-$ (see formulas (29),(75)), that the following hold: $X(t_1, t, x^*) \neq \emptyset$ and

$$\mathcal{I}_*(t, t_1, \mathcal{I}_-(t_1, t, 0)) \supseteq \mathcal{I}_-(t_1, t, 0) + \mathcal{I}_*(t, t_1, 0) = \mathcal{R}(t, t_1),$$

where

$$0 \in \mathcal{R}(t, t_1),$$

and we arrive at Lemma 5.3.

LEMMA 5.3. *Under Assumptions* 4.2 (a) *taken for* $X^0 = \{0\}$ *and* 5.1 *for* $\mathcal{M} = \{0\}$, *or under Assumption* 4.1, *the following inclusion is true:*

$$x^* \in \mathcal{W}(t, t_1, X(t_1, t, x^*)); \tag{79}$$

moreover,

$$X(s, t, x^*) \subseteq \mathcal{W}(s, t_1, X(t_1, t, x^*)) \quad \forall s \in [t, t_1].$$

Inclusion (79) implies the existence of a feedback strategy $\mathcal{U}_*(t, x)$ that brings system (10) from $x^* = x(t)$ to $x(t_1) \in X(t_1, t, x^*)$.

THEOREM 5.1. *Under Assumptions* 4.2 (b), $X^0 = \{0\}$, *and* 5.1, $\mathcal{M} = \{0\}$, *or under Assumption* 4.1, *there exists a closed-loop strategy* $\mathcal{U}_*(t, x) \subseteq U_{\mathcal{C}}$ *that steers system* (10) *from any position* $\{t, x^*\}$ $(x^* = x(t))$ *to* $X(t_1, t, x^*)$.

The strategy $\mathcal{U}_*(t, x)$ *may be found through the solution* $\mathcal{V}(t, x)$ *of* (71), *with boundary condition* (78), *as*

$$\mathcal{U}_*(t, x) = \arg\min\{(\partial \mathcal{V}_*(t, x)/\partial x, u) | u \in \mathcal{P}(t)\} \tag{80}$$

(if the gradient $\partial \mathcal{V}_*(t, x)/\partial x$ *does exist at* $\{t, x\}$*), or, more generally, as*

$$\mathcal{U}_*(t, x) = \left\{u : \max_v \{dd^2(x, \mathcal{W}^*[t])/dt | v \in \mathcal{Q}(t)\} \leq 0\right\}, \tag{81}$$

where $\mathcal{W}^*[t] = \{x : \mathcal{V}_*(t, x) \leq 0\}$.

This is verified by differentiating $\mathcal{V}(t, x)$ with respect to t and checking that almost everywhere

$$\frac{d\mathcal{V}}{dt}|_{u=\mathcal{U}_*(t,x)} \leq 0 \quad \forall v(t) \in \mathcal{Q}(t)$$

(see [11], [16]).

The previous theorem ensures merely that *some* point of $X(t_1, t, x^*)$ may be reached from x^*. In order to demonstrate that *any point* $x^\star \in X(t_1, t, x^*)$ may be reached from position $\{t, x^*\}$, we have to prove the inclusion

$$x^* \in \mathcal{W}(t, t_1, x^\star) \tag{82}$$

for any $x^\star \in X(t_1, t, x^*)$ or

$$x^* \in \{x : \mathcal{V}_\star(t, x | \mathcal{V}_\star(t_0, x)) \leq 0\}, \tag{83}$$

provided $\mathcal{V}_\star(t_1, x^\star) \leq 0$. Here $\mathcal{V}_\star(t, x)$ is a solution to (66) with boundary condition

$$\mathcal{V}_\star(t_0, x) = d^2(x, x^\star), \tag{84}$$

where $x^\star \in X(t_1, t, x^*)$.

However, inclusions (82) and (83) again follow from the properties of $\mathcal{I}_-(t_1, t, 0), \mathcal{I}_*(t, t_1, x^*)$, assuming that both of these set-valued integrals are nonempty. The latter, in its turn, is again ensured by either Assumptions 4.2 (a), $X^0 = 0$, and 5.1, $\mathcal{M} = 0$, or Assumption 4.1. This leads to Theorem 5.2.

THEOREM 5.2. *Under either Assumptions* 4.1 (a), $X^0 = 0$, *and* 5.1, $\mathcal{M} = 0$, *or* 4.1 *there exists a closed-loop strategy* $\mathcal{U}_\star(t, x) \subseteq U_{\mathcal{C}}$ *that steers system* (10) *from any position* $\{t, x^*\}$ $(x^* = x(t))$ *to point* $x^\star \in X(t_1, t, x^*)$.

The strategy $\mathcal{U}_\star(t,x)$ *may be found through the solution* $\mathcal{V}_\star(t,x)$ *of* (71), *with boundary condition* (84), *as*

$$\mathcal{U}_\star(t,x) = \arg\min\{(\partial\mathcal{V}_\star(t,x)/\partial x, u) | u \in \mathcal{P}(t)\} \tag{85}$$

(if the gradient $\partial\mathcal{V}_\star(t,x)/\partial x$ *does exist at* $\{t,x\}$*) or, more generally, as*

$$\mathcal{U}_\star(t,x) = \left\{u : \max_v\{dd^2(x,\mathcal{W}^\star[t])/dt | v \in \mathcal{Q}(t)\} \leq 0\right\}, \tag{86}$$

where $\mathcal{W}^\star[t] = \{x : \mathcal{V}_\star(t,x) \leq 0\}$.

Remark 5.1. Assumptions 4.1 (a), $X^0 = 0$, and 5.1, $\mathcal{M} = 0$, are ensured by Assumption 4.1. If this does not hold, it is possible to go through all the procedures taking *μ-neighborhoods* of sets $X(\cdot), W(\cdot)$ rather than the sets themselves. Then one has to look for the $\mu(\cdot)$-reach sets $X(t,t_0,x^*,\mu(\cdot))$ and μ-solvability sets $W(t,t_1,x^\star,\mu(\cdot))$ with μ sufficiently large so that $X(t,t_0,X^0,\mu(\cdot)), W(t,t_1,x^\star,\mu(\cdot))$ would surely be nonempty.

Remark 5.2. The emphasis of this paper is to discuss the issue of reachability under uncertainty governed by unknown but bounded disturbances. This topic was studied here through a reduction to the calculation of value functions for a successive problem on sequential minmax and maxmin of certain *distance functions* or their squares. The latter problems were dealt with via techniques of convex analysis and set-valued calculus. However, the solution schemes of this paper naturally allow a more general situation, which is to substitute the distance function $d(x,\mathcal{M})$ by *any proper convex function* $\phi(x)$, for example, with similar results passing through. The more general problems then reduce to those of this paper.

Thus, given terminal cost function $\phi(x)$, it may readily generate a terminal set $\mathcal{M}$ as a level set $\mathcal{M} = \{x : \phi(x) \leq \alpha\}$ for some α, with support function [25]

$$\rho(l|\mathcal{M}) = \inf\{\lambda(\phi^*(l/\lambda) + \alpha) | \lambda > 0\}.$$

The given formalisms for describing reachability are not the only ones available. We further indicate yet another formal scheme.

6. Reachability and the funnel equations. In this section, we briefly indicate some connections between the previous results and those that can be obtained through evolution equations of the "funnel type" [2], [16].

Consider the evolution equations

$$\lim_{\sigma\to\infty} \sigma^{-1} h_+(\mathcal{X}^-(t+\sigma), (\mathcal{X}^-(t) + \sigma B(t)\mathcal{P}(t)) \dot{-} \sigma(-C(t)\mathcal{Q}(t))) = 0, \tag{87}$$

with initial condition

$$\mathcal{X}^-(t_0) = X^0$$

and

$$\lim_{\sigma\to\infty} \sigma^{-1} h_+(\mathcal{X}^+(t+\sigma), (\mathcal{X}^-(t) \dot{-} \sigma(-C(t)\mathcal{Q}(t))) + \sigma B(t)\mathcal{P}(t)) = 0, \tag{88}$$

with

$$\mathcal{X}^+(t_0) = X^0.$$

Under some regularity assumptions (similar to Assumption 4.1) which ensure that all the sets that appear in (87), (88) are nonempty, these equations have solutions which turn out to be *set-valued*. The solutions $\mathcal{X}^-(t), \mathcal{X}^+(t)$ satisfy (87), (88) almost everywhere. But they need not be unique. However, the property of uniqueness may be restored if we presume that $\mathcal{X}^-(t), \mathcal{X}^+(t)$ are the *(inclusion) maximal* solutions (see[16, sections 1.3, 1.7]). (A solution $\mathcal{X}_0(t)$ to a funnel equation of type (87), (88) is maximal if it satisfies the inclusion $\mathcal{X}_0(t) \supseteq \mathcal{X}(t)$ for any other solution $\mathcal{X}(t)$ to the respective equation with the same initial condition.)

Equations (87), (88) may be interpreted as some limit form of the recurrence equations

$$\mathcal{X}^-(t+\sigma) = (\mathcal{X}^-(t) + \sigma B(t)\mathcal{P}(t)) \dot{-} \sigma(-C(t)\mathcal{Q}(t)),\ X^-(t_0) = X^0 \tag{89}$$

and

$$\mathcal{X}^+(t+\sigma) = (\mathcal{X}^+(t) \dot{-} \sigma(-C(t)\mathcal{Q}(t))) + \sigma B(t)\mathcal{P}(t), \tag{90}$$

$$\mathcal{X}^+(t_0) = X^0 + r\sigma\mathcal{B}_1(0),\ r\mathcal{B}_1(0) \dot{-} \sigma(-C(t)\mathcal{Q}(t)) \neq \emptyset.$$

Indeed, taking, for example, $\sigma = (\tau - t_0)/k$ and solving the recurrence equation (89) for values of time $t_0 = 0, t_0 + \sigma = 1, \ldots, \tau - \sigma, \tau = k$, from $\mathcal{X}^-(t_0) = X^0 = X^-_\sigma(0|\tau)$ to $\mathcal{X}^-(\tau) = X^-_\sigma(k|\tau)$, we observe that $\mathcal{X}^-_\sigma(k|\tau)$ is similar to $\mathcal{I}^-(\tau, t_0, X^0, \Sigma_k) = X^-_k(\tau, t_0, X^0)$, provided Σ_k is selected with constant $\sigma_i = \sigma = (\tau - t_0)/k$. Namely, formula

$$\mathcal{X}^-_\sigma(k) = (\ldots(X^0 + B(\sigma)\mathcal{P}(\sigma)) \dot{-} (-C(\sigma)\mathcal{Q}(s)) + \cdots$$

$$+B(\tau-\sigma)\mathcal{P}(\tau-\sigma) \dot{-} (-C(\tau-\sigma)\mathcal{Q}(\tau-\sigma)) + \cdots + B(\tau)\mathcal{P}(\tau)) \dot{-} (-C(\tau)\mathcal{Q}(\tau))$$

is similar to (29)(when $\mu = 0$) and to (43).

Under Assumtion 4.1, a direct calculation leads to the next conclusions.

LEMMA 6.1. *The following relations are true with* $k \to \infty, (\sigma(k) \to 0)$*:*

(i) $\lim h(\mathcal{X}^-_\sigma(k|\tau), X^-_k(\tau, t_0, X^0)) = 0.$

(ii) $\lim h(\mathcal{X}^-_\sigma(k|\tau), \mathcal{X}^-(\tau)) = 0.$

(iii) *Function* $\mathcal{I}^-[\tau] = \mathcal{I}^-(\tau, t_0, X^0) = \lim \mathcal{I}^-(\tau, t_0, X^0, \Sigma_k) = \mathcal{X}^-(\tau)$ *is a maximal solution to* (87) *with* $\mathcal{X}^-(t_0) = X^0$.

Therefore, the closed-loop reach set $\mathcal{X}(\tau, t_0, X^0)$ may also be calculated through the funnel equation (87), which therefore also describes *the dynamics of the level sets of the value function* $\mathcal{V}(\tau, x)$*—the solution to the forward HJBI equation* (66).

Remark 6.1. As we have seen, (87) describes the evolution of the alternated integral $\mathcal{I}^-(\tau, t_0, X^0)$. Similarly to that, (88) describes the evolution of the alternated integral $\mathcal{I}^+(\tau, t_0, X^0)$. The recurrence equations (89), (90) may then serve to be the basis of numerical schemes for calculating the reach sets.

7. Example. Consider the system

$$\dot{x_1} = x_2 + v, \tag{91}$$

$$\dot{x_2} = u,$$

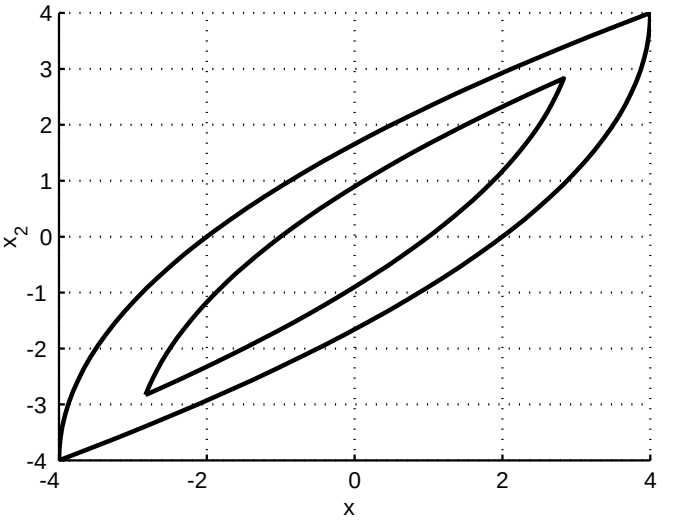


FIG. 1.

defined on the interval $[0, \tau]$, with hard bounds

$$|u| \le r_1,\ |v| \le r_2,\ r_1 > 0,\ r_2 > 0,$$

on the control u and the uncertain disturbance v.

As is known (see, for example, [17]), a parametric representation of the boundary of the reach set $X(\tau, t_0, x^0 | \mathcal{P}(\cdot), \{0\})$ of system (91) without uncertainty ($v(t) \equiv 0$) is given by two curves (see the *external* set in Figure 1, generated for $x^0 = 0, \tau = 2, r_1 = 2, r_2 = 0$):

$$x_1(t) = x_1^0 + x_2^0 t \pm r_1(t^2/2 - \sigma^2),$$

$$x_2(t) = x_2^0 \pm r_1(2\sigma + t),$$

and where $\sigma \le 0$ is the parameter (the values $\sigma > 0$ correspond to the vertices of $X(\tau, t_0, x^0 | \mathcal{P}(\cdot), \{0\})$).

Similarly, the reach set $X(\tau, t_0, X^0 | \{0\}, \mathcal{Q}(\cdot))$ in the variable v is given by the curves

$$x_1(t) = x_1^0 \pm x_2^0 t \pm r_2 t,$$

$$x_2(t) = x_2^0.$$

According to (7) , the set

$$X^-(\tau, 0, x^0, 0) = X(\tau, t_0, x^0 | \mathcal{P}(\cdot), \{0\}) \dot{-} X(\tau, t_0, 0 | \{0\}, \mathcal{Q}(\cdot)),$$

which leads to a parametrization of the boundary of this set in the form

$$x_1(t) = x_1^0 + x_2^0 t \pm r_1(t^2/2 - \sigma^2) \pm r_2 t,$$

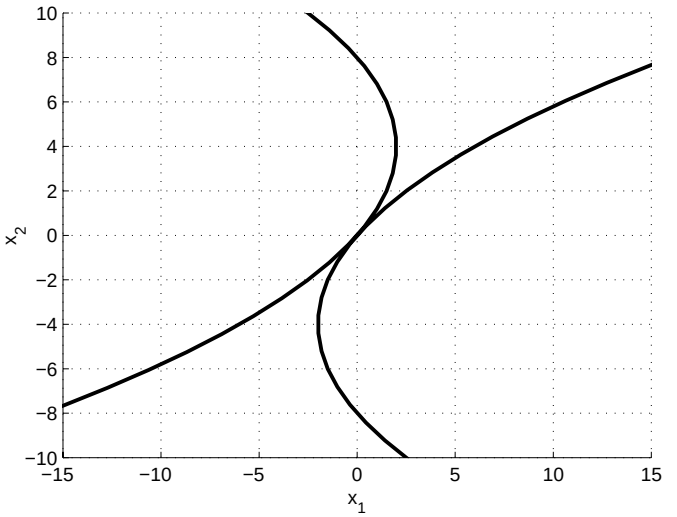


FIG. 2.

$$x_2(t) = x^0 - 2 \pm r_1(2\sigma + t)$$

(see the *internal set* in Figure 1, generated for $X^0 = 0, \tau = 2, r_1 = 2, r_2 = 1$). Clearly,

$$X^-(\tau, 0, x^0, 0) \subset X(\tau, 0, x^0 | \mathcal{P}(\cdot), \{0\})$$

so that the OLRS $X^-(\tau, 0, x^0, 0)$ under uncertainty is smaller than $X(\tau, 0, x^0 | \mathcal{P}(\cdot), \{0\})$—the reach set without uncertainty.

Let us now look for the OLRS $X^-(\tau, 0, x^0, 0)$ with one correction at time $t = \tau_1$. Taking $\tau = 2, \tau_1 = 1, x^0 = 0, r_1 = 1, r_2 = 1/2$, one may figure out that set $X^-(\tau_1, 0, 0, 0)$ has to be bounded by two curves (Figure 2)

$$x_1(t) = \pm(1/2 - 2\sigma^2),$$

$$x_2(t) = \pm 2(2\sigma + 1),$$

which gives $X^-(\tau_1, 0, 0, 0) = \{0\}$. Then for $r_2 > 1/2$ we have $X^-(\tau_1, 0, 0, 0) = \emptyset$, and for $r_2 < 1/2$ we come to $\mathrm{int} X^-(\tau_1, 0, 0, 0) \neq \emptyset$.

Continuing with $r_2 = 1/2$, we have

$$X^-(2, 1, X^-(1, 0, 0, 0), 0) = \{0\} \in X^-(2, 0, 0, 0).$$

We also observe that with $r_2 > 1/2$ we have $X^-(\tau_1, 0, 0, \mu) = \{0\} \neq \emptyset$ if $\mu > 0$ is sufficiently large.

As indicated above, sets $X^+(\tau, t_0, x^0, \mu)$ turn out to be empty unless μ is sufficiently large. Continuing our example further, for $\tau = 2, \tau_1 = 1, x^0 = 0, r_1 = 2, r_2 = 1$, we have

$$X^-(2, 0, 0, \mu) = (\mathcal{B}_\mu(0) + X(2, 0, 0 | \mathcal{P}(\cdot), \{0\})) \dot{-} X(2, 0, 0 | \{0\}, \mathcal{Q}(\cdot)),$$

and

$$X^+(2, 0, 0, \mu) = (\mathcal{B}_\mu(0) \dot{-} X(2, 0, 0 | \{0\}, \mathcal{Q}(\cdot))) + X(2, 0, 0, | \mathcal{P}(\cdot), \{0\}).$$

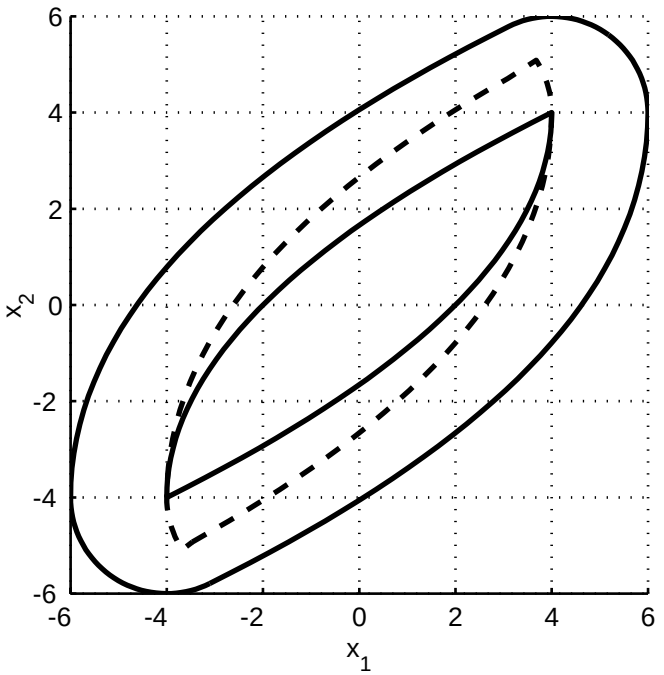


FIG. 3.

The last set is nonvoid if μ is such that $\mathcal{B}_\mu(0) \dot{-} X(2,0,0,|\{0\},\mathcal{Q}(\cdot)) = \mathcal{X}^+[\mu] \neq \emptyset$. The smallest value μ^0 of all such μ ensures $\mathcal{X}^+[\mu^0] = \{0\}$.

For all $\mu \geq \mu^0$ it is then possible to compare sets $X^-(2,0,0,\mu)$ and $X^+(2,0,0,\mu)$, observing that the latter is smaller than the former (see Figure 3, where $X^-(2,0,0,0)$ is shown by the *internal continuous curve*, $X^-(2,0,0,1)$ by the *external continuous curve,* and $X^+(2,0,0,1)$ by the dashed curve).

8. Conclusion. In this paper, we deal with one of the recent problems in reachability analysis which is to specify the sets of points that can be reached by a controlled system *despite the unknown but bounded disturbances in the system inputs.* The paper gives a description of several notions of such reachability and indicates schemes to calculate various types of reach sets. We consider systems with linear structure and closed-loop controls that are generally nonlinear. In particular, we emphasize the difference between reachability under open-loop and closed-loop controls. We distinguish open-loop controls of the anticipative type, which presume the disturbances to be known in advance, and of the nonanticipative type, which presume no such knowledge. The nonanticipative OLRSs is smaller than the one for anticipative open-loop controls, and the closed-loop reach set (which is always nonanticipative) lies in between. Intermediate reach sets are those generated by piecewise closed-loop controls that allow on-line measurements of the state space variable at isolated instants of time—the points of correction. Increasing the number of corrections to infinity and keeping them dense within the interval under consideration, we came to the case of continuous corrections—the solution to the problem of reachability under closed-loop (feedback) control.

The various types of reach sets introduced here *were calculated* through two alternative presentations, namely, *either through operations on set-valued integrals or as level sets for value functions* in sequential problems on maxmin or minmax for certain distance functions.

For the closed-loop reachability problem the corresponding value function defines a mapping that satisfies *the semigroup property.* This property allowed us to formulate *the principle of optimality under uncertainty* for the class of problems considered here. The last principle allowed us to demonstrate that the closed-loop reach set under uncertainty is the level set for the solution to a *forward* HJBI equation. On the other hand, the feedback control strategy that steers a point to its closed-loop reach set (whatever the disturbance is) may be found from the solution to a *backward* HJBI equation whose boundary condition is taken from the solution of the earlier mentioned forward HJBI equation.

This paper leaves many issues for further investigation. For example, there is a strong demand from many applied areas to calculate reach sets under uncertainty. However, the given solutions to the problem are not simple to calculate. Among the nearest issues may be the calculation of the reach sets of this paper through ellipsoidal approximations along the schemes of [16], [17]. Then, of course, comes the propagation of the results to nonlinear systems. Here the application of the HJBI technique seems to allow some progress. Needless to say, similar problems could also be posed for systems with uncertainty in its parameters or in the model itself as well as for other types of controlled transition systems.

Acknowledgment. We thank Oleg Botchkarev for the figures.

REFERENCES

[1] R. Alur, T. A. Henzinger, and O. Kupferman, *Alternating time temporal logic*, in Proceedings of the IEEE Symposium on Foundations of Computer Science, Miami Beach, FL, 1997, pp. 100–109.

[2] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser Boston, Boston, 1990.

[3] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser Boston, Boston, 1997.

[4] T. Basar and P. Bernhard, H^∞ *Optimal Control and Related Minimax Design Problems*, 2nd ed., Birkhäuser Boston, Boston, 1995.

[5] L. De Alfaro, *Stochastic transition systems*, in Proceedings of the 9th International Conference on Concurrency Theory, Springer-Verlag, New York, 1998, pp. 423–438.

[6] V. F. Demyanov and A. M. Rubinov, *Constructive Nonsmooth Analysis*, Peter Lang, Frankfurt, 1995.

[7] K. Y. Fan, *Minmax theorems*, Proc. Natl. Acad. Sci. USA, 39 (1993), pp. 42–47.

[8] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[9] G. E. Ivanov and E. C. Polovinkin, *On strongly convex linear differential games*, Differential Equations, 31 (1995), pp. 1603–1612.

[10] H. Knobloch, A. Isidori, and D. Flockerzi, *Topics in Control Theory*, DMV Sem. 22, Birkhäuser, Basel, 1993.

[11] N. N. Krasovskii, *Game-Theoretic Problems on the Encounter of Motions*, Nauka, Moscow, 1970 (in Russian); English translation: *Rendezvous Game Problems,* Nat. Tech. Inf. Serv., Springfield, VA, 1971.

[12] N. N. Krasovski and A. N. Subbotin, *Positional Differential Games*, Springer-Verlag, New York, 1988.

[13] A. B. Kurzhanski, *Control and Observation Under Uncertainty*, Nauka, Moscow, 1977.

[14] A. B. Kurzhanski, *Pontryagin's alternated integral in the theory of control synthesis*, Tr. Mat. Inst. Steklova, 224 (1999), pp. 234–248 (in Russian).

[15] A. B. Kurzhanski and N. B. Melnikov, *On the problem of control synthesis: The alternated integral of Pontryagin*, Sb. Math., 191 (2000), pp. 849–881.

[16] A. B. Kurzhanski and I. Vályi, *Ellipsoidal Calculus for Estimation and Control*, Birkhäuser Boston, Boston, 1997.
[17] A. B. Kurzhanski and P. P. Varaiya, *Ellipsoidal techniques for reachability analysis*, in Hybrid Systems: Computation and Control, Lecture Notes in Comput. Sci. 1790, B. Krogh and N. Lunch, eds., Springer-Verlag, New York, 2000, pp. 202–214.
[18] A. B. Kurzhanski and P. Varaiya, *Dynamic optimization for reachability problems*, J. Optim. Theory Appl., 108 (2000), pp. 227–251.
[19] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, Wiley, New York, 1961.
[20] G. Leitmann, *Optimality and reachability with feedback controls*, in Dynamical Systems and Microphysics, A. Blaquiere and G. Leitmann, eds., Academic Press, New York, 1982, pp. 119–141.
[21] P.-L. Lions and P. E. Souganidis, *Differential games, optimal control and directional derivatives of viscosity solutions of Bellman's and Isaacs' equations*, SIAM J. Control Optim., 23 (1985), pp. 566–583.
[22] J. Lygeros, C. Tomlin, and S. Sastri, *Controllers for reachability specifications for hybrid systems*, Automatica J. IFAC, 35 (1999), pp. 349–370.
[23] L. S. Pontryagin, *Linear differential games of pursuit*, Mat. Sb. (N.S.), 112 (1980), pp. 307–330 (in Russian).
[24] A. Puri, V. Borkar, and P. Varaiya, *ϵ-approximations of differential inclusions*, in Hybrid Systems, Lecture Notes in Comput. Sci. 1201, R. Alur, T.A. Henzinger, and E.D. Sontag, eds., Springer-Verlag, New York, 1998, pp. 109–123.
[25] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis*, Springer-Verlag, New York, 1998.
[26] A. I. Subbotin, *Generalized Solutions of First Order PDEs: The Dynamic Optimization Perspective*, Birkhäuser Boston, Boston, 1995.
[27] P. P. Varaiya, *On the existence of solutions to a differential game*, SIAM J. Control, 5 (1967), pp. 153–162.
[28] P. P. Varaiya and J. Lin, *Existence of saddle points in differential games*, SIAM J. Control, 7 (1969), pp. 141–157.
[29] P. P. Varaiya, *Reach set computation using optimal control*, in Proceedings of the KIT Workshop on Verification of Hybrid Systems, Verimag, Grenoble, 1998.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 1, pp. 217–252

CONTROL OF POLLING IN PRESENCE OF VACATIONS IN HEAVY TRAFFIC WITH APPLICATIONS TO SATELLITE AND MOBILE RADIO SYSTEMS*

EITAN ALTMAN† AND HAROLD J. KUSHNER‡

Abstract. Consider a queueing system with many queues, each with its own input stream, but with only one server. The server must allocate its time among the queues to minimize or nearly minimize some cost criterion. The allocation of time among the queues is often called polling and is the subject of a large literature. Usually, it is assumed that the queues are always available, and the server can allocate at will. We consider the case where the queues are not always available due to disruption of the connection between them and the server. Such occurrences are common in wireless communications, where any of the mobile sources might become unavailable to the server from time to time due to obstacles, atmospheric or other effects. The possibility of such "vacations" complicates the polling problem enormously. Due to the complexity of the basic problem we analyze it in the heavy traffic regime where the server has little idle time over the average requirements. It is shown that the suitable scaled total workloads converge to a controlled limit diffusion process with jumps. The jumps are due to the effects of the vacations. The control enters the dynamics only via its value just before a vacation begins; hence it is only via the jump value that the control affects the dynamics. This type of model has not received much attention. The individual queued workloads and job numbers can be recovered (asymptotically) from the limit scaled workload. This state space collapse is critical for the effective numerical and analytical work, since the limit process is one dimensional. It is also shown, under appropriate conditions, that the arrival process during a vacation can be approximated by the scaled "fluid" process. With a suitable nonlinear discounted cost rate, it is shown that the optimal costs for the physical problems converge to that for the limit problem as the traffic intensity approaches its heavy traffic limit. Explicit solutions are obtained in some simple but important cases, and the $c\mu$-rule is asymptotically optimal if there are no vacations. The stability of the queues is analyzed via a perturbed Liapunov function method, under quite general conditions on the data. Finally, we extend the results to unreliable channels where the data might be received with errors and need to be retransmitted.

Key words. heavy traffic analysis, queueing networks, scheduling queues, communication networks, wireless communications, mobile communications, polling, optimal stochastic control

AMS subject classifications. 90B22, 60K25, 60K30, 60F17, 93E20, 93E25

PII. S0363012999358464

1. Introduction.

Consider a queueing system with several queues and a single server. The problem of assigning service among the competing queues in an optimal way has been studied extensively in the last half of the century, starting with [9, pp. 84–85]. The assignment is often called "polling." For linear holding costs, the fixed-priority policy known as the $c\mu$-rule (and other rules closely related to it) has been shown to be an optimal policy under a variety of statistical assumptions and cost structures (see, e.g., [1, 3, 4, 9] and references therein). Due to Little's rule, this policy turns out to minimize also the overall average expected waiting time in the system. The problem can be considered to be one in optimal stochastic control.

*Received by the editors July 6, 1999; accepted for publication (in revised form) November 22, 2000; published electronically May 14, 2002. This work was supported by contracts DAAD 19-99-1-00223 from the Army Research Office and NSF grant ECS9703895.

http://www.siam.org/journals/sicon/41-1/35846.html

†INRIA, 2004 Route des Lucioles, B.P. 93, 06902 Sophia-Antipolis Cedex, France (Altman@sophia.inria.fr).

‡Division of Applied Mathematics, and Lefschetz Center for Dynamical Systems, Brown University, Providence, RI 02912 (hjk@dam.brown.edu).

We are concerned with this assignment or polling problem when the connections between each queue and the server are broken at random times and for random durations. Such intervals during which a queue is not available to the server (even if it wished to poll it) are often called vacations in the queueing literature. The possibility of such vacations complicates the control problem considerably, since the possibility that any queue might not be available to the server at any future time needs to be accounted for in choosing the current server allocation.

This problem is of considerable importance in contemporary wireless communications, where the queues are in the mobile sources which generate data to be transmitted and the server is the channel or antenna of the base station. At each time, one assigns the channel to one of the sources (i.e., points the antenna in the direction of that source). In more complex cases with so-called smart antennas, the channel can be shared among the sources in a controlled way, but that possibility will not be considered here.

Recently, Tassiulas and Ephremides [28] have considered this problem of how to assign service to competing queues in the presence of random connectivity. The motivation concerned the dynamic assignment of transmission access to a channel between mobile terminals, any of which might be unavailable from time to time due to physical obstacles or to propagation problems (atmospheric attenuation, interference, noise, fading, etc.). In the context of satellite communications, a survey of such problems can be found in [11].

The classical $c\mu$-rule turns to be not only far from optimal for this system, but in fact the system may be unstable when any fixed priority policy is used. Tassiulas and Ephremides [28] and Tassiulas and Papavassiliou [29] considered the problem of obtaining a dynamic assignment policy that maximizes the throughput. The solution methodology is based on stability analysis using Liapunov functions; first a necessary condition for stability is identified, which holds under any policy. Then a particular policy is identified for which a sufficient stability condition coincides with the above necessary stability condition. It then turns out that this policy stabilizes the system under the largest range of input rates and is thus shown to maximize the throughput that the system can handle. Such a policy has a very simple form [28]: assign a transmission opportunity to the longest connected queue.

It turns out that there is a very large class of policies other than the one above which also achieve that maximum stability region and maximum throughput. (For example, if we first multiply the length of each queue by some [queue dependent] positive constant and then assign transmission to the queue with longest *weighted* length, still an optimal throughput is achieved; this is suggested by our stability analysis in section 6.) The aim of this paper is thus to consider control and optimal control under more sensitive cost criteria, which can not only maximize the throughput but can also minimize some expected holding costs (or in particular, discounted mean values of a broad class of functions of the queue lengths or expected workload in the system).

Due to the complexity of the system and the generality of the statistical assumptions, we consider the optimal control problem only in an asymptotic sense; i.e., the one obtained by an appropriate scaling, corresponding to the heavy traffic regime, where the server has little spare capacity over the mean requirements. As usual with heavy traffic analysis, the limit system is substantially simpler than the original physical system. The aim is to use the limit model to get nearly optimal controls and approximations to the optimal value functions for the actual physical system under heavy traffic conditions. For some large class of policies, we establish the convergence

of the total workload processes to a one dimensional diffusion process with jumps, where the jumps are due to the possibility that the server will have no work to do during part of a vacation of some source. The individual workloads and queue lengths can be approximated in terms of this limit process; hence there is a substantial reduction in dimension from that of the original problem to unity. This limit result is then used to obtain a closed form solution for the asymptotic problem for several types of cost functions. In particular, a closed form solution is obtained for the case when the cost corresponds to the total workload in the system.

A problem related to the one we solve here has been treated in [8] and references therein. There too, optimal scheduling of service opportunities is considered. However, the problem of random (unpredictable) disconnectivity is not considered there; instead, there are predictable instances in which service opportunities appear. These correspond to transmission opportunities between adjacent satellites which use intersatellite links within a satellite constellation. As in [28], the criterion is to maximize the throughput. Several policies are proposed there and their performance is compared.

The structure of the paper is as follows. Section 2 describes the problem and lists the assumptions which are needed to get the weak convergence results of section 3. Jobs (batches of data) arrive at the queues of the individual sources at random times, and in random amounts. It is assumed that the vacations are relatively "rare" and that the ratio of the vacation intervals to the intervacation intervals is small. Nevertheless they have a very important effect on the performance. For notational simplicity, only two sources are considered in the details. However, all of the results hold irrespective of the number of sources, and we comment on the extensions. The analysis contains results which are of broader interest. Examples are the proof that the arrival (and, indeed, the [suitably time scaled] workload and queue processes) process *during a vacation* can be well approximated by a "fluid" process under heavy traffic and that the individual queue sizes can be approximated by linear functions of the individual queued workloads. The policy affects the limit process only via the magnitude of the jumps. In particular, the jumps depend only on the "control" values just before a vacation begins. If we restrict the policy to being a member of a large class of piecewise continuous feedback policies, then Theorem 3.2 shows that the individual workloads can be well approximated by linear functions of the total workload, under heavy traffic.

The discounted cost function is introduced in section 4. The limit control problem appears to be nonstandard, owing to the special way that the control appears in the dynamics, even though it appears in the cost function in a standard way. To get weak convergence of the cost or optimal cost values, one needs a uniform integrability condition as well as weak convergence, and this is dealt with as well.

The optimal control is computed under some assumptions on the costs in section 5. In section 6 we establish the stability conditions for a large class of policies. We extend our model and results in section 7, where we consider the case of unreliable channels which may require retransmissions of erroneous information.

2. The problem formulation. [1] There are two sources with inputs which generate data in some random way. Any number of sources can be used, but we stick to two for notational simplicity. The results for the general case will be apparent

[1] The book by H. Kushner, *Heavy Traffic Analysis of Controlled Queueing and Communication Networks*, Vol. 2, Springer, New York, 2001, has much information on related problems.

from the results for the two source case. The data (or jobs) created by each source are queued there, and the server alternates service (i.e., polling) between them in some way to be determined later. There is no switchover time in going from one source to another. Each of the sources will become unavailable to the server from time to time. The periods of unavailability are called vacations, in accordance with current terminology. Strictly speaking, it is the connection from the source to the server which is "on vacation," but we simply say that the source itself is on vacation. A source that is on vacation cannot be polled, but the data or job inputs which it creates still arrive to its queue. In that case, the content of the unavailable queue grows, but the server can only work on the available queue. Service is nonpreemptive, and first-come-first-served (FCFS): i.e., a job once started is completed, assuming that there is no intervening vacation. Also, the system is assumed to be "work conserving" in that the server will not idle if there is work to do on an available queue. Suppose that a vacation starts in the middle of a job. We suppose *either* (a) that the job is allowed to be completed, or (b) that it is stopped, but when that source is next served the job needs only its residual time, or (c) that the entire job needs to be redone. Because of the "rarity" of vacations under the assumptions (A2.3) and (A2.4) to be introduced below, the results will be the same for all cases. For specificity, and without loss of generality in the results, we suppose that both sources are available at time 0.

Our approach is that of heavy traffic analysis, where the "spare capacity" of the system is small. As usual in heavy traffic analysis, the problem is embedded in a sequence of problems, indexed by n. As $n \to \infty$, the spare capacity of the server goes to zero, and this is quantified in (A2.2). Let $\{\Delta_{i,l}^{a,n}, l < \infty\}$ denote the interarrival times for jobs at source $i = 1, 2$, and let $\{\Delta_{i,l}^{d,n}, l < \infty\}$ denote the corresponding work (real time) requirements.

Comment on weak convergence. Let $\mathbb{R}^r$ denote r-dimensional Euclidean space. The path space for all of the processes will be $D(\mathbb{R}^r; 0, \infty)$, the space of $\mathbb{R}^r$-valued functions which are right continuous and have left-hand limits, for the appropriate values of r. If $r = 1$, we write simply $D(\mathbb{R}; 0, \infty)$. The Skorohod topology will be used on this space. All of the concepts concerning weak convergence which will be used can be found in the standard references [5, 12]. Summaries with applications to stochastic systems can be found in [17, 19]. The following is a convenient criterion for tightness in $D(\mathbb{R}; 0, \infty)$. It will be used implicitly, without specific mention. Let $Y^n(\cdot)$ be a sequence of processes with paths in $D(\mathbb{R}; 0, \infty)$. Let $\mathcal{T}^n(t)$ denote the stopping times with respect to the filtration engendered by $Y^n(\cdot)$ and which are no larger than t. If, for each t,

$$\lim_{\delta \to 0} \sup_n \sup_{\tau \in \mathcal{T}^n(t)} E\left(1 \wedge |Y^n(\tau + \delta) - Y^n(\tau)|\right) = 0 \tag{2.1a}$$

and

$$\{Y^n(s) : n < \infty, s \le t\} \text{ is tight in } \mathbb{R}, \tag{2.1b}$$

then $\{Y^n(\cdot)\}$ is tight [12].

Notation and assumptions. For some centering constants $\bar{\Delta}_i^{\alpha,n}, \alpha = a, d$, whose properties will be specified below, define the processes

$$w_i^{\alpha,n}(t) = \frac{1}{\sqrt{n}} \sum_{l=1}^{nt} \left[1 - \frac{\Delta_{i,l}^{\alpha,n}}{\bar{\Delta}_i^{\alpha,n}}\right], \ t \ge 0, \ \alpha = a, d, i = 1, 2. \tag{2.2}$$

When an index of summation is nt we mean the *integer* part $[nt]$. Let $x_i^n(t)$ denote $1/\sqrt{n}$ times the number of jobs in queue i at *real time* nt, including the one in service, if any. Let $WL_i^n(t)$ (called the workload at queue i) denote $1/\sqrt{n}$ times the real time that the server must work to complete all of the jobs which are in queue i at *real time* nt. Thus, time is scaled by $1/n$ and the state by $1/\sqrt{n}$. Define the *total workload* $WL^n(t) = \sum_i WL_i^n(t)$. By *scaled work* we mean $1/\sqrt{n}$ times the actual physical work in question. The expression *scaled time* of some event always refers to the real time of that event divided by n.

Define the index of a job at queue i as one plus the number of jobs that arrived or were there before it, starting with the ordered $\sqrt{n}x_i^n(0)$ jobs which are in that queue at time zero. Let $L_i^n(t) \geq 0$ denote the index of the last customer to enter service in queue i at or before real time nt. For future use, note that $x_i^n(\cdot)$ and $WL_i^n(\cdot)$ are related by

$$WL_i^n(t) \in \left[\frac{1}{\sqrt{n}} \sum_{l=L_i^n(t)}^{L_i^n(t)+\sqrt{n}x_i^n(t)-1} \Delta_{i,l}^{d,n}, \frac{1}{\sqrt{n}} \sum_{l=L_i^n(t)+1}^{L_i^n(t)+\sqrt{n}x_i^n(t)-1} \Delta_{i,l}^{d,n} \right]. \tag{2.3}$$

We will use the following conditions. By "driving random variables," we mean the set of initial conditions, the arrival times and service requirements, and the starting and stopping times of the vacations.

A2.0. *For each n, $x_i^n(0), WL_i^n(0), i = 1, 2$, are independent of all of the "future" driving random variables. None of the sources is on vacation at $t = 0$.* (The last sentence is used only to simplify the notation.)

A2.1. *For $\alpha = a, d; i = 1, 2$, there are constants $\bar{\Delta}_i^\alpha$ such that*

$$\bar{\Delta}_i^{\alpha,n} \to \bar{\Delta}_i^\alpha.$$

As $n \to \infty$, the sequences $w_i^{\alpha,n}(\cdot), n < \infty, \alpha = a, d; i = 1, 2$, converge weakly to mutually independent Wiener processes $w_i^\alpha(\cdot), \alpha = a, d, i = 1, 2$, with variance parameters $\sigma_{\alpha,i}^2$, respectively.

Define $\bar{\lambda}_i^{\alpha,n} = 1/\bar{\Delta}_i^{\alpha,n}$, which will be used interchangeably, and similarly for $\bar{\lambda}_i^\alpha = 1/\bar{\Delta}_i^\alpha$. Define the traffic intensities

$$\rho_i^n = \bar{\Delta}_i^{d,n}/\bar{\Delta}_i^{a,n} = \bar{\Delta}_i^{d,n}\bar{\lambda}_i^{a,n}, \quad \rho_i = \bar{\Delta}_i^d/\bar{\Delta}_i^a = \bar{\Delta}_i^d\bar{\lambda}_i^a.$$

A2.2. *There is a real number b such that*

$$\lim_n \sqrt{n}\left[\sum_i \rho_i^n - 1\right] = b.$$

Note that (A2.2) implies that $\sum_i \rho_i = 1$.

A2.3. *For each n, i, the intervals between the end of the lth vacation and the start of the next one for source i are denoted by $n\tau_{i,l}^{s,n}, l = 1, \ldots$. They are mutually independent, exponentially distributed, independent of all the other "driving" random variables and have rate $\bar{\lambda}_i^{s,n}/n$, where $\bar{\lambda}_i^{s,n}$ converges to $\bar{\lambda}_i^s > 0$ as $n \to \infty$. The intervals for the different sources are mutually independent.*

A2.4. *For each n, i, there are mutually independent and identically distributed random variables $\tau_{i,l}^{v,n}, l = 1, \ldots$, such that the duration of the lth vacation interval for source i is $\sqrt{n}\tau_{i,l}^{v,n}$. Also, $\tau_{i,l}^{v,n}$ converges weakly to a random variable τ_i^v as $n \to \infty$.*

For each i, the $\tau_{i,l}^{v,n}, l = 1, \ldots,$ are independent of all other "driving" random variables. The intervals for the different sources are mutually independent. $\sup_{i,n} E\tau_{i,l}^{v,n} < \infty$. *Define* $\tau_{i,0}^{v,n} = 0$.

The convergence to the Wiener process in A2.1 is a convenient way of covering many common models, the simplest being the independent and identically distributed cases. The extension of A2.4 which covers correlated (between the sources) vacation intervals is discussed at the end of section 3. The added difficulties are only algebraic. We can work with different information structures, in that the server controller can know either the numbers queued or the work queued. We will confine ourselves to the first case, but it will be seen that the results are asymptotically (as $n \to \infty$) equivalent in that the minimum costs are the same and a good policy for one is equivalent to a good policy for the other. This holds since (Theorem 3.1) the scaled number queued and scaled work queued are (asymptotically) linearly related except on an arbitrarily small (scaled) time interval. Thus, unless mentioned otherwise, the server does not know the work in the queues, only the number of jobs in each queue. It also knows the entire past history, which is the set of past polling decisions, the work done for each job already served and the timing, as well as the starting and ending times of the vacations to date, for each source. The *admissible control* (or polling) policy is defined in the following way.

A2.5. *The server can select the queue served in any nonanticipative way at all, provided that it does not switch while a job is being processed. By nonanticipative we mean the following. Suppose that a job has been completed at real time nt and both sources are available. Then, the next source to be polled is determined by the value* (0 *or* 1) *of a measurable function of the initial queue sizes, all arrival and service data, and the record of vacation starts and completions up to real time nt for each queue.*

Later we will also deal with the set of policies which satisfy either of the following special but important conditions.

A2.6a. (In terms of queued numbers.) *There is a real-valued function $\phi(\cdot)$, which is continuous and nondecreasing, such that between vacations the server polls source* 1 *at real time nt if $x_2^n(t) < \phi(x_1^n(t))$, and polls source* 2 *otherwise. The server does not switch while a job is being processed.*

A2.6b. (In terms of queued workload, a less restrictive function.) *There is a real-valued function $\theta(\cdot)$ which is continuous at all but a finite number of values such that between vacations the server polls source* 1 *if $WL_1^n(t) \geq \theta(WL^n(t))$, and polls source* 2 *otherwise. The server does not switch while a job is being processed.*

3. Weak convergence of the workload and content processes: Arbitrary controls. For $l > 0$, define

$$\nu_{i,l}^n = \sum_{k=1}^{l} \left[\tau_{i,k}^{s,n} + \tau_{i,k-1}^{v,n}/\sqrt{n}\right],$$

which is $1/n$ times the real time of the start of the lth vacation at source i. That is, it is the *scaled* time of the start of the lth vacation at source i. The (scaled) lth vacation interval for source i is the half open (scaled) interval $[\nu_{i,l}^n, \nu_{i,l}^n + \tau_{i,l}^{v,n}/\sqrt{n})$.

Discussion of the control problem. The weak convergence result will be in terms of the total workload (rather than in terms of the workload in each queue), which will be seen to be enough to get the desired results. In fact, except under special policies such as A2.6, in general there is no weak convergence result for the $(WL_i^n(\cdot), i = 1, 2)$ or $(x_i^n(\cdot), i = 1, 2)$. Also, the use of $WL(\cdot)$ yields a one dimensional

limit control problem, a considerable advantage. Next, we introduce some needed notation, and set the problem up and discuss some of its features in a way that facilitates the proof of Theorem 3.1.

Let $S_i^{a,n}(t)$ (resp., $S_i^{d,n}(t)$) denote $1/n$ times the number of jobs that arrived to (resp., were completely served at) queue i by real time nt. Let $Z^n(t)$ denote $1/\sqrt{n}$ times the total real time that both queues are empty *and* neither source is on vacation up to real time nt. Let $T^{v,n}(t)$ denote $1/\sqrt{n}$ times the total time up to real time nt that the server could not work due to a vacation (i.e., where the contents of the available queue, if any, is zero, or where there are no available queues). Then we can write (remaining work = arrived work − work done)

$$WL^n(t) = WL^n(0) + \frac{1}{\sqrt{n}} \sum_i \sum_{l=1}^{nS_i^{a,n}(t)} \Delta_{i,l}^{d,n} - \frac{1}{\sqrt{n}} \left[\text{real time of all service by real time } nt\right], \tag{3.1}$$

where the last term on the right is

$$-\left[\sqrt{n}t - Z^n(t) - T^{v,n}(t)\right]. \tag{3.2}$$

The effect of the vacations: A heuristic discussion. We need to examine $T^{v,n}(t)$ more carefully, since (as will be seen) it is through this term that the control affects the paths of the process in the limit. By A2.4, in scaled time the vacations last $\tau_{i,l}^{v,n}/\sqrt{n}$ units of time, an amount which vanishes as $n \to \infty$.

By A2.3, the intervacation times are $\tau_{i,l}^{v,n}$ in scaled time. Owing to the mutual independence of the intervacation times for the same queue and for the different queues, for any $T > 0$ the probability that vacations will overlap at some point on the scaled time interval $[0, T]$ is of the order of $1/\sqrt{n}$. Since weak convergence on the time interval $[0, \infty)$ is implied by weak convergence on all intervals $[0, T]$, the possibility of overlapping can be ignored in the weak convergence proofs. Thus, in the following discussion, which evaluates the effects of the vacations on the paths for arbitrarily large n, we will suppose (without loss of generality) that only one source can be on vacation at a time. More particularly, if one or more vacations overlap, ignore all but the first. Since this modification alters the paths on each interval $[0, T]$ with a probability of the order $O(1/\sqrt{n})$, it does not affect the distribution of the limit quantities. While the possibility of overlapping vacations is not important for the purely weak convergence aspects in this section, it will have to be taken into account when dealing with the convergence of the costs in section 4. This is because the convergence of the costs (which are not bounded functions) requires both weak convergence of the processes and uniform integrability of the cost functions, so that events of small probability cannot necessarily be neglected. Define $u^n(t) = WL_1^n(t)$. Hence, $WL_2^n(t) = WL^n(t) - u^n(t)$. It will turn out that $u^n(\cdot)$ has the effect of a control. Its value will be seen to be the mechanism for controlling the values of the jumps.

Consider the lth vacation of source i. It starts at (scaled) time $\nu_{i,l}^n$, and the total workload just before that is $WL^n(\nu_{i,l}^n-)$. Define $\bar{A}_{i,l}^{j,n}$ to be $1/\sqrt{n}$ times the work arriving at queue i during the lth vacation of source j. Define

$$\bar{A}_{i,l}^{j,n}(t) = \frac{1}{\sqrt{n}} \sum_{l=nS_i^{a,n}(\nu_{j,l}^n-)+1}^{nS_i^{a,n}((\nu_{j,l}^n+(\tau_{j,l}^{v,n}\wedge t)/\sqrt{n})-)} \Delta_{i,l}^{d,n}. \tag{3.3}$$

The minuses $(-)$ in the lower and upper limits of summation in (3.3) are due to the fact that (recall that the interval is half open) we count $\nu^n_{1,l}$ as part of the (scaled) vacation period, but not $\nu^n_{1,l} + \tau^{v,n}_{1,l}/\sqrt{n}$, for specificity, although the exact accounting procedure used at the end-points is asymptotically (as $n \to \infty$) irrelevant.

The time scale used in (3.3) will be called the *local fluid scale*. In this scale, t denotes an interval of real time of length $\sqrt{n}t$, or, equivalently, an interval in scaled time of length $t/\sqrt{n}$.

Until further notice, for notational simplicity in the motivational discussion, let us fix our attention on the lth vacation of source 1. Thus, $\bar{A}^{1,n}_{i,l} = \bar{A}^{1,n}_{i,l}(\infty)$, the scaled arriving work at queue i during this vacation. Also, $\bar{A}^{1,n}_{i,l}(t)$ is $1/\sqrt{n}$ times the work arriving at queue i in the *real* time interval $[n\nu^n_{1,l}, n\nu^n_{1,l} + \sqrt{n}t)$ for $t < \tau^{v,n}_{1,l}$.

Let $\xi^{v,n}_{i,l}$ denote the change in the *total workload* during the lth vacation of source i. If

$$\tau^{v,n}_{1,l} < WL^n_2(\nu^n_{1,l}-) + \bar{A}^{1,n}_{2,l} = WL^n(\nu^n_{1,l}-) - u^n(\nu^n_{1,l}-) + \bar{A}^{1,n}_{2,l}, \tag{3.4a}$$

then the vacation ends before queue 2 is emptied, and the vacation would not seem to have an immediate effect on the total idle time and workload and (asymptotically, as $n \to \infty$) $\xi^{v,n}_{1,l} = 0$. This is not quite obvious, and the point is both important and subtle, since it is possible that the scaled work that arrives during a vacation all arrives close to the end in which case there might be idle time. However, as seen from Theorem 3.1, it turns out that (asymptotically, as $n \to \infty$ and in the local fluid time scale) the scaled work can be supposed to arrive "continuously" and at the mean rate during the vacation, analogously to a "fluid." This implies that, asymptotically, as $n \to \infty$ and under (3.4a), the vacation has no effect on the total workload.

On the other hand, if

$$\tau^{v,n}_{1,l} > WL^n(\nu^n_{1,l}-) - u^n(\nu^n_{1,l}-) + \bar{A}^{1,n}_{2,l} \equiv \hat{\tau}^{v,n}_{1,l}, \tag{3.4b}$$

then queue 2 is emptied before the vacation ends. The proof of Theorem 3.1 allows us to suppose that (asymptotically, as $n \to \infty$, and in the local fluid time scale) the scaled work arrives continuously, as a fluid, at the mean rate as noted above. However, the heavy traffic condition A2.2 implies that the service rate is so much faster than the arrival rate of work at queue 2, that (asymptotically, as $n \to \infty$) the workload in queue 2 is zero for a nonvanishing fraction of the vacation time. Thus, there is (asymptotically) forced idle time and an increase in the total workload due to the vacation. This increase will depend on the value of the workload at queue 2 at the time that the vacation at queue 1 starts. In turn, that value depends on the control policy. This is the *only* way that the control policy affects the workload: via the sizes of the jumps due to the vacations, which (in turn) is determined by the distribution of the total workload just before the vacation starts. It will be seen that the difference between the scaled work that arrives at queue j during this lth vacation of source 1 and $\bar{\lambda}^a_j \bar{\Delta}^d_j \tau^{v,n}_{1,l} = \rho_j \tau^{v,n}_{1,l}$ converges weakly to zero. Obviously, one can reverse sources 1 and 2 in the above discussion. From the above discussion, we see that the control can be viewed as the division of the total workload among the two queues.

Suppose, formally, that n indexes a weakly convergent subsequence of

$$\left(\tau^{v,n}_{1,l}, u^n(\nu^n_{1,l}-), WL^n(\nu^n_{1,l}-)\right),$$

use the same notation (dropping the n superscript) for the weak sense limits, and formally use the asymptotic fluid approximation to (3.3). This fluid approximation

simply replaces the terms in (3.3) by their asymptotic mean value and is $\rho_i[\tau^v_{1,l} \wedge t]$. We see, formally, that (in the limit) the increase in $T^{v,n}(\cdot)$ (equivalently, in the total workload) during the lth vacation of source 1 can be written as

$$\xi^v_{1,l} = \left[(1-\rho_2)\,\tau^v_{1,l} - [WL(\nu_{1,l}-) - u(\nu_{1,l}-)]\right]^+$$

which equals

$$\left[\rho_1\tau^v_{1,l} - [WL(\nu_{1,l}-) - u(\nu_{1,l}-)]\right]^+. \tag{3.5a}$$

The analogue for the lth vacation of source 2 is

$$\xi^v_{2,l} = \left[\rho_2\tau^v_{2,l} - u(\nu_{2,l}-)\right]^+. \tag{3.5b}$$

Here, $u(t) \le WL(t)$ can be considered to be the control function for the limit system (3.7), (3.8a). It appears only via the discrete values: $u(\nu_{i,l}-), l < \infty, i = 1, 2$. The notation in (3.5) can be misleading since the use of the symbol $u(t-)$ suggests either left continuity or that the left-hand limit exists at t, or that (some subsequence of) $u^n(\cdot)$ converges weakly. *We do not make these claims for general polling policies,* but the sequence $u^n(\nu^n_{i,l}-)$ will always be tight in n. If the control policy is of the type in (A2.6), then roughly speaking (see Theorem 3.2) the "intervacation sections" of $u^n(\cdot)$ will be tight and have continuous weak sense limits.

The intervacation sections. Let $\bar\nu^n_l$ and ν^n_l denote, respectively, the (scaled) time of the beginning and end (respectively) of the lth intervacation interval, irrespective of the source. Thus, by our conventions, $\bar\nu^n_l$ is $1/n$ times the real time of the beginning of the lth vacation. By our convention, no source is on vacation at the initial time, so that $\bar\nu^n_1 = 0$. Define the *intervacation sections* as the functions

$$WL^n\left((\bar\nu^n_l + t) \wedge \nu^n_l\right),\ t \ge 0. \tag{3.6}$$

It is constant for $t \ge \nu^n_l - \bar\nu^n_l$.

State space collapse. The physical dimension of the original problem is the number of sources. Mathematically, the dimension is even higher since the set of queue lengths is not Markovian. Theorem 3.1 is an example of what is called state space collapse [7, 24, 25, 26, 32] in the heavy traffic literature. The dimension of the approximating problem is unity and the original $x^n_i(\cdot)$ (which do not necessarily converge weakly) can be asymptotically approximated by a constant (depending on i) times the total workload process $WL^n(\cdot)$ (which does converge weakly in the sense described in the theorem). Such state space collapse is obviously very helpful in the control problem and for numerical procedures.

Comment on tightness and Theorem 3.1. During a vacation, $WL^n(\cdot)$ changes in steps of size $O(1/\sqrt{n})$ over an interval of scaled size $O(1/\sqrt{n})$. While the effect of the vacation is (asymptotically) a jump in a well-defined sense, because of the way that the jump is realized, $WL^n(\cdot)$ is not tight in the Skorohod topology. Since the parts of the path between vacations are well behaved, it is convenient to work with the effects of the vacations and the intervacation parts separately. Suppose that the set in (3.6) is tight for each l. It is easy to show this (Theorem 3.1) for $l = 1$. Then the set of its "terminal" conditions $WL^n(\nu^n_1-)$ is also tight, as is $u^n(\nu^n_l-)$. Thus the set (3.5a) or (3.5b) for the first vacation is also tight (Theorem 3.1). Then, the set of initial conditions $WL^n(\bar\nu_2)$ for the next intervacation interval is tight. Then repeat, as for the first section, etc. In this way, taking an appropriate subsequence and working

section by section, one constructs a "limit" process $WL(\cdot)$. We call this procedure "concatenation." A primary aim is showing the convergence of the discounted cost functions (4.3b) to (4.3a) for a well-defined "limit" process $WL(\cdot)$. One does not need full weak convergence of $WL^n(\cdot)$ for this and the piecewise or concatenation approach is adequate. The various conclusions of the theorem are denoted by (a), (b), etc.

As indicated above, in the proof of the theorem one works step by step. After some preliminary details concerning convergence of the $S_i^{a,n}(\cdot)$ and representations of the workload process, it is shown that the sequence of sections up to the time of the first vacation is tight and its limit is characterized. Thus, the sequence of states at the time at which the first vacation starts is tight. Then we deal with the first vacation and characterize its limits. Now, we have that the sequence of states at the end of the first vacation is tight, so we can analyze the paths between the end of the first vacation and the beginning of the second, just as the path up to the first vacation was handled, etc. In this way, we can see that there is nothing special about the first intervacation interval or the first vacation. Thus, all of the intervacation sections and vacation jumps can be dealt with. The appropriate limit process puts these together in sequence. This type of convergence is sufficient to get the convergence result for the discounted cost function later on. The procedure is analogous to a common method of constructing the solution to a jump-diffusion process.

THEOREM 3.1. *Assume* A2.0–A2.5*, and suppose that* $(x_i^n(0), i = 1, 2)$ *converges weakly to* $(x_i(0), i = 1, 2)$. (a) *Then* $WL^n(0)$ *converges weakly to* $\sum_i \bar{\Delta}_i^d x_i(0)$. (b) *For* $i = 1, 2$, *the set*

$$\Psi^n = \Big(WL^n(\nu_{i,l}^n-), u^n(\nu_{i,l}^n-), \tau_{i,l}^{v,n}, \tau_{i,l}^{s,n}, \xi_{i,l}^{v,n}, i = 1, 2, l < \infty\Big)$$

is tight in n. (c) *The sequence of intervacation sections of* $WL^n(\cdot)$ *defined by* (3.6) *and of* $Z^n(\cdot)$ *are tight for each* i *and* l, *the weak sense limit of any weakly convergent subsequence has continuous paths.*

Fix a weakly convergent subsequence of the set Ψ^n *and the set of intervacation sections of* $WL^n(\cdot)$ *and* $Z^n(\cdot)$, *and index it by* n *also* (*abusing terminology*). *The weak sense limits are denoted by dropping the* n *superscript.* (d) *Then* $(\tau_{i,l}^{s,n}, l < \infty)$ *converges weakly to* $(\tau_{i,l}^s, l < \infty)$, *where the* $\tau_{i,l}^s$ *are exponentially distributed with rate* $\bar{\lambda}_i^s$. (e) *The differences* $WL_i^n(\cdot) - \bar{\Delta}_i^{d,n} x_i^n(\cdot)$ *converge weakly to the "zero" process.* (f) *Define* $WL(\cdot)$ *by concatenating the weak sense limits of the successive intervacation sections of* $WL^n(\cdot)$. *The weak sense limits of any weakly convergent subsequence are related by*

$$WL(t) = WL(0) + bt + w(t) + \sum_i J_i(t) + Z(t), \tag{3.7}$$

where

$$J_i(t) = \sum_{l:\nu_{i,l} \le t} \xi_{i,l}^v, \quad \nu_{i,l} = \sum_{k=1}^{l} \tau_{i,k}^s. \tag{3.8a}$$

In (3.7) *the process* $WL(\cdot)$ *between its* $(l-1)$*st and* l*th jump is the value of the weak sense limit of the process defined by* (3.6) *on the interval* $[0, \nu_l^n - \bar{\nu}_l^n)$. (g) *Also*

$$\big(WL(0), w(\cdot), \tau_{i,l}^v, \tau_{i,l}^s; i = 1, 2, l < \infty\big) \tag{3.8b}$$

are mutually independent, and $w(\cdot)$ is a Wiener process with variance parameter

$$\sigma^2 = \sum_i \left[\rho_i \sigma_{a,i}^2 + \sigma_{d,i}^2\right], \tag{3.9}$$

which we assume to be positive. (h) *$Z(\cdot)$ is the reflection term. It is continuous, nondecreasing, can increase only at t, where $WL(t) = 0$ and satisfies $Z(0) = 0$.* (i) *The $\xi_{i,l}^v$ have the representation* (3.5). (j) *Define the Poisson processes $N_i(\cdot), i = 1, 2$, to be the process with a unit jump at $\nu_{i,l}, l \geq 1$. For each t,*

$$w(t+\cdot) - w(t), N_i(t+\cdot) - N_i(t),\ i = 1, 2, \tag{3.10}$$

is independent of

$$\begin{aligned} &w(s), N_i(s), s \leq t; s \leq t; i = 1, 2, \\ &\quad \left(u(\nu_{i,l}-) I_{\{\nu_{i,l} \leq t\}}, \xi_{i,l}^v I_{\{\nu_{i,l} \leq t\}}, i = 1, 2, l < \infty\right). \end{aligned} \tag{3.11}$$

(k) *The process* (3.3) *converges weakly to the process with values $\rho_i(t \wedge \tau_{j,l}^v)$.*

Comment on the control $u(\nu_{i,l}-)$. Under the general conditions that are used in this theorem to get the weak convergence, we cannot get convergence of the random processes $u^n(\cdot)$, only of the random variables which are the values at selected points. However, in the next section the class of polling policies will be restricted to be in some very reasonable class, and for this class there will be tightness of $u^n(\cdot)$ in an appropriate sense. Then, the weak sense limits will be well-defined admissible control functions for the weak sense limit $WL(\cdot)$ process.

Proof. As noted below (3.2), without loss of generality we can suppose that at most one source is on vacation at a time. Given the current real time nt, the real time since the current service started or has to go, or the real time since or until the next arrival are called *residual times*. We define a *residual time error term* to be a random process (to be denoted by $\epsilon^n(\cdot)$) which is bounded by [constant/$\sqrt{n}$] times a [finite sum of such residual time terms plus a constant]. Successive uses of $\epsilon^n(\cdot)$ might refer to different residual time error terms. Assumption A2.1 implies that the $\epsilon^n(\cdot)$ converge weakly to the "zero" process, since the continuity of the limit there implies that the maximum of the first $[nt]$ summands, divided by $\sqrt{n}$, goes to zero in probability as $n \to \infty$.

The difference between the terms in (2.3) is a residual time error term, thus the process defined by the difference converges weakly to the "zero" process. The proof that $WL^n(0)$ converges as asserted follows from A2.1 and the representation (2.3), where $t = 0$.

For specificity, we will suppose that the preempt-resume discipline holds for any job which is being served when a vacation of its source starts. Assumption A2.3 implies that, for any t, the number of vacations on any real time interval $[0, nt]$ is bounded in probability, uniformly in n. Due to this and the fact that (by A2.1) the maximum of the first $[nt]$ workloads divided by $\sqrt{n}$ goes to zero in probability as $n \to \infty$, any of the disciplines cited in section 2 will yield the same result.

We next prove the weak convergence of $S_i^{a,n}(\cdot)$ to the process with constant values $\bar{\lambda}_i^a t$. Define $\mathcal{T}_i^{a,n}(t) = \sum_{l=1}^{nt} \Delta_{i,l}^{a,n}/n$. By A2.1, $\mathcal{T}_i^{a,n}(\cdot)$ converges weakly to the process with values $\bar{\Delta}_i^a t = t/\bar{\lambda}_i^a$. Also, possibly modulo a residual time error term,

$$\begin{aligned} S_i^{a,n}(\mathcal{T}_i^{a,n}(t)) &= t, \\ \mathcal{T}_i^{a,n}(S_i^{a,n}(t)) &= t. \end{aligned}$$

This and the weak convergence of $\mathcal{T}_i^{a,n}(\cdot)$ imply the asserted weak convergence of $S_i^{a,n}(\cdot)$.

The next step is to show the tightness and asymptotic continuity of $WL^n(\cdot)$ *when there are no vacations.* In the absence of vacations, we can write

$$WL^n(t) = WL^n(0) + \frac{1}{\sqrt{n}} \sum_i \sum_{l=1}^{nS_i^{a,n}(t)} \Delta_{i,l}^{d,n} - t\sqrt{n} + Z^n(t). \tag{3.12}$$

The term $T^{v,n}(t)$ of (3.2) is not included since, in this part of the proof, we have assumed that there are no vacations. For each i, expand the inner sum in (3.12) as

$$\frac{1}{\sqrt{n}} \sum_{l=1}^{nS_i^{a,n}(t)} \left[\Delta_{i,l}^{d,n} - \bar{\Delta}_i^{d,n}\right] + \frac{1}{\sqrt{n}} \sum_{l=1}^{nS_i^{a,n}(t)} \bar{\Delta}_i^{d,n}. \tag{3.13}$$

The first term of (3.13) is $-\bar{\Delta}_i^{d,n} w_i^{d,n}(S_i^{a,n}(t))$. Expand the last term in (3.13) as

$$\frac{1}{\sqrt{n}} \bar{\Delta}_i^{d,n} \sum_{l=1}^{nS_i^{a,n}(t)} \left[1 - \frac{\Delta_{i,l}^{a,n}}{\bar{\Delta}_i^{a,n}}\right] + \frac{1}{\sqrt{n}} \frac{\bar{\Delta}_i^{d,n}}{\bar{\Delta}_i^{a,n}} \sum_{l=1}^{nS_i^{a,n}(t)} \Delta_{i,l}^{a,n}. \tag{3.14}$$

The right-hand *sum* in (3.14) equals nt minus the time between nt and the last arrival at or before real time nt. Hence, the right-hand term equals $\rho_i^n \sqrt{n} t$ plus a residual time error term.

Summarizing,

$$\begin{aligned} WL^n(t) = WL^n(0) &+ \sum_i \bar{\Delta}_i^{d,n} \left[w_i^{a,n}(S_i^{a,n}(t)) - w_i^{d,n}(S_i^{a,n}(t))\right] \\ &+ \sqrt{n}\left[\sum_i \rho_i^n - 1\right] t + Z^n(t) + \epsilon^n(t). \end{aligned} \tag{3.15}$$

The hypotheses and the weak convergence of $S_i^{a,n}(\cdot)$ imply the weak convergence of the processes on the right of (3.15) to those of (3.7), except possibly that of $Z^n(\cdot)$, with the given definitions of $w(\cdot)$ and b, but without the jump term. If $\{Z^n(\cdot), n < \infty\}$ were not tight and have continuous weak sense limits, then we would have a contradiction to the facts that $Z^n(\cdot)$ can increase only when $WL^n(t) = 0$ and has jumps of size $1/\sqrt{n}$ only. Thus, by taking a further subsequence if necessary, we can suppose that $Z^n(\cdot)$ converges to the reflection term $Z(\cdot)$ in (3.7). The sequence of processes defined by (3.15) converges weakly and the limit satisfies (3.7) without the jump term.

Now, return to the *original problem, where there are vacations,* and recall that ν_1^n is the scaled time of the start of the first vacation. For the remainder of this proof, continue using the assumption that the vacations do not overlap. As noted below (3.2), this is accomplished by ignoring all but the first if there are overlaps. The alteration does not change the distribution of the limit processes, since the probability of such a change on any finite interval goes to zero as $n \to \infty$.

By A2.3 and A2.4, the various sequences of times $\tau_{i,l}^{s,n}, \tau_{i,l}^{v,n}, i = 1, 2, l = 1, \ldots,$ converge weakly and $\bar{\nu}_{l+1}^n - \nu_l^n = \tau_{i,l}^{v,n}/\sqrt{n}$ converges weakly to zero. By A2.3 and the weak convergence of the sequence of processes defined by (3.15), $WL^n(\nu_1^n -)$ also converges weakly. Denote the weak sense limits by dropping the superscript n. By A2.3,

the $(\tau^s_{i,l}, l = 1, \ldots, i = 1, 2)$ are mutually independent, and exponentially distributed, with rate $\bar\lambda^s_i$ for $\tau^s_{i,l}$. By A2.4, $(\tau^v_{i,l}, l = 1, \ldots, i = 1, 2)$ are mutually independent. By A2.0, A2.1, A2.3, and A2.4, the

$$WL(0), w^a_i(\cdot), w^d_i(\cdot), \tau^v_{i,l}, \tau^s_{i,l}, i = 1, 2,\ l = 1, \ldots, \tag{3.16}$$

are mutually independent. For each l, ν^n_l converges weakly and $\nu^n_{l+1} - \nu^n_l$ converges weakly to an exponentially distributed random variable, with rate $\sum_i \bar\lambda^s_i$, and the limits are mutually independent and are independent of the random variables in (3.16) other than $\{\tau^s_{i,l}; i, l\}$.

Suppose for the moment that the jumps $\xi^{v,n}_{i,l}$ are tight for each i, l. Abusing notation, let n index a further subsequence along which all of the $\xi^{v,n}_{i,l}, i = 1, 2, l = 1, \ldots,$ also converge weakly, and denote the weak sense limits by dropping the superscript n. Then, by repeating the analysis which led to (3.15) on each successive intervacation interval, we are led to (3.7) with $J_i(\cdot)$ defined by (3.8a) and the independence in (3.8b). Equation (3.7) represents the limit of $WL^n(\cdot)$ in the particular sense that its interjump sections are the weak sense limits of the intervacation sections (3.6) and its jumps are the limits of the $WL^n(\bar\nu^n_{l+1}) - WL^n(\nu^n_l-)$ (for the chosen subsequence).

By taking a further subsequence, if necessary, we can also suppose that, together with the other convergences, $u^n(\nu^n_{i,l}-), i = 1, 2, l = 1, \ldots,$ converges weakly to random variables which we call $u(\nu_{i,l}-), i = 1, 2, l = 1, \ldots$. From the weak convergence of the $u^n(\nu^n_{i,l}-), \nu^n_{i,l}, i = 1, 2, l = 1, \ldots,$ we have the weak convergence of the $u^n(\nu^n_l-), l = 1, \ldots$.

We will next show that $\xi^{v,n}_{i,l}$ is tight for each i and l and that (3.5) characterizes the weak sense limits. To simplify the notation, we will start with the first vacation and let the first vacation be that of source 1. With this simplifying assumption, we can write $\tau^{s,n}_{1,1} = \nu^n_{1,1} = \nu^n_1$, and we will use these variables (and their weak sense limits) interchangably. By the weak convergence of $WL^n(\cdot)$ (with the weak sense limit of $WL^n(\cdot)$ being continuous) when there are no vacations and the weak convergence of $\tau^{s,n}_{1,1}$, $WL^n_1(\tau^{s,n}_{1,1}-) = WL^n_1(\nu^n_1-) = u^n_1(\nu^n_1-)$ converges weakly to the random variable which we denote by $u(\nu^s_{1,1}-)$. We will show that, under (3.4a),

$$WL^n(\nu^n_{1,1} + \tau^{v,n}_{1,1}/\sqrt{n}) - WL^n(\nu^n_{1,1}) \Rightarrow 0, \tag{3.17}$$

and under (3.4b),

$$WL^n_2(\nu^n_{1,1} + \tau^{v,n}_{1,1}/\sqrt{n}) \Rightarrow 0, \tag{3.18}$$

$$WL^n_1(\nu^n_{1,1} + \tau^{v,n}_{1,1}/\sqrt{n}) - WL^n_1(\nu^n_{1,1}) \text{ is tight}, \tag{3.19}$$

and the weak sense limit (along the selected weakly convergent subsequence) of (3.19) is defined by (3.5). Since the (scaled) vacation interval in question is $[\nu^n_{1,1}, \nu^n_{1,1} + \tau^{v,n}_{1,1}/\sqrt{n})$, strictly speaking, the arguments in the functions in (3.19) should be $(\nu^n_{1,1} + \tau^{v,n}_{1,1}/\sqrt{n})-$, and analogously for (3.17) and (3.18). However, since the processes defined by $WL^n(t) - WL^n(t-)$ and $u^n(t) - u^n(t-)$ converge weakly to the "zero" process, one can always replace $t-$ by t without changing any of the weak sense limits. We will do this to simplify the notation.

Since source 2 is being polled during this vacation, for $t \le \tau^{v,n}_{1,1}$ we can write

$$WL^n_2(\nu^n_{1,1} + t/\sqrt{n}) = \left[WL^n(\nu^n_{1,1}) - u^n(\nu^n_{1,1})\right] - t + T^{v,n}(\nu^n_{1,1} + t/\sqrt{n}) + \bar A^{1,n}_{2,1}(t). \tag{3.20}$$

We will next show that the process defined by

$$\bar{A}_{i,1}^{1,n}(t \wedge \tau_{1,1}^{v,n}) - \rho_i^n(t \wedge \tau_{1,1}^{v,n})$$

converges weakly to the "zero" process. This is what was meant by the statement below (3.4a) to the effect that work can be assumed to arrive continuously during a vacation and it is the last assertion (k) of the theorem. Note that the local fluid time scale defined below (3.3) is used in (3.20), so that t denotes an interval of length $\sqrt{n}t$ in real time or $t/\sqrt{n}$ in scaled time.

Use the representation (3.3) (dropping the $-$ in the indices of summation without changing the end result) to write $\bar{A}_{i,1}^{1,n}(t)$ as

$$\begin{aligned}\frac{1}{\sqrt{n}} \sum_{l=nS_i^{a,n}(\nu_{1,1}^n)+1}^{nS_i^{a,n}(\nu_{1,1}^n+(t\wedge\tau_{1,1}^{v,n})/\sqrt{n})} \left[\Delta_{i,l}^{d,n} - \bar{\Delta}_i^{d,n}\right] \\ + \frac{1}{\sqrt{n}} \sum_{l=nS_i^{a,n}(\nu_{1,1}^n)+1}^{nS_i^{a,n}(\nu_{1,l}^n+(t\wedge\tau_{1,1}^{v,n})/\sqrt{n})} \bar{\Delta}_i^{d,n}.\end{aligned} \tag{3.21}$$

The first term in (3.21) goes weakly to zero by the tightness of $\nu_{i,l}^n, \tau_{i,l}^{v,n}$ in n, the weak convergence of $S_i^{a,n}(\cdot)$, and condition A2.1. We need to characterize the right-hand term of (3.21). Write it as

$$\begin{aligned}\frac{\bar{\Delta}_i^{d,n}}{\sqrt{n}} \sum_{l=nS_i^{a,n}(\nu_{1,1}^n)+1}^{nS_i^{a,n}(\nu_{1,1}^n+(t\wedge\tau_{1,1}^{v,n})/\sqrt{n})} \left[1 - \frac{\Delta_{i,l}^{a,n}}{\bar{\Delta}_i^{a,n}}\right] \\ + \frac{\bar{\Delta}_i^{d,n}}{\bar{\Delta}_i^{a,n}\sqrt{n}} \sum_{l=nS_i^{a,n}(\nu_{1,1}^n)+1}^{nS_i^{a,n}(\nu_{1,1}^n+(t\wedge\tau_{1,1}^{v,n})/\sqrt{n})} \Delta_{i,l}^{a,n}.\end{aligned}$$

Just as for the first term in (3.21), the first term in the above expression goes weakly to the "zero" process as $n \to \infty$. The real time difference between the arguments in the upper and lower indices in the last expression is $\sqrt{n}[t \wedge \tau_{1,1}^{v,n}]$. Hence, the sum in the second term times $1/\sqrt{n}$ is $t \wedge \tau_{1,1}^{v,n}$, modulo a residual time error term. Thus, the difference between the second term and

$$\frac{\bar{\Delta}_i^d}{\bar{\Delta}_i^a}\left(t \wedge \tau_{1,1}^{v,n}\right) = \rho_i\left(t \wedge \tau_{1,1}^{v,n}\right) \tag{3.22}$$

converges weakly to the "zero" process.

The above computations concerning the arriving scaled work during the vacation show that the net change $\xi_{1,1}^{v,n}$ in the total workload is tight, and that we can suppose, asymptotically and in the local fluid time scale defined below (3.3), that scaled work arrives at the queues continuously (i.e., as a fluid process) at the mean rate ρ_i during the vacation.

Consider the case (3.4a). By what has just been proved, the scaled work process arriving to queue 1 during the first vacation is arbitrarily well approximated (in the local fluid time scale) by (3.22) for $i = 1$. Similarly, the scaled work that departs queue 2 during that time (local fluid time scale) is (asymptotically) equal to $\tau_{1,1}^{v,n}$ minus the idle time in the local fluid time scale. However, due to the fluid approximation

and the condition (3.4a), this idle time is zero, asymptotically. Thus, by the above computation, the net increase in the workload (input minus output) of queue 2 during the vacation is (asymptotically) equal to $[\rho_2 - 1]\tau_{1,1}^{v,n}$. Thus, by the heavy traffic condition A2.2, adding the changes in the two queues, we see that the net change in the total workload during the vacation converges weakly to zero as $n \to \infty$.

Now, consider the condition (3.4b), continue to use the approximation (3.22), and recall the definition of $\hat{\tau}_{i,l}^{v,n}$ from (3.4b). It is then clear that $\tau_{i,l}^{v,n} - \hat{\tau}_{i,l}^{v,n}$ is the net contribution to $T^{v,n}(\cdot)$ during this vacation interval, and (3.5) follows from this and the results of the last paragraph. Let $\xi_{1,1}^{v}$ denote the weak sense limit of $\xi_{1,1}^{v,n}$. Thus, we have obtained (3.7) and (3.8) and verified (i) up to and including the time of the first vacation.

Now, with $WL_i^n(\nu_{1,1}^n + \tau_{1,1}^{v,n}/\sqrt{n})$ well defined and tight, restart the $WL_i^n(\cdot)$ at scaled time $\nu_{1,1}^n + \tau_{1,1}^{v,n}/\sqrt{n}$ and repeat the above approximation and limit procedure. Then an induction argument yields the asserted limit relations for all of the intervacation sections and jumps. Take a weakly convergent subsequence of Ψ^n in the theorem statement, and obtain (3.7) and (3.8) by concatenating the intervacation sections.

Next, we will prove the (asymptotic) linear relationship (e) between $WL_i^n(\cdot)$ and $x_i^n(\cdot)$. We say that a sequence of real-valued processes $q^n(\cdot)$ is *bounded in probability* if, for each $T > 0$,

$$\lim_{N\to\infty} \limsup_n P\left\{\sup_{t\le T} |q^n(s)| \ge N\right\} = 0. \tag{3.23}$$

It will be seen that the tightness of $WL^n(\cdot)$ implies that $x_i^n(\cdot)$ satisfies (3.23). Let us assume this for the moment. We will use the representation (2.3). Write $\Delta_{i,l}^{d,n} = \bar{\Delta}_{i,l}^{d,n} + [\Delta_{i,l}^{d,n} - \bar{\Delta}_{i,l}^{d,n}]$. With this representation, expand each of the two terms in the brackets in (2.3) into two components, analogous to what was done to get the expression below (3.21). The first component of the expansion of (say) the first term inside the brackets in (2.3) is

$$\bar{\Delta}_i^{d,n} x_i^n(t). \tag{3.24}$$

The second component of (again) the first term inside the brackets of (2.3) is

$$\frac{1}{\sqrt{n}} \sum_{l=L_i^n(t)}^{L_i^n(t)+\sqrt{n}x_i^n(t)-1} \left[\Delta_{i,l}^{d,n} - \bar{\Delta}_{i,l}^{d,n}\right], \tag{3.25}$$

and it converges weakly to the "zero" process by the weak convergence assumption A2.1 on $w_i^{d,n}(\cdot)$ since $x_i^n(\cdot)$ is assumed to satisfy (3.23).

Note that the difference between the two terms inside the brackets of (2.3) is a residual time error term $\epsilon^n(t)$, where $\epsilon^n(\cdot)$ converges weakly to the "zero" process, and this is true irrespective of whether or not (3.23) holds for $x_i^n(\cdot)$.

By the tightness of the sections of $WL^n(\cdot)$ between the $(l-1)$st and lth vacations (for each l) and of the associated set of jumps, (3.23) holds for $WL_i^n(\cdot)$, hence it holds for the process defined by the first term in the brackets in (2.3). Suppose that (3.23) does not necessarily hold for $x_i^n(\cdot)$. The expansion of the first term in (2.3) into (3.24) and (3.25) still holds. By the assumption A2.1 on the $w_i^{d,n}(\cdot)$, the fact that the upper index of summation in (3.25) is no bigger than $nS_i^{a,n}(\cdot) + \sqrt{n}x_i^n(0)$ and the weak convergence of $S_i^{a,n}(\cdot)$ and of $x_i^n(0)$, the process defined by (3.25) satisfies (3.23).

Then, since the sum of (3.24) and (3.25) is the first term in (2.3), which satisfies (3.23), so must the process defined by (3.24). Hence $x_i^n(\cdot)$ satisfies (3.23). The fact that (3.23) holds for $x_i^n(\cdot)$ and the assumption A2.1 imply that (3.25) converges weakly to the "zero" process. Hence, $WL_i^n(\cdot)$ is asymptotically equivalent to $\bar{\Delta}_i^{d,n} x_i^n(\cdot)$.

Only the nonanticipativity (j) needs to be proved. But this is a consequence of the independence in (3.16) □

More than 2 sources. Suppose that there is an arbitrary number of sources, with the natural extensions of the assumptions A2.0–A2.5 and the notation holding. Then, on any (scaled) interval $[0, T]$, with a probability that goes to one as $n \to \infty$, there is still only a finite number of vacations and at most one source can be on vacation at a time. Because of this, the analogues of (3.4) and (3.5) can easily be written.

Consider the lth vacation of source i. If

$$\tau_{i,l}^{v,n} < \left[WL^n(\nu_{i,l}^n-) - WL_i^n(\nu_{i,l}^n-)\right] + \sum_{j\neq i} \bar{A}_{j,l}^{i,n},$$

then the method of Theorem 3.1 can be used to show that (asymptotically) the vacation at source i ends before the other queues are emptied, and the vacation has no immediate effect on the total workload. On the other hand, if

$$\tau_{i,l}^{v,n} > \left[WL^n(\nu_{i,l}^n-) - WL_i^n(\nu_{i,l}^n-)\right] + \sum_{j\neq i} \bar{A}_{j,l}^{i,n} \equiv \hat{\tau}_{i,l}^{v,n},$$

then there is (asymptotically) a forced idle time during the vacation. Dropping the n superscripts, the increase in scaled work during the lth vacation of source i has the asymptotic form

$$\xi_{i,l}^v = \left[\tau_{i,l}^v - [WL(\nu_{i,l}-) - WL_i(\nu_{i,l}-)] - \tau_{i,l}^v \sum_{j\neq i} \rho_j\right]^+$$

which equals

$$\left[\rho_i \tau_{i,l}^v - [WL(\nu_{i,l}-) - WL_i(\nu_{i,l}-)]\right]^+,$$

where $WL_i(\nu_{i,l}-)$ is the weak sense limit of (a suitable weakly convergent subsequence of) $WL_i^n(\nu_{i,l}^n-)$. This is the only change in Theorem 3.1.

The control form A2.6a. The control form specified by A2.6a seeks to force the relationship (asymptotic) $x_2^n(t) \sim \phi(x_1^n(t))$, at least when possible between vacations, when both sources are available. The control, in practice, might be based on either the queue sizes or on the workloads, depending on what information is available to the controller at the server. Theorem 3.1 implies that we can do either, due to the asymptotic equivalence $WL_i^n(\cdot) \sim \bar{\Delta}_i^{d,n} x_i^n(\cdot)$. To a control represented by the function $\phi(\cdot)$ in A2.6a, there is one in terms of the workload in the sense of asymptotic equivalence. We will now see that A2.6a is asymptotically equivalent to the existence of a continuous and nondecreasing function $\theta(\cdot)$ such that we poll source 1 if $WL_1^n(t) \geq \theta(WL^n(t))$ and poll source 2 otherwise, provided that the source is available.

To get $\theta(\cdot)$, we use the asymptotic equivalence $x_2^n(t) \sim \phi(WL_1^n(t)/\bar{\Delta}_1^d)$ and

$$WL_2^n(t) \sim \bar{\Delta}_2^d \phi(WL_1^n(t)/\bar{\Delta}_1^d).$$

Since, asymptotically, using the policy $\phi(\cdot)$ between vacations,

$$WL^n(t) \sim \sum_i WL_i^n(t) = WL_1^n(t) + \bar{\Delta}_2^d \phi(WL_1^n(t)/\bar{\Delta}_1^d),$$

we can define an "asymptotic inverse" $\theta(\cdot)$ to the function $\phi(\cdot)$ in that (between vacations) we poll source 1 if $WL_1^n(t) \geq \theta(WL^n(t))$ and poll source 2 otherwise. The inverse is obtained from

$$WL - WL_1 = \bar{\Delta}_2^d \phi(WL_1/\bar{\Delta}_1^d).$$

Note that we could have started with $\theta(\cdot)$ and derived $\phi(\cdot)$ from it: i.e., suppose that we are given a nonnegative, continuous, and nondecreasing function $\theta(\cdot)$ satisfying $\theta(WL) \leq WL$, and use the following rule: between vacations, poll source 1 if $WL_1^n(t) \geq \theta(WL^n(t))$ and poll source 2 otherwise. This can be turned into an (asymptotic) rule based on the $x_i^n(\cdot)$ by polling source 1 if

$$x_1^n(t)\bar{\Delta}_1^d \geq \theta(x_1^n(t)\bar{\Delta}_1^d + x_2^n(t)\bar{\Delta}_2^d)$$

and polling source 2 otherwise.

The function $\phi(\cdot)$ in terms of the numbers queued is often (but certainly not always) the more pertinent in applications. However, the dynamic programming equation will be in terms of the total workload and the system (3.7), since the basic weak convergence result is in terms of the total workload. The total workload formulation is also much more convenient from the computational point of view due to the "state space collapse" for which, no matter how many sources there are, the problem is one dimensional. Thus, it is important to be able to travel back and forth between the queued number and total workload forms.

Note on realizing the relationship $WL_1^n(t) \sim \theta(WL^n(t))$, or its equivalent in terms of the number queued, under A2.6. Suppose that both sources are available at scaled time t, and that we wish to change $WL_1^n(\cdot)$ to the value $WL_1^{*,n} > WL_1^n(t)$ as quickly as possible. In heavy traffic, by polling queue 2, the scaled queue of source 1 increases at a mean rate of $\bar{\lambda}_1^a \bar{\Delta}_1^d \sqrt{n}$ in scaled time. Thus, it takes approximately $[WL_1^{*,n} - WL_1^n(t)]/[\bar{\lambda}_1^a \bar{\Delta}_1^d \sqrt{n}]$ units of scaled time for the transition. Thus, in the heavy traffic limit, with neither source on vacation, any desired change can be realized instantaneously.

The relationship $WL_1^n(t) \sim \theta(WL^n(t))$ cannot be realized arbitrarily well, uniformly (for large n) on the entire interval between vacations. This is because the uncontrollable changes in the $WL_i^n(\cdot)$ during a vacation will cause it to be violated for a short interval just after the vacation ends, while we "catch up." But there are $\epsilon_n \to 0$ as $n \to \infty$ such that the sections of the differences

$$x_1^n(\cdot) - \frac{\theta(WL^n(\cdot))}{\bar{\Delta}_1^d} \tag{3.26a}$$

and

$$x_2^n(\cdot) - \frac{WL^n(\cdot) - \theta(WL^n(\cdot))}{\bar{\Delta}_2^d} \tag{3.26b}$$

starting (scaled time) ϵ_n after a vacation *begins* and stopping at the start of the next vacation converge to the zero process as $n \to \infty$. This will be sufficient for our purposes.

Using the control $\theta(\cdot)$ in A2.6b, we can write the $\xi^v_{i,l}$ of (3.5) as

$$\xi^v_{1,l} \equiv \left[(1-\rho_2)\,\tau^v_{1,l} - [WL(\nu_{1,l}-) - \theta(WL(\nu_{1,l}-))]\right]^+ \tag{3.27a}$$

and

$$\xi^v_{2,l} \equiv \left[(1-\rho_1)\,\tau^v_{i,l} - \theta(WL(\nu_{2,l}-))\right]^+. \tag{3.27b}$$

The following theorem codifies the last part of the above discussion and the proof follows from the computations done in Theorem 3.1. The last sentence of the theorem holds because of the weak convergence of the intervacation sections and the fact that, between vacations, $WL(\cdot)$ behaves like a Wiener process, provided that $\sigma^2_{\alpha,i} > 0$ for some α, i.

THEOREM 3.2. *Assume the conditions of Theorem* 3.1 *and* A2.6a *as well. Then there are positive real numbers* $\epsilon_n \to 0$ *such that* $x^n_2(\cdot) - \phi(x^n_1(\cdot))$ *converges weakly to the "zero" process on each interval* $[\nu^n_{i,l} + \epsilon_n, \nu^n_{i,l+1}]$. *So do* $WL^n_1(\cdot) - \theta(WL^n(\cdot))$ *and the processes defined in* (3.26), *where* $\theta(\cdot)$ *is defined from* $\phi(\cdot)$ *as above the theorem.*

Now assume A2.6b *in lieu of* A2.6a. *Then, excluding an arbitrarily small neighborhood of the times where* $WL^n(t)$ *is a point of discontinuity of* $\theta(\cdot)$, *the last sentence of the previous paragraph holds for* $\theta(\cdot)$. *Assume that at least one of the* $\sigma^2_{\alpha,i}, \alpha = a, d, i = 1, 2$, *is positive. Given* $\epsilon > 0$ *and* $t_1 > 0$, *let* $T^n_\epsilon(t_1)$ *denote the Lebesgue measure of the closure set of time points on* $[0, t_1]$ *at which* $WL^n(t)$ *is within* ϵ *of a point of discontinuity of* $\theta(\cdot)$. *Then, for each* $\delta > 0$ *and* $t_1 > 0$,

$$\lim_{\epsilon \to 0} \limsup_n P\left\{T^n_\epsilon(t_1) \geq \delta\right\} = 0. \tag{3.28}$$

Correlated vacations of the sources. Up to now, we have supposed that the vacation processes of the two sources are independent of each other. This would be the case if they were due to movement in independent environments or to extraneous interference if the sources were far apart. If the vacations were due to extraneous interference which affected the sources in a similar manner, then the vacation intervals would be correlated. The main problem in introducing such correlation is algebraic, in that it complicates the expressions.

Let us first suppose that, in addition to the mutually independent vacations specified by A2.3 and A2.4, there are also simultaneous vacations of the two sources, as defined by the following condition.

A3.1. *For each* n, *the intervals between the end of a simultaneous vacation and the start of the next one are denoted by* $n\tau^{m,n}_l, l = 1, \ldots$. *They are mutually independent, exponentially distributed, independent of all the other "driving" random variables and have rate* $\bar\lambda^{m,n}/n$, *where* $\bar\lambda^{m,n}$ *converges to* $\bar\lambda^m > 0$ *as* $n \to \infty$.

A3.2. *For each* n, *there are mutually independent and identically distributed random variables* $\tau^{mv,n}_l, l = 1, \ldots$, *such that the duration of the* l*th simultaneous vacation interval is* $\sqrt{n}\tau^{mv,n}_{i,l}$. *Also,* $\tau^{mv,n}_l$ *converges weakly to a random variable* τ^{mv}_l *as* $n \to \infty$. *The* $\tau^{mv,n}_l, l = 1, \ldots$, *are independent of all other "driving" random variables.*

If A3.1 and A3.2 are added to the conditions of Theorem 3.1 or Theorem 3.2, then the results would be the same, except for the addition of another (independent) jump process $J^m(\cdot)$. Let $\nu^{m,n}_l$ denote the (scaled) starting times of the successive mutual vacations, and let ν^m_l denote the weak sense limits. The weak sense limit (in the sense used in Theorem 3.1) equation is

$$WL(t) = WL(0) + bt + w(t) + \sum_i J_i(t) + J^m(t) + Z(t), \tag{3.29}$$

where

$$J^m(t) = \sum_{l:\nu_l^m \le t} \xi_l^m, \tag{3.30}$$

where ξ_l^m is the weak sense limit of (see (3.3))

$$\bar{A}_l^{m,n} = \frac{1}{\sqrt{n}} \sum_{l=nS_1^{a,n}(\nu_l^{m,n}-)+1}^{nS_1^{a,n}((\nu_l^{m,n}+\tau_l^{m,n}/\sqrt{n})-)} \Delta_{1,l}^{d,n} + \frac{1}{\sqrt{n}} \sum_{l=nS_2^{a,n}(\nu_l^{m,n}-)+1}^{nS_2^{a,n}((\nu_l^{m,n}+\tau_l^{m,n}/\sqrt{n})-)} \Delta_{2,l}^{d,n}, \tag{3.31}$$

and the limit is just τ_l^m, owing to the analysis done for the $\bar{A}_{i,l}^{j,n}$ in Theorem 3.1 and A2.2.

4. The limit control problem. The limit dynamical model. Theorem 3.1 enables us to write the correct limit control problem. As usual in heavy traffic modeling, the aim is to use the limit control problem to get good controls for the physical problem and approximations to its optimal costs, under heavy traffic. The limit dynamics are defined by (3.7) and (3.8), where the jumps are defined by (3.5), where an admissible control $u(\cdot)$ satisfies $u(t) \le WL(t)$ and is nonanticipative in the sense that it is a measurable process, and, for each t, $u(t)$ is independent of $w(t+\cdot) - w(t), N_i(t+\cdot) - N_i(t-), i = 1, 2$. Since $WL(\cdot)$ is continuous at all t where there are no jumps and has a left-hand limit at t if there is a jump there, $WL(t-)$ is well defined for all t. However, $u(t-)$ is not necessarily defined. If the control for (3.7), (3.8) is defined via a function such as the $\theta(\cdot)$ in A2.6b, then we would have $u(t) = \theta(WL(\cdot))$ and $u(t-)$ is well defined for almost all t, which ensures that $u(\nu_{i,l}-)$ is well defined with probability 1. This is one of the main reasons for our interest in control functions such as $\theta(\cdot)$. Alternatively, we could write the jumps as

$$\xi_{1,l}^v = \left[\rho_1 \tau_{1,l}^v - [WL(\nu_{1,l}-) - u(\nu_{1,l})]\right]^+, \tag{4.1a}$$

$$\xi_{2,l}^v = \left[\rho_2 \tau_{2,l}^v - u(\nu_{2,l})\right]^+, \tag{4.1b}$$

where $u(\cdot)$ is a "predictable" process [15, 23]. All that is important is the nonanticipativity as defined above, so that, for each t, $u(t)$ is independent of any jump that might occur at t.

The cost function. Let $c_i(\cdot)$ be a strictly increasing and continuous real-valued function on $[0, \infty)$ with $c_i(0) = 0$, and satisfying $c_i(x) \le Kx + K$ for some $K < \infty$. We will work with a discounted cost function. The cost rate will depend on whether we are penalizing queued jobs or queued workload. In the latter case, we penalize the workloads individually and simply use the cost rate

$$\sum_i c_i(WL_i^n(\cdot)) = c_1(u^n(\cdot)) + c_2(WL^n(\cdot) - u^n(\cdot)) \equiv c(WL^n(\cdot), u^n(\cdot)).$$

In the former case, we would like to penalize the queue sizes individually, i.e., with a cost rate $\sum_i c_i(x_i^n(t))$. Since, in general, we do not have a weak convergence result for the $x_i^n(\cdot)$ for an arbitrary admissible control policy, we are still forced to work with the workload formulation. Then we asymptotically approximate in terms of workload as

$$\sum_i c_i(x_i^n(t)) \approx c_1\left(\frac{u^n(t)}{\bar{\Delta}_1^{d,n}}\right) + c_2\left(\frac{WL^n(t) - u^n(t)}{\bar{\Delta}_2^{d,n}}\right) \equiv c(WL^n(t), u^n(t)). \tag{4.2}$$

Let $\beta > 0$ be the discount factor. The cost function for the limit system will be

$$W_\beta(WL(0), u(\cdot)) = E \int_0^\infty e^{-\beta t} c(WL(t), u(t)) dt, \tag{4.3a}$$

and for the physical system it will be

$$W_\beta^n(WL^n(0), u^n(\cdot)) = E \int_0^\infty e^{-\beta t} c(WL^n(t), u^n(t)) dt. \tag{4.3b}$$

Define $V_\beta(WL(0)) = \inf_u W_\beta(WL(0), u(\cdot))$ and $V_\beta^n(WL^n(0)) = \inf_{u^n} W_\beta^n(WL(0), u^n(\cdot))$, where the $u^n(\cdot)$ and $u(\cdot)$ are admissible. Under a feedback control $\theta^n(\cdot)$ and associated polling policy satisfying A2.6b, we can write

$$W_\beta^n(WL^n(0), \theta^n(\cdot)) = E \int_0^\infty e^{-\beta t} c(WL^n(t), \theta^n(WL^n(t))) dt. \tag{4.4}$$

A restriction of the class of controls and a redefinition of the inf. As noted, we would like to show that a nearly optimal control for the limit problem is nearly optimal for the physical problem for large n and that

$$V_\beta^n(WL^n(0)) \to V_\beta(WL(0)) \tag{4.5}$$

if $WL^n(0)$ converges weakly to $WL(0)$, analogously to what was done in [2, 20, 21]. This is hard to do, since the control appears in the dynamics (3.7) only via the magnitude of the jumps.

The usual method [2, 20, 21] for showing (4.5) involves writing the control in some form such that the sequence of optimal (or ϵ-optimal) controls for the physical problem is tight, and any weak sense limit of the (state process, control process) is an admissible limit control problem. For example, suppose that we have a problem where the control is a vector-valued function which takes values in a compact set. Then, we would write the controls in relaxed control form [19] with the weak topology on them. Since any sequence of such relaxed controls is tight (in the weak topology which is normally used), there is always a weakly convergent subsequence. One could attempt the same thing here. The sequence of relaxed control representations of the control will be tight. The problem is that our $u(\cdot)$ is the derivative of the relaxed control. This derivative is defined only almost everywhere, and in particular, it is not guaranteed that the weak sense limit of $u^n(\nu_{i,l}^n -)$ would be $u(\nu_{i,l} -)$, where this $u(\cdot)$ is the derivative of the weak sense limit of the relaxed control representations. The problem is that, while the cost rate can be written as a linear function of the relaxed control, the jump distribution depends only on specific values of the $u^n(\cdot)$.

One can circumvent these difficulties. However, in order to apply any control which is nearly optimal for the limit system to the physical system, the form in which $u(\cdot)$ appears in (3.5) essentially implies that it should be a feedback control which is continuous "most of the time." Because of these facts, we will be concerned with the more restricted class of controls of A2.6b defined by the following assumption. It is seen from the examples in section 5 and numerical data (taken without the restriction in A4.1) that A4.1 is not restrictive.

A4.1. *Given an integer M, let Θ denote a class of functions, each member of which satisfies* A2.6b, *has at most M points of discontinuity, and on each finite interval $[0, WL]$, the functions in Θ are equicontinuous between discontinuities. The controls will be restricted to such a class, with the polling policy being as defined in* A2.6b.

Fix the class Θ for some M and modulus of equicontinuity. *Redefine* $V_\beta^n(WL(0))$ and $V_\beta(WL(0))$ to be the infima over controls in the class Θ.

A slight alteration of the proof of Theorem 3.1 yields the following. Assume the conditions of Theorem 3.1 and let $\theta^n(\cdot) \in \Theta$. Choose the weakly convergent subsequence such that $\theta^n(\cdot)$ also converges (in the Skorohod topology) to, say, $\theta(\cdot)$. Then the conclusions of Theorems 3.1 and 3.2 hold. Suppose, in addition, that the function whose expectation is being taken in (4.4) is uniformly (in n) integrable. Then the weak convergence in Theorems 3.1 and 3.2 implies that the expected value in (4.4) converges to the expected value for the controlled limit system. We state this in the following more restrictive way, since that is the way it will be verified. The proof is simpler than those in [2, 20, 21] for other control problems under heavy traffic, owing to the more restricted class of allowed controls. The proof implies that a good control for the limit problem will be good for the physical problem under heavy traffic.

THEOREM 4.1. *Assume* A4.1 *and the conditions of Theorem* 3.1. *Let* $c_i(\cdot)$ *satisfy the conditions imposed at the beginning of this section. Suppose that there is a real* C_1 *such that*

$$\sup_{\theta(\cdot)\in\Theta} E\,|WL^n(t)|^2 \le C_1 t + C_1. \tag{4.6}$$

Then the function whose expectation is being taken in (4.4) *is uniformly integrable and*

$$V_\beta^n(WL^n(0)) \to V_\beta(WL(0)).$$

Comments on the proof. Let $\epsilon > 0$ be small and arbitrary. Let $\theta^{\epsilon,n}(\cdot)$ and $\theta^\epsilon(\cdot)$ be ϵ-optimal controls in Θ for the processes $WL^n(\cdot)$ and $WL(\cdot)$, with the initial conditions $WL^n(0)$ and $WL(0)$, respectively. Condition A4.1 implies that, by choosing a subsequence if necessary, we can suppose that $\theta^{\epsilon,n}(\cdot)$ converges to some $\bar\theta(\cdot) \in \Theta$ in the Skorohod topology. Then

$$\epsilon + V_\beta^n(WL^n(0)) \ge W_\beta^n(WL^n(0), \theta^{\epsilon,n}(\cdot)) \to W_\beta(WL(0), \bar\theta(\cdot)) \ge V_\beta(WL(0)).$$

Thus,

$$\liminf_n V^n(WL^n(0)) \ge V_\beta(WL(0)).$$

Now, apply $\theta^\epsilon(\cdot)$ to $WL^n(\cdot)$ to get

$$V_\beta^n(WL^n(0)) \le W_\beta^n(WL^n(0), \theta^\epsilon(\cdot)) \to W_\beta(WL(0), \theta^\epsilon(\cdot)) \le V_\beta(WL(0)) + \epsilon.$$

These inequalities yield the theorem. □

On the condition (4.6). Condition (4.6) is not a consequence of the other conditions. Write (3.15) for the general case where the vacations are included as

$$WL^n(t) = h^n(t) + Z^n(t).$$

It follows from the estimates given for the general Skorohod problem in [10, Theorem 2.2.] that there is a real C such that

$$WL^n(t) \le C \sup_{s\le t} |h^n(s)| \quad \text{for all } t.$$

Thus, a sufficient condition for (4.6) is that the following inequalities hold.

$$\sup_n E\,|WL^n(0)|^2 < \infty, \tag{4.7}$$

$$E \sup_{s\le t} |w_i^{\alpha,n}(S_i^{a,n}(s))|^2 = O(t),\ \alpha = a,d, i = 1,2, \tag{4.8}$$

$$\bar{J}^n(t) = \sum_j E \left| \sum_{k:\nu_{j,k}^n \le t} \xi_{j,k}^{v,n} \right|^2 = O(t^p) \text{ for some } p > 0. \tag{4.9}$$

Dealing with (4.8) and (4.9) in detail will take us far afield, but they do hold under quite broad conditions. To illustrate one of the possibilities, we will give some of the details under the following condition.

A4.2. *For each n, the random variables $(\Delta_{i,l}^{\alpha,n}, l < \infty)$ are mutually independent and identically distributed for each $i = 1,2, \alpha = a,d$, and the absolute third moments are uniformly bounded. There are $\bar{\Delta}_i^{\alpha,n}$ and $\bar{\Delta}_i^{\alpha}$ such that $E\Delta_{i,l}^{a,n} = \bar{\Delta}_i^{\alpha,n} \to \bar{\Delta}_i^{\alpha}$, $\alpha = a,d$. Also, the second moments of $\tau_{i,l}^{v,n}$ are uniformly bounded.*

THEOREM 4.2. *Assume* A2.3, A2.4, *and* A4.2. *Then* (4.8) *and* (4.9) *hold.*

Proof. First, we consider (4.8). Fix α and i and define $\psi_{i,l}^{\alpha,n} = (1 - \Delta_{i,l}^{\alpha,n}/\bar{\Delta}_i^{\alpha,n})$. Let $\mathcal{F}_{i,l}^{\alpha,n}$ denote the minimal ν-algebra which measures $\{\psi_{i,j}^{\alpha,n}, j \le l\}$, and write $E_{i,l}^{\alpha,n}$ for the associated conditional expectation. The $\psi_{i,l}^{\alpha,n}$ are martingale differences in that $E_{i,l}^{\alpha,n}\psi_{i,l+1}^{\alpha,n} = 0$ with probability one for all l. There is $C_2 < \infty$ such that

$$E_{i,l}^{\alpha,n} \left|\psi_{i,l+1}^{\alpha,n}\right|^2 \le C_2.$$

Define

$$N_i^{\alpha,n}(t) = \frac{1}{n} \times \min\left\{ n : \sum_{l=1}^{n} \Delta_{i,l}^{\alpha,n} \ge nt \right\}.$$

The $S_i^{\alpha,n}(t)$ and $N_i^{\alpha,n}(t)$ will differ by at most $1/n$. The $N_i^{\alpha,n}(t)$ have the advantage that they are stopping times with respect to the filtrations $\mathcal{F}_{i,l}^{\alpha,n}$. In particular, $\{\omega : N_i^{\alpha,n}(t) \ge l\} \in \mathcal{F}_{i,l-1}^{\alpha,n}$. We have

$$E \max_{s\le t} |w_i^{\alpha,n}(S_i^{\alpha,n}(s))|^2 \le E \max_{m \le nN_i^{\alpha,n}(t)} \frac{1}{n} \left| \sum_{l=1}^{m} \psi_{i,l}^{\alpha,n} \right|^2. \tag{4.10}$$

Owing to the martingale properties, the right-hand side of (4.10) is bounded by $C_2 E N_i^{\alpha,n}(t)$. Thus, we need to bound $EN_i^{\alpha,n}(t)$.

For an integer $m > 0$, write

$$\frac{m \wedge nN_i^{a,n}(t)}{n} = \frac{1}{n} \sum_{l=1}^{m\wedge(nN_i^{a,n}(t))} 1 = \frac{1}{n} \sum_{l=1}^{m\wedge(nN_i^{a,n}(t))} \psi_{i,l}^{a,n} + \frac{1}{n\bar{\Delta}_i^{a,n}} \sum_{l=1}^{m\wedge(nN_i^{a,n}(t))} \Delta_{i,l}^{a,n}. \tag{4.11}$$

The expectation of the first term on the right is zero. Dropping that term and letting $m \to \infty$ yields

(4.12)
$$EN_i^{a,n}(t) = \frac{t}{\bar{\Delta}_i^{a,n}} + \frac{1}{n} E\left[(\text{first time of arrival to source } i \text{ at or after } nt - nt)\right].$$

Thus

$$EN_i^{a,n}(t) = \frac{t}{\bar{\Delta}_i^{a,n}} + \delta_n,$$

where $\lim_n \delta_n = 0$ and (4.8) holds. (The proof of the renewal theorem for the "excess life" in [13, pp. 192–193] implies that $\delta_n \to 0$.)

Now turn to the proof of (4.9). To bound (4.9), we can use the expression

$$E\left|\xi_{j,k}^{v,n}\right|^2 \le \sum_i \frac{1}{n} E \left| \sum_{l=nS_i^{a,n}(\nu_{j,k}^n)+1}^{nS_i^{a,n}(\nu_{j,k}^n+\tau_{j,k}^{v,n}/\sqrt{n})} \Delta_{i,l}^{d,n} \right|^2. \tag{4.13}$$

Writing $\Delta_{i,l}^{d,n} = [\Delta_{i,l}^{d,n} - \bar{\Delta}_i^{d,n}] + \bar{\Delta}_i^{d,n}$ in (4.13) and splitting the upper bound in (4.13) into the two corresponding parts yields a bound on $E|\xi_{j,k}^{v,n}|^2$ as (twice) the sum of

$$\sum_i \left[\bar{\Delta}_i^{d,n}\right]^2 E\left|w_i^{d,n}(\nu_{j,k}^n + \tau_{j,k}^{v,n}/\sqrt{n}) - w_i^{d,n}(\nu_{j,k}^n)\right|^2 \le C_2 \sum_i \left[\bar{\Delta}_i^{d,n}\right]^2 E\tau_{j,k}^{v,n}/\sqrt{n}$$

and

$$\sum_i \frac{[\bar{\Delta}_i^{d,n}]^2}{n} E\left[\#\text{arrivals at queue } i \text{ in real time } \left[n\nu_{j,k}^n, n\nu_{j,k}^n + \sqrt{n}\tau_{j,k}^{v,n}\right]\right]^2.$$

The first expression is $O(1/\sqrt{n})$. Following the idea in the expansion (4.11), for the second expression we get the bound, for some real C_3,

$$C_3 E\left[\tau_{j,k}^{v,n}\right]^2 + C_3 E\left[\text{a residual time term}\right]^2/n.$$

However, by the cited proof of the renewal theorem for the "excess life" in [13, pp. 192–193], and the third moment condition in A4.2, the mean square value of the residual time term is bounded, uniformly in all indices.

We have obtained a bound for the mean square value of each of the jumps. To complete the proof, we need to average over the number of vacations on real time $[0, nt]$. We do this by ignoring the vacation durations in computing the distribution of the number of vacations on any real time interval $[0, nt]$, which gives an upper bound. Then the number has a Poisson distribution for each n, with the rate parameter being bounded in n, and the number is independent of the jump sizes. If there are L vacations on $[0, nt]$, then (4.9) is bounded by L^2 times the bound on the mean square value of each jump. Finally, using the Poisson distribution, average over L to get (4.9) for $p = 2$. □

The Bellman equation for the limit system. Let $\mathcal{L}$ denote the differential generator of the pure diffusion part of (3.7): i.e., for smooth real-valued $f(\cdot)$, $\mathcal{L}f(WL) = \sigma^2 f_{ww}(WL)/2 + bf_w(WL)$. Write the control in feedback form $u(t) =$

$\theta(WL(t)) \le WL(t)$ for some measurable function $\theta(\cdot)$. The jump part of the differential operator acting on a measurable real-valued function $f(\cdot)$ is

$$\sum_i \bar{\lambda}_i^s E\left[f(WL + \xi_i^v) - f(WL)\right], \tag{4.14}$$

where ξ_i^v is the jump due to a vacation of source i, and E denotes the expectation of the jump given the WL and the control just before the start of the jump. The boundary condition is $f_w(0) = 0$.

Define $\bar{V}_\beta(WL)$ to be the inf of the cost over all admissible controls, not only those of the form in (A4.1). Define the function

$$H(\bar{V}_\beta, WL) = \min_{\theta(WL) \le WL} \left\{ c(WL, \theta(WL)) + \sum_i \bar{\lambda}_i^s E\left[\bar{V}_\beta(WL + \xi_i^v) - V_\beta(WL)\right] \right\}. \tag{4.15}$$

The formal Bellman equation is the partial differential integral equation

$$\mathcal{L}\bar{V}_\beta(WL) - \beta \bar{V}_\beta(WL) + H(\bar{V}_\beta, WL) = 0, \tag{4.16}$$

with the boundary condition $\bar{V}_{\beta,WL}(0) = 0$. The subscript WL denotes the derivative.

Conjecture and assumption. We have not been able to find anything in the literature concerning the PDE (4.16), where the jump magnitudes are controlled. To fully justify the restriction A4.1, it is necessary to show both that (4.16) has a unique (either classical or viscosity sense) solution which is the minimal cost and that the minimizing $\theta(\cdot)$ in (4.15) is of the type in A2.6b. This seems to be a very reasonable expectation, although we have not been able to demonstrate it. As noted in the next section, it is essentially obvious in certain special cases, e.g., where $c_i(x) = x_i$, and we expect that it holds under broad conditions on $c(\cdot)$. Note that this is not an impulse control problem. The jump times are those of a Poisson process.

Thus, we assume that our conjecture is true, namely, that the minimum cost satisfies (4.16) and that the optimal control, given by the minimizer in (4.15), satisfies A2.6b. Under this assumption, an optimal control for the limit problem is nearly optimal for the physical problem under heavy traffic if the controls for the physical problem are restricted to a large enough class of the type in A4.1.

5. Extensions and comments. In special cases, the weak convergence results and the form of the limit problem suggest nearly optimal strategies for the physical problem, without much additional analysis. A case of current interest will be discussed.

Minimizing the total expected workload. Suppose that the cost rates $c_i(\cdot)$, written in terms of workload, satisfy $c_i(WL_i) = WL_i$. Then $c(WL, \theta(WL)) = WL$ and the control problem is the minimization of the expectation of the integral of the discounted total workload. The mean total workload $EWL(t)$ for the limit problem is minimized, uniformly in t, by using the policy $\theta(\cdot)$ that minimizes the mean jump, namely,

$$Q := \bar{\lambda}_1^s E\left[\rho_1 \tau_1^v - [WL - \theta(WL)]\right]^+ + \bar{\lambda}_2^s E\left[\rho_2 \tau_2^v - \theta(WL)\right]^+. \tag{5.1}$$

Example: The case of exponentially distributed vacation intervals. As an example, assume that τ_i^v is exponentially distributed with parameter v_i. Note that for any real

number y and any random variable τ, exponentially distributed with parameter w, we have

$$E(\tau - y)^+ = w \int_y^\infty e^{-wx}(x-y)dx = w \int_0^\infty e^{-w(y+z)} z dz = \frac{e^{-wy}}{w}.$$

Denote for $i = 1, 2$, $j = 1, 2$, $j \neq i$,

$$w_i = \frac{v_i}{1 - \bar{\lambda}_j^a / \bar{\lambda}_j^d}.$$

Then we obtain

$$Q = \frac{\bar{\lambda}_1^s e^{-w_1(WL-\theta)}}{w_1} + \frac{\bar{\lambda}_2^s e^{-w_2\theta}}{w_2}.$$

Thus, for each WL, Q is convex with respect to θ, and its minimum is obtained at θ for which $dQ(\theta)/d\theta = 0$, provided that this solution satisfies $\theta \in (0, WL)$. If it does not, then the minimum over $\theta \in [0, WL]$ is obtained on one of the boundaries. Differentiating with respect to θ yields

$$\bar{\lambda}_1^s e^{-w_1(WL-\theta)} - \bar{\lambda}_2^s e^{-w_2\theta} = 0.$$

Solving this equation yields

$$\theta(WL) = \frac{\log(\bar{\lambda}_2^s / \bar{\lambda}_1^s)}{w_1 + w_2} + \frac{w_1}{w_1 + w_2} WL.$$

A nearly optimal policy $\theta(\cdot)$ for the limit problem should be defined by

$$\theta^*(WL) = \Big(\min \Big(WL, \frac{\log(\bar{\lambda}_2^s / \bar{\lambda}_1^s)}{w_1 + w_2} + \frac{w_1}{w_1 + w_2} WL \Big) \Big)^+,$$

and this is borne out by numerical solutions.

Example: The symmetrical case. In the special case where the two sources have the same rates, it is obvious that $\theta(WL) = WL/2$. Thus, under the conditions of Theorem 3.1 and the uniform integrability conditions of Theorem 4.1, the minimization of (5.1) yields a nearly optimal strategy for large n.

No vacations. The asymptotic optimality of the $c\mu$-rule. Suppose that there are no vacations and the basic desired cost rate is $\bar{c}_1 x_1^n + \bar{c}_2 x_2^n$, where $\bar{c}_i > 0$. Write the limit form of the cost rate in terms of the workload as

$$\bar{\lambda}_1^d \bar{c}_1 \theta(WL) + \bar{\lambda}_2^d \bar{c}_2 [WL - \theta(WL)]. \tag{5.2}$$

The minimizer of (5.2) is just the $c\mu$-rule. Namely, poll source 1 if $\bar{\lambda}_1^d \bar{c}_1 > \bar{\lambda}_2^d \bar{c}_2$ and there are jobs there, and conversely for source 2. Under the conditions of Theorem 3.1 and the uniform integrability conditions, such a rule would be asymptotically optimal for the physical system. In this case, the limit workload does not depend on the polling policy, only the cost rate does. This is an asymptotic form of the well-known $c\mu$-rule [31]. The asymptotic optimality of this rule under heavy traffic was given in [30].

The $c\mu$-rule gives priority to one of the queues, and this might lead to unacceptably long waits in the nonpriority queue. This can be alleviated with a nonlinear

weighting. For example, queue 1 might have a smaller cost rate than queue 2 for moderate queue lengths. But to discourage the complete priority of queue 1, we might use a nonlinear cost rate for queue 2.

Remark. Note that the optimal policies in all the examples in this section indeed satisfy assumption A4.1, which is required in Theorem 4.1. Thus the class of policies described by A4.1 is rich enough to contain an optimal policy within it for these asymptotic problems.

6. Stability.

DEFINITION: STABILITY, UNIFORMLY IN $\boldsymbol{n}$ FOR LARGE $\boldsymbol{n}$. Suppose that there are real n_0 and $\bar{W}$ such that

$$E\big[\text{time for } WL^n(t), t \geq t_0, \text{ to return to the value } WL^n(t) \leq \bar{W} \big| \text{ data to real time } nt_0, WL^n(t_0) = q\big] \leq F(q)$$

for all $n \geq n_0, t_0, q$, where $F(q)$ is bounded on each bounded q-set. Then we say that $WL^n(\cdot)$ is *stable, uniformly in* n.

DEFINITION: STABILITY FOR FIXED $\boldsymbol{n}$. Fix n, and suppose that the above conditional mean return time property holds for all q and t_0. Then, for that value of n, the queue is said to be *stable.*

We will use the following assumption, a modification of A2.2.

A6.1. *There is a real* $b_0 < 0$*:* $\sqrt{n}\left[\sum_i \rho_i^n - 1\right] \leq b_0$ *for all* n.

Comments on stability. Under A6.1, it is trivial to prove the stability of the weak sense limit system (3.7) using classical stochastic Liapunov function methods, as in [14, 16, 17]. Stability is one of the most important properties of physical systems, and should be proved under broad conditions. It is essentially an assertion on the robustness of the system, and should hold under reasonable perturbations of the basic data. Stability of the physical system is not automatically guaranteed by stability of the weak sense limit. The technique to be employed is versatile and gets the desired stability property, uniformly in reasonable perturbations of the basic data, in the sense that the function $F(q)$ will not depend on the exact form of the data, under a reasonable mixing-type condition. The first definition above concerns large n. It will be seen that if there are no vacations then we can set $n_0 = 1$ in that definition, under broad conditions on the data.

Under the conditions of Theorem 3.1, the ratio of time on vacation to total time goes to zero as $n \to \infty$. If n is fixed and small, then it is conceivable that this ratio would be large enough so that the accumulation of data during the vacations will not be offset by the processing between vacations, as is necessary for stability. However, from the point of view of stability with vacations, there is an equivalence between large n and small $\bar{\lambda}_i^s$. This explains the last assertion of Theorem 6.2.

First, we will provide the motivation for the perturbed Liapunov function approach. Then it is used for the problem without vacations and stability uniformly in n (not just in large n) is proved. Then vacations are added. We will simplify the algebra by supposing that arrivals to the queues can occur only at multiples of (real time) $\delta > 0$, which can be as small as desired. Otherwise, we would use integrals in lieu of sums, but the results would be the same. Also, again for notational simplicity and with little loss of generality, we also suppose that vacations start and stop only at integral multiples of δ, and modify the assumptions A2.3 and A2.4 appropriately.

We will also use A2.5 plus other (weak) conditions to be imposed below. Let $E^n_{k\delta}$ denote the expectation, given all of the system data up to and including *real time* $k\delta$. Let $I^{a,n}_{i,k\delta}$ be the indicator function of an arrival at real time $k\delta$ from source i, and let $\Delta^{d,n}_{i,k\delta}$ be the associated work, if there is an arrival.

Motivation and background: Perturbed Liapunov functions. A perturbed Liapunov function method will be used [6, 17, 18, 22]. The classical Liapunov function method is quite limited for problems such as ours, since there is not usually a "contraction" at each step to yield the local supermartingale property of a classical Liapunov function. The perturbed Liapunov function method is a powerful extension of the classical method. In the perturbed Liapunov function method, one adds a small perturbation to the original Liapunov function. As will be seen, when this perturbation can be well defined it provides an "averaging" which is needed to get the local supermartingale property.

The primary Liapunov function will be simply $WL(\cdot)$. The final Liapunov function will be of the form $W^n(\cdot) = WL^n(\cdot) + \delta W^n(\cdot)$, where $\delta W^n(\cdot)$ is bounded. Suppose that there is no vacation at real time $k\delta$. Then, for $WL^n(k\delta/n) \geq \delta$, we can write

$$E^n_{k\delta} WL^n(k\delta/n + \delta/n) - WL^n(k\delta/n) = -\frac{\delta}{\sqrt{n}} + \frac{1}{\sqrt{n}} \sum_i E^n_{k\delta} I^{a,n}_{i,k\delta+\delta} \Delta^{d,n}_{i,k\delta+\delta}. \tag{6.1}$$

The right-hand term needs to be "averaged," and this is done with the use of a perturbation function $\delta W^n(\cdot)$.

Motivation using a simpler problem. Before defining the actual perturbation which will be used, for motivation we will discuss the general principle with a simpler form when there are no vacations. Even for this problem, stability of the physical queues is not guaranteed by stability of the limit system. Let $\bar{\Delta}^{a,n}_i = 1/\bar{\lambda}^{a,n}_i$ and $\bar{\Delta}^{d,n}_i$ be centering constants such that the corresponding ρ^n_i satisfy A6.1 for some $b_0 < 0$. More will be said about them later.

Proceeding formally until further notice, define the first suggested perturbation:

$$\delta\tilde{W}^n(k\delta/n) = \frac{1}{\sqrt{n}} \sum_i \sum_{j=k+1}^{\infty} E^n_{k\delta} \left[I^{a,n}_{i,j\delta} \Delta^{d,n}_{i,j\delta} - \delta\bar{\lambda}^{a,n}_i \bar{\Delta}^{d,n}_i \right]. \tag{6.2}$$

Clearly, the centering constants must be such that the sum is well defined, and we return to this point below. If $WL^n(k\delta/n) \geq \delta$, then we get

$$\begin{aligned} &E^n_{k\delta}\delta\tilde{W}^n(k\delta/n + \delta/n) - \delta\tilde{W}^n(k\delta/n) \\ &\qquad = -\frac{1}{\sqrt{n}} \sum_i E^n_{k\delta} \left[I^{a,n}_{i,k\delta+\delta} \Delta^{d,n}_{i,k\delta+\delta} - \delta\bar{\lambda}^{a,n}_i \bar{\Delta}^{d,n}_i \right]. \end{aligned}$$

Define $\tilde{W}^n(k\delta/n) = WL^n(k\delta/n) + \delta W^n(k\delta/n)$. Then (6.1) and the last expression yield

$$E^n_{k\delta}\tilde{W}^n(k\delta/n + \delta/n) - \tilde{W}^n(k\delta/n) = -\frac{\delta}{\sqrt{n}} + \frac{1}{\sqrt{n}} \sum_i \left[\delta\bar{\lambda}^{a,n}_i \bar{\Delta}^{d,n}_i \right].$$

By the condition A6.1, the right side of the last expression is asymptotically $\leq b_0\delta/n$. Thus, it is less than the negative constant b_0 times the scaled time interval δ/n. Hence, when $WL^n(t) \geq \delta$, $\tilde{W}^n(k\delta/n)$ has the supermartingale property and we can use this to get the desired (uniform in n) stability if $\tilde{W}^n(\cdot)$ is "well defined and bounded."

Let us examine the sum in (6.2) more closely to see why it is well defined and bounded under broad mixing conditions. Since $\bar{\Delta}_i^{d,n}$ and $\bar{\lambda}_i^{a,n}$ are merely *centering* constants for the *entire* sequence, the actual mean values or rates can vary with time (say, being periodic, etc.). Fix k and let $\mu_{i,1}\delta$ and $\mu_{i,2}\delta$ be the real times of the first two arrivals to queue i after real time $k\delta$. Formally, consider the part of the inner sum in (6.2) given by

$$E_{k\delta}^n \sum_{j=\mu_{i,1}+1}^{\mu_{i,2}} \left[I_{i,j\delta}^{a,n} \Delta_{i,j\delta}^{d,n} - \delta \bar{\lambda}_i^{a,n} \bar{\Delta}_i^{d,n} \right].$$

This equals

$$E_{k\delta}^n \left[\Delta_{i,\mu_{i,2}\delta}^{d,n} - (\mu_{i,2} - \mu_{i,1})\delta \bar{\lambda}_i^{a,n} \bar{\Delta}_i^{d,n} \right]. \tag{6.3}$$

Next, suppose that the interarrival times and workloads are mutually independent, with the members of each set being mutually independent and identically distributed, with finite second moments, and means $\bar{\Delta}_i^{a,n}, \bar{\Delta}_i^{d,n}$. Then (6.3) equals zero, since $E_{k\delta}^n(\mu_{i,2} - \mu_{i,1})\delta = \bar{\Delta}_i^{a,n}$. Obviously, for any integer m, $\mu_{i,1}, \mu_{i,2}$ can be the mth and $(m+1)$st arrival times with the same result. Thus, under the independence assumptions, (6.2) is just

$$E_{k\delta}^n \left[\Delta_{i,\mu_{i,1}\delta}^{d,n} - \delta(\mu_{i,1} - k)\bar{\lambda}_i^{a,n} \bar{\Delta}_i^{d,n} \right] = \bar{\Delta}_i^{d,n} E_{k\delta}^n \left[1 - \frac{\delta(\mu_{i,1} - k)}{\bar{\Delta}_i^{a,n}} \right],$$

where $E_{k\delta}^n(\mu_{i,1} - k)\delta$ is just the conditional expectation of the mean time to the next arrival after $k\delta$, given the data to time $k\delta$. For use below, keep in mind that this quantity is bounded uniformly in k, under the above assumptions on the independence and the moments. Hence, formally, $\delta\tilde{W}^n(t)$ is of the order of $1/\sqrt{n}$, uniformly in all variables.

Now, suppose that the interarrival times are as in the last paragraph, but the service times are correlated, still with centering constant $\bar{\Delta}_i^{d,n}$. Let $\mu_{i,j}, j = 1, \ldots,$ denote the sequence of arrival times after $k\delta$. Then (6.3) equals

$$E_{k\delta}^n \left[\Delta_{i,\mu_{i,2}\delta}^{d,n} - \bar{\Delta}_i^{d,n} \right]. \tag{6.4}$$

Then, grouping terms and formally speaking, we see that the sum (6.2) is just (6.3)$/\sqrt{n}$, plus a series

$$\frac{1}{\sqrt{n}} \sum_i \sum_{l=\mu_{i,2}}^{\infty} E_{k\delta}^n \left[\Delta_{i,l}^{d,n} - \bar{\Delta}_{i,l}^{d,n} \right].$$

Clearly, the inner sum is bounded under quite broad mixing conditions. All that is needed is that $E_{k\delta}^n[\Delta_{i,l}^{d,n} - \bar{\Delta}_{i,l}^{d,n}] \to 0$ is a summable way as $l - k \to \infty$. A similar computation can be done if the $\Delta_{i,l}^{a,n}$ are correlated.

If the inner sum in (6.2) is well defined and bounded (uniformly in n, k, ω), then Theorem 6.1, which summarizes the above discussion, proves stability, uniformly in (all) n and the discounting which is used there is not needed. While the inner sums of (6.2) are well defined under broad conditions, there are interesting examples where they are not. For example, consider the case where $\Delta_{i,l}^{a,n} = H\delta = 1/\bar{\lambda}_i^{a,n}$, where H is

an integer and the work in all jobs is just the constant $\bar{\Delta}_i^{d,n}$. Then the inner sum, taken from $k+1$ to m, oscillates between zero and $-(H-1)\delta\bar{\lambda}_i^{a,n}\bar{\Delta}_i^{d,n}/\sqrt{n}$ as $m \to \infty$. The most convenient way of circumventing this problem is to suitably discount the defining sums [22, 27]. Thus, for some small $\beta > 0$, consider the alternative "discounted" perturbation

$$\delta W_\beta^n(k\delta/n) = \frac{1}{\sqrt{n}} \sum_i \sum_{j=k+1}^{\infty} E_{k\delta}^n e^{-(j-k-1)\beta\delta/n} \left[I_{i,j\delta}^{a,n} \Delta_{i,j\delta}^{d,n} - \delta\bar{\lambda}_i^{a,n}\bar{\Delta}_i^{d,n} \right]. \tag{6.5}$$

The sum (6.5) is well defined if $E|\Delta_{i,l}^{d,n}|$ is uniformly bounded, and then the conditional expectation can be taken either inside or outside of the summation.

Stability without vacations. We now proceed to prove the stability results. It is simpler to start with the assumption that there are no vacations. The following additional assumption will be used. The above discussion shows that the assumption covers many cases of interest.

A6.2. *There is real B such that w,p.1 $|\delta W_\beta^n(k\delta/n)| \le B/\sqrt{n}$ for all $\beta > 0$ and all n, k, where $\delta W_\beta^n(\cdot)$ is defined by* (6.5).

Define the final perturbed Liapunov function

$$W_\beta^n(k\delta/n) = WL^n(k\delta/n) + \delta W_\beta^n(k\delta/n). \tag{6.6}$$

THEOREM 6.1. *Let $WL^n(0)$ be tight, suppose that there are no vacations, and assume* A2.5, A6.1, *and* A6.2. *Then the process $WL^n(\cdot)$ is stable, uniformly in n.*

Proof. We have

$$\begin{aligned} &E_{k\delta}^n \delta W_\beta^n(k\delta/n + \delta/n) - \delta W_\beta^n(k\delta/n) \\ &\quad = -\frac{1}{\sqrt{n}} \sum_i E_{k\delta}^n \left[I_{i,k\delta+\delta}^{a,n} \Delta_{i,k\delta+\delta}^{d,n} - \delta\bar{\lambda}_i^{a,n}\bar{\Delta}_i^{d,n} \right] + \epsilon_k^n, \end{aligned} \tag{6.7}$$

where

$$\epsilon_k^n = E_{k\delta}^n \left[1 - e^{-\beta\delta/n} \right] \delta W_\beta^n(k\delta/n + \delta/n). \tag{6.8}$$

Thus, adding (6.1) and (6.7),

$$E_{k\delta}^n W_\beta^n(k\delta/n + \delta/n) - W_\beta^n(k\delta/n) = -\frac{\delta}{\sqrt{n}} + \frac{1}{\sqrt{n}} \sum_i \left[\delta\bar{\lambda}_i^{a,n}\bar{\Delta}_i^{d,n} \right] + \epsilon_k^n. \tag{6.9}$$

By the condition A6.1, the right side of (6.9) is asymptotically no greater than

$$\frac{b_0\delta}{n} + \epsilon_k^n. \tag{6.10}$$

Assumption A6.2 implies that $|\epsilon_k^n| = B[\beta\delta/n]\gamma^n$, where $\gamma^n \to 0$. Thus, for small β

$$E_{k\delta}^n W_\beta^n(k\delta/n + \delta/n) - W_\beta^n(k\delta/n) \le \frac{b_0\delta}{2n}, \quad \text{for } WL^n(k\delta/n) \ge \delta. \tag{6.11}$$

Inequality (6.11) implies that $W_\beta^n(k\delta/n)$ has the supermartingale property when $WL^n(k\delta/n) \ge \delta$. Suppose that $W_\beta^n(k\delta/n) = B_2 > B + \delta$ and let $B_2 > B_1 > B + \delta$. Then, by standard stability arguments [16, 17], the mean number of steps (of real

time length δ and conditioned on the data to real time $k\delta$) for $W_\beta^n(j\delta/n), j \geq k$, to return to the set where $W_\beta^n(j\delta/n) \leq B_1$ is bounded by

$$\frac{W_\beta^n(k\delta/n)}{[-b_0\delta/2n]} \leq \frac{B + WL^n(k\delta/n)}{[-b_0\delta/2n]}.$$

Since $|\delta W_\beta^n(t)| \leq B/\sqrt{n}$, the return time estimate also holds for $WL^n(\cdot)$. Thus, in the time scale which is used to define $WL^n(\cdot)$, where time is compressed by a factor of n, the conditional mean return time is asymptotically bounded by $2[B + WL^n(t)]/[-b_0]$. Hence, we have stability, uniformly in n. □

Stability, with vacations. Now, we add the vacations. Again, to simplify the notation, suppose a preempt-resume discipline, so that we do not have to concern ourselves with redoing all of an interrupted job. The analysis for the latter case follows similar lines.

The polling policy is subject only to the unrestrictive condition A6.3. The condition is motivated by the fact that the $\xi_{i,l}^v$ defined by (3.5) go to zero as the individual workloads go to infinity, since the larger the work remaining in the nonvacationing sources, the less likely it is that the server will have idle time during a vacation. If A6.3 does not hold, then there might not be stability for *each* $b < 0$, uniformly in large n. For example, suppose that the polling policy is to give source 1 priority. Then $WL_1^n(t)$ will be arbitrarily close to zero, except possibly during and for a short interval just after a vacation of that source. Consequently, the mean or conditional mean jump in the total workload during a vacation of source 2 which starts at scaled time t will not go to zero as $WL^n(t)$ goes to infinity. The condition A6.3 excludes exhaustive polling (but only when the workload is very large), where a source is polled until its queue is empty, unless a vacation of that source intervenes. However, when the total workload is large, we might not want to use exhaustive polling anyway. While we work with two sources for notational simplicity, the idea is the same no matter what the number of sources. By our convention, for a vacation that starts at real time $k\delta$, the real time vacation interval is the half open interval $[k\delta + \delta, k\delta + \delta + \sqrt{n}\tau_{i,k\delta}^{v,n})$.

A6.3. *The polling policy is unrestricted, except for the following. There are constants $\bar{W}_a \ll \bar{W}_b$, which will be as large as we wish. If $WL^n(t) \leq \bar{W}_b$, then there are no restrictions. If $WL^n(t) > \bar{W}_b$, then the only restriction is that if*

$$WL_i^n(t) \geq \bar{W}_a \tag{6.12}$$

is not satisfied for some i and the other source is not on vacation, then we poll the other source.

A6.4. *For each $\epsilon > 0$, there is $\bar{W} < \infty$ such that for $i, j : i \neq j$,*

$$E\left[\xi_{i,l}^{v,n} \middle| \text{data to scaled time } \nu_{i,l}^n, WL_j^n(\nu_{i,l}^n-) \geq \bar{W}\right] \leq \epsilon.$$

The assumption A6.4 holds under the conditions of Theorem 4.2 if A6.3 holds. The inequality (6.12) (when $WL^n(t) > \bar{W}_b$) cannot be guaranteed for all time. It will sometimes not hold during a vacation or for a vanishingly short (in scaled time) interval after. The real time interval between vacations is $O(n)$. Let $\alpha(\cdot)$ be a real-valued function on $[0, \infty)$ such that $\alpha(n)/\sqrt{n} \to \infty$ and $\alpha(n)/n \to 0$. Suppose that a vacation ends at real time t_0 and the next one begins at real time t_1, with $t_1 - t_0 = O(n)$. Then with a probability (conditioned on the data up to t_0) that goes to unity as $n \to \infty$, one can poll such that (6.12) is guaranteed on $[t_0 + \alpha(n), t_1)$ when

$WL^n(t_0) > \bar{W}_b$. The excluded interval is just $\alpha(n)/n$ in scaled time. Condition A6.3 works since the probability that two successive vacations will be within $\alpha(n)$ in real time is $O(\alpha(n)/n)$.

THEOREM 6.2. *Let* $\{WL^n(0)\}$ *be tight and assume* A2.3–A2.5 *and* A6.1–A6.4. *Then the process* $WL^n(\cdot)$ *is stable, uniformly in* n. *Fix* n. *Then for small enough* $\bar{\lambda}_i^s, i = 1, 2$, $WL^n(\cdot)$ *is stable.*

Note on the stability of the limit problem (3.7). Let $\mathcal{L}$ denote the differential generator of (3.7) when $WL > 0$. Then

$$\mathcal{L}WL(t) = b + \sum_i \bar{\lambda}_i^s E_{WL(t-),u(t-)}\xi_i^v.$$

Since, for the limit problem, the condition (6.12) can always be guaranteed for $WL(t) > \bar{W}_b$ (except at the jump instants) if $\bar{W}_b \gg \bar{W}_a$ are as large as we wish, it can always be assured that the sum in the above expression is arbitrarily small for large $WL^n(t)$. Then, since $b < 0$, $WL(\cdot)$ is stable. The proof below attempts to duplicate this idea.

Proof. In this proof, it is more convenient to work in scaled time. Thus, let $E_t^{s,n}$ denote the expectation conditioned on all data to scaled time t. All scaled times are integral multiples of δ/n. Suppose that no source is on vacation at scaled time t and a vacation of some source starts at scaled time $t + \delta/n$. Then, let $\tau_t^{v,n}/\sqrt{n}$ denote the scaled time which passes until *no* source is on vacation, and let $\xi_t^{v,n}$ denote the total jump in the workload due to all vacations which start at scaled time $t + \delta/n$ and end at scaled time $t + \delta/n + \tau_t^{v,n}/\sqrt{n}$. Thus, it might cover a single vacation, or several overlapping or abutting vacations.

Define $\sigma_k^n, k \geq 0$, recursively as follows. Start with $\sigma_0^n = 0$. Given σ_k^n, if no vacation starts at scaled time $\sigma_k^n + \delta/n$, then set $\sigma_{k+1}^n = \sigma_k^n + \delta/n$. If a vacation starts at scaled time $\sigma_k^n + \delta/n$, then set $\sigma_{k+1}^n = \sigma_k^n + \delta/n + \tau_{\sigma_k^n}^{v,n}/\sqrt{n}$. Thus, the σ_k^n are the sequence of scaled times $k\delta/n$, but with the instants where some source is on vacation skipped. To prove the stability it is sufficient to work with $WL^n(\sigma_k^n)$ and $WL^n(\sigma_k^n) \geq \bar{W}_a$ only.

Until further notice, suppose that the condition A6.3 holds at the start of each vacation. The event that this is not the case is very rare for large n and will be accounted for later. Recall the definition of $W_\beta^n(\cdot)$ in (6.6). Let $E_t^{v,n}$ denote the expectation, conditioned on all data to scaled time t and the event that a vacation starts at scaled time $t + \delta/n$. By the computations in Theorem 6.1, we have (whether or not (6.12) holds)

(6.13)
$$E_{\sigma_k^n}^{s,n} W_\beta^n(\sigma_{k+1}^n) - W_\beta^n(\sigma_k^n) \leq \prod_i \left(1 - \frac{\bar{\lambda}_i^{s,n}}{n} + o\left(\frac{\bar{\lambda}_i^{s,n}}{n}\right)\right) \frac{b_0\delta}{2n} + \left[\sum_i \frac{\delta}{n}\bar{\lambda}_i^{s,n} + o\left(\frac{\delta\sum_i \bar{\lambda}_i^{s,n}}{n}\right)\right] E_{\sigma_k^n}^{v,n}\xi_{\sigma_k^n}^{v,n}.$$

The $o(\cdot)$ will be ignored henceforth. Now, by A6.4 the term $E_{\sigma_k^n}^{v,n}\xi_{\sigma_k^n}^{v,n}$ in (6.13) goes to zero as $WL^n(\sigma_k^n)$ goes to infinity, which yields the stability, uniformly in n for large n.

Next let us consider the possibility that we might not always have $WL_i^n(t) \geq \bar{W}_a, i = 1, 2$, at the start of a vacation, when $WL^n(t) \geq \bar{W}_b$, for $\bar{W}_b \gg \bar{W}_a$, both being sufficiently large. Let $I_{t+\delta/n}^{v,n}$ denote the event that a vacation starts at scaled time $t + \delta/n$, with $WL_i^n(t) \leq \bar{W}_a$ for some i and $WL^n(t) \geq \bar{W}_b$. Let μ_l^n denote the scaled time of the lth occurrence of this event. We exploit the fact that this event is "rare"

for large n, by introducing another perturbation to the Liapunov function. We need to add the term

$$E^{s,n}_{\sigma^n_k} I^{v,n}_{\sigma^n_k+\delta/n} \tau^{v,n}_{\sigma^n_k} \tag{6.14}$$

to the right side of (6.13) (and multiply the current right-hand term there by $(1 - I^{v,n}_{\sigma^n_k+\delta/n})$, which leaves the estimates for that term unchanged. By A2.4, (6.14) is bounded by $C_1 E^{s,n}_{\sigma^n_k} I^{v,n}_{\sigma^n_k+\delta/n}$ for some constant C_1. Introduce the additional perturbation to the Liapunov function:

$$\delta \bar{W}^n_\beta(k\delta/n) = C_1 \sum_{l=k+1}^{\infty} e^{-(l-k-1)\beta\delta/n} E^{s,n}_{k\delta/n} I^{v,n}_{l\delta/n}. \tag{6.15}$$

This equals (dropping $o(\delta/n)$ terms for simplicity)

$$C_1 \frac{\delta \sum_i \bar{\lambda}^{s,n}_i}{n} \sum_{l:\mu^n_l > k\delta/n} E^{s,n}_{k\delta/n} e^{-\beta(\mu^n_l - k\delta/n - \delta/n)}. \tag{6.16}$$

Write the sum as $K^n_\beta(k)$. Then, for each $\beta > 0$, there is $n(\beta) < \infty$ such that $K^n_\beta(k) \le 2$ for $n \ge n(\beta)$ and all k.

Note the difference of the conditional expectations:

$$E^{s,n}_{\sigma^n_k} \delta \bar{W}^n_\beta(\sigma^n_{k+1}) - \delta \bar{W}^n_\beta(\sigma^n_k) = -C_1 E^{s,n}_{\sigma^n_k} I^{v,n}_{\sigma^n_k+\delta/n} + \epsilon^{v,n}_k, \tag{6.17}$$

where

$$\epsilon^{v,n}_k = C_1 \left[1 - e^{-\beta\delta/n}\right] \sum_{l=k+1}^{\infty} e^{-(l-k-2)\beta\delta/n} E^{s,n}_{k\delta/n} I^{v,n}_{l\delta/n}. \tag{6.18}$$

For $n \ge n(\beta)$,

$$\epsilon^{v,n}_k \le 2C_1 \beta \sum_i \frac{\bar{\lambda}^{s,n}_i}{n}.$$

Now, use the new perturbed Liapunov function defined by $W^n_\beta(k\delta/n) + \delta \bar{W}^n_\beta(k\delta/n)$. The conditional difference (6.17) cancels (6.14), modulo the error $\epsilon^{v,n}_k$, which is $O(\delta\beta/n)$ and β can be made as small as desired for large enough n. The proof is then completed as in Theorem 6.1 □

7. Unreliable channels. Up to now, it was supposed that any data sent from any source to the server arrived without error. In this section, we suppose that an error during transmission (as distinct from a vacation) is possible. Suppose that the server time is divided into "slots," of duration $\delta > 0$. That is, the work in each arrival is an integral multiple of δ, and each δ-interval is devoted to either a job from one of the sources, or to idling if there is no work present at the beginning of the slot. If a vacation starts during a slot, it is assumed that the data in that slot (if any) is retransmitted later. However, this has no effect on the heavy traffic limit. In case of an error in transmission, the data in the slot must be retransmitted. A variety of such situations can be readily incorporated into our general model.

If the sequence of errors is mutually independent in time, or if it does not depend on the source, then the modeling and analytical problem is relatively simple. The

errors would not depend on the source if the corrupting noise were at the server/base station, or was due to, say, a general atmospheric condition which affects all sources in the same way. On the other hand, if the errors are correlated (say, channels with bursty noise) and are *source dependent* as well, then the modeling problem is complicated by the fact that the server is allowed to poll the sources in a rather arbitrary way. For example, suppose that the noise is bursty for the channel connecting to one of the sources but that the channel connecting to the other is noise-free. Then, depending on how the sources are sequenced, the correlation between the errors can take many forms. Thus, it is hard to know the relation between the sequencing of the polling of the sources and the channel noise. One could try to poll taking into account the correlation. But, even if this were feasible, it is beyond our goals. For these reasons, we will assume that the disturbing noise does not depend on the source, despite the importance of the general problem.

The error model. Let $I_l^{e,n}$ denote the indicator function of the event that the data transmitted in the lth time slot was not acceptable and needed to be retransmitted. Let $S_i^{e,n}(t)$ (resp., $S^{e,n}(t)$) denote $1/n$ times the number of slots transmitted (successfully or not) from source i (resp., from both sources) by real time nt. Let $I_l^{a,n}$ denote the indicator function of the event that there is available data to be transmitted in time slot l from any source not on vacation. The (scaled) work that must be retransmitted by real time nt is

$$L^n(t) = \frac{\delta}{\sqrt{n}} \sum_{l=1}^{nt/\delta} I_l^{e,n} I_l^{a,n}. \tag{7.1}$$

For some centering constant $p^{e,n}$, write $L^n(t)$ as

$$L^n(t) = \frac{\delta}{\sqrt{n}} \sum_{l=1}^{nt/\delta} \left[I_l^{e,n} - p^{e,n}\right] I_l^{a,n} + \sqrt{n} p^{e,n} S^{e,n}(t)\delta. \tag{7.2}$$

The last term on the right of (7.2) is (see (3.2))

$$p^{e,n}\left[\sqrt{n}t - T^{v,n}(t) - Z^n(t)\right]. \tag{7.3}$$

Define

$$w^{e,n}(t) = \frac{\delta}{\sqrt{n}} \sum_{l=1}^{nt/\delta} \left[I_l^{e,n} - p^{e,n}\right] I_l^{a,n}. \tag{7.4}$$

We will use the following assumptions.

A7.1. *The error process is independent of all of the other driving processes. Also, $p^{e,n}$ converges to the constant p^e as $n \to \infty$ and $w^{e,n}(\cdot)$ converges weakly to a Wiener process, which will be denoted by $w^e(\cdot)$, and whose variance is $\delta\sigma_e^2$.*

A7.2. (The new heavy traffic condition.) *There is a constant b such that*

$$\lim_n \sqrt{n}\left[\sum_i \rho_i^n + p^{e,n} - 1\right] = b.$$

A7.1 is an assumption on the channel and will be returned to below. By adding

the work to be retransmitted, (3.15) becomes

$$WL^n(t) = WL^n(0) + \sum_i \bar{\Delta}_i^{d,n}\left[w_i^{a,n}(S_i^{a,n}(t)) - w_i^{d,n}(S_i^{a,n}(t))\right] + w^{e,n}(t)$$
$$+\sqrt{n}\left[\sum_i \rho_i^n + p^{e,n} - 1\right]t + (1-p^{e,n})\left[Z^n(t) + T^{v,n}(t)\right] + \epsilon^n(t), \tag{7.5}$$

where $\epsilon^n(\cdot)$ is a residual time error process.

Under the conditions of Theorem 3.1, with A7.1 added and the new heavy traffic condition A7.2 used, Theorem 3.1 continues to hold, with the following changes. The process $w^e(\cdot)$ is added to $w(\cdot)$. The jumps are computed by first showing that (in the local fluid time scale) the processes of completed work *during a vacation* can be asymptotically approximated by a fluid process with slope $1-p^e$, and they are

$$\begin{aligned}\xi_{1,l}^v &= \left[((1-p^e)-\rho_2)\,\tau_{1,l}^v - [WL(\nu_{1,l}-) - u(\nu_{1,l}-)]\right]^+ \\ &= \left[\rho_1 - [WL(\nu_{1,l}-) - u(\nu_{1,l}-)]\right]^+,\end{aligned} \tag{7.6a}$$

$$\xi_{2,l}^v = \left[((1-p^e)-\rho_1)\,\tau_{2,l}^v - u(\nu_{2,l}-)\right]^+ = \left[\rho_2\tau_{2,l}^v - u(\nu_{2,l}-)\right]^+. \tag{7.6b}$$

Also, $w^e(\cdot)$ is independent of $w(\cdot)$. With these changes, Theorem 3.2 also holds. Theorem 4.2 will continue to hold with these changes, provided that $E|w^{e,n}(t)|^2 = O(t)$. Similarly, the analogues of the stability results hold.

Remark. It is not possible to account for the retransmissions by simply enlarging the work in each job by an amount that has the same distribution as the retransmitted work does. This is because the controls are based on either the current queued work or queued numbers, and not what might be expected due to future errors and retransmissions.

Comments concerning $w^{e,n}(\cdot)$. First, suppose that the errors are independent from slot to slot with $P\{I_l^{e,n} = 1\} = p^{e,n}$. Then Donsker's theorem [12] implies that $w^{e,n}(\cdot)$ is tight and converges weakly to a Wiener process with variance $\delta p^e(1-p^e)$.

Now, turn to the correlated error problem. The error process concerns the channel, and is defined whether or not there is something to be transmitted. Suppose that the error process is Markov and doesn't depend on n, for notational simplicity. In particular, assume that

$$P\{I_{l+1}^{e,n} = 1 | I_l^{e,n} = 0\} = p,\ P\{I_{l+1}^{e,n} = 0 | I_l^{e,n} = 1\} = q,$$

where p and q are in $(0,1)$. Then $p^e = p/(p+q)$. Let I_l^e denote the stationary error process. Again, it is not hard to verify that $w^{e,n}(\cdot)$ converges weakly to a Wiener process with variance

$$\delta E\left[I_l^e - p^\epsilon\right]^2 + 2\delta E\sum_{l=1}^{\infty}\left[I_l^e - p^\epsilon\right]\left[I_0^e - p^\epsilon\right]$$

[12, 17].

We have $I_l^{a,n} = 0$ if both sources are on vacation, both queues are empty, or one source is on vacation and the other queue is empty at real time $l\delta$. These possibilities have negligible effect asymptotically.

Lévy processes. Many other models are possible for the error process (7.1) and a couple of other possibilities will be outlined. One approach, which does not require

the addition of A7.1 and uses (3.5) for the jumps, is to simply suppose that $L^n(\cdot)$ converges weakly to a general Lévy jump process. For example, suppose that the noise occurs in occasional bursts, where the rate at which the bursts occur (in real time) is $\bar{\lambda}^{e,n}/n$ and the duration (in real time) is $\sqrt{n}\tau_l^{e,n}$, where the durations (and the process of starting) are mutually independent and independent of the other "driving" random variables in the system. In this model the bursts are rare, and occur at a rate which is of the order of that of the vacations. But $\bar{\lambda}^{e,n}$ might be much larger than $\bar{\lambda}^{s,n}$ and the $\tau_l^{e,n}$ much smaller than $\tau_{i,l}^{v,n}$.

The scheme in the last paragraph supposed a finite rate $\bar{\lambda}^{e,n}$ for the bursts. The rate could depend on the duration, so that shorter durations have higher rates, with the rate going to infinity as the duration goes to zero, but in such a way that there is a limit Lévy process.

REFERENCES

[1] E. Altman, D. Kofman, and U. Yechiali, *Discrete time queues with delayed information*, Queueing Systems, 19 (1995), pp. 361–376.

[2] E. Altman and H. J. Kushner, *Admission control for combined guaranteed performance and best effort communications systems under heavy traffic*, SIAM J. Control Optim., 37 (1999), pp. 1780–1807.

[3] E. Altman and A. Shwartz, *Optimal priority assignment: A time sharing approach*, IEEE Trans. Automat. Control, 34 (1989), pp. 1098–1102.

[4] J. S. Baras, D. J. Ma, and A. M. Makowski, *K competing queues with geometric service requirements and linear costs: The μc rule is always optimal*, Systems Control Lett., 6 (1985), pp. 173–180.

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

[6] G. Blankenship and G. C. Papanicolaou, *Stability and control of systems with wide-band noise disturbances*. I, SIAM J. Appl. Math., 34 (1978), pp. 437–476.

[7] M. Bramson, *State space collapse with application to heavy traffic limits for multiclass queueing networks*, Queueing Systems, 30 (1999), pp. 89–148.

[8] M. Carr and B. Hajek, *Scheduling with asynchronous service opportunities with applications to multiple satellite systems*, IEEE Trans. Automat. Control, 38 (1993), pp. 1820–1833.

[9] D. R. Cox and W. L. Smith, *Queues*, Methuen, London, 1961.

[10] P. Dupuis and H. Ishii, *On Lipschitz continuity of the solution mapping to the Skorokhod problem, with applications*, Stochastics Stochastics Rep., 35 (1991), pp. 31–62.

[11] B. R. Elbert, *The Satellite Communication Applications Handbook*, Artech House, Boston, London, 1997.

[12] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*, Wiley, New York, 1986.

[13] S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes*, 2nd ed., Academic Press, New York, 1975.

[14] R. Z. Khasminskii, *Stochastic Stability of Differential Equations*, Sijthoff, Noordhoff, Alphen aan den Rijn, Amsterdam, 1982.

[15] H. Kunita and S. Watanabe, *On square integrable martingales*, Nagoya Math. J., 30 (1967), pp. 209–245.

[16] H. J. Kushner, *Stochastic Stability and Control*, Academic Press, New York, 1967.

[17] H. J. Kushner, *Approximation and Weak Convergence Methods for Random Processes with Applications to Stochastic Systems Theory*, MIT Press, Cambridge, MA, 1984.

[18] H. J. Kushner, *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, Systems and Control, Vol. 3, Birkhäuser, Boston, 1990.

[19] H. J. Kushner and P. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, 2nd ed., Springer-Verlag, Berlin, New York, 2001.

[20] H. J. Kushner, D. Jarvis, and J. Yang, *Controlled and optimally controlled multiplexing systems: A numerical exploration*, Queueing Systems, 20 (1995), pp. 255–291.

[21] H. J. Kushner and K. M. Ramachandran, *Optimal and approximately optimal control policies for queues in heavy traffic*, SIAM J. Control Optim., 27 (1989), pp. 1293–1318.

[22] H. J. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, Berlin, New York, 1997.

[23] R. Liptser and A. N. Shiryaev, *Statistics of Random Processes*, Springer-Verlag, Berlin, New York, 1977.
[24] W. P. Peterson, *Diffusion approximations for networks of queues with multiple customer types*, Math. Oper. Res., 9 (1951), pp. 90–118.
[25] M. I. Reiman, *Some diffusion approximations with state space collapse*, in Proceedings of the Int. Seminar on Modelling and Performance Evaluation Methodology, Paris, 1983, F. Baccelli and G. Fayolle, eds., Springer-Verlag, New York, 1983, pp. 209–240.
[26] M. I. Reiman, *A multiclass feedback queue in heavy traffic*, Adv. Appl. Probab., 20 (1998), pp. 179–207.
[27] V. Solo and X. Kong, *Adaptive Signal Processing Algorithms*, Prentice–Hall, Englewood Cliffs, NJ, 1995.
[28] L. Tassiulas and A. Ephremides, *Dynamic server allocation to parallel queues with randomly varying connectivity*, IEEE Trans. Automat. Control, 39 (1993), pp. 466–478.
[29] L. Tassiulas and S. Papavassiliou, *Optimal anticipative scheduling with asynchronous transmission opportunities*, IEEE Trans. Automat. Control, 40 (1995), pp. 2052–2062.
[30] J. A. van Mieghem, *Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule*, Ann. Appl. Probab., 5 (1995), pp. 809–833.
[31] J. Walrand, *An Introduction to Queuing Networks*, Prentice–Hall, Englewood Cliffs, NJ, 1988.
[32] R. J. Williams, *Diffusion approximations for open multiclass queueing networks: Sufficient conditions involving state space collapse*, Queueing Systems, 30 (1998), pp. 27–88.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 1, pp. 253–277

DUALITY IN $\mathbf{H}^\infty$ CONE OPTIMIZATION*

ANDREY GHULCHAK† AND ANDERS RANTZER†

Abstract. Positive real cones in the space $\mathbf{H}^\infty$ appear naturally in many optimization problems of control theory and signal processing. Although such problems can be solved by finite-dimensional approximations (e.g., Ritz projection), all such approximations are conservative, providing one-sided bounds for the optimal value. In order to obtain both upper and lower bounds of the optimal value, a dual problem approach is developed in this paper. A finite-dimensional approximation of the dual problem gives the opposite bound for the optimal value. Thus, by combining the primal and dual problems, a suboptimal solution to the original problem can be found with any required accuracy.

Key words. quasi-convex optimization, $\mathbf{H}^\infty$ space, convex duality

AMS subject classifications. 90C25, 32A35, 46A20

PII. S0363012900369617

1. Introduction.

Many analysis and synthesis problems of control theory have recently been stated in terms of convex optimization [3, 18, 22, 24]. This often gives great benefit for both theoretical analysis and practical computation. However, when an optimization problem is infinite-dimensional, a reduction to finite-dimensional form (like Ritz projection, Galerkin finite element scheme, grid methods, etc.) is needed [3]. Such an approximation introduces conservatism to the problem. The gap between the true optimal value and its finite-dimensional counterpart can be arbitrarily large. To overcome this difficulty and to estimate the conservatism, convex duality has been widely used [3, 4, 8, 11, 12, 16, 17, 24]. Once the dual problem is stated, a similar finite-dimensional scheme can be applied to obtain an opposite bound for the optimal value. Provided that there is no duality gap, a suboptimal solution with any degree of accuracy can be computed by increasing the dimension of the approximations.

In this paper, we consider a cone optimization in the Hardy space $\mathbf{H}^\infty$, by which we mean the optimization of a quasi-convex functional on $\mathbf{H}^\infty$ whose level sets are positive real cones $\{h \in \mathbf{H}^\infty : \mathrm{Re}\,(g^T h) > 0,\ g \in \mathbf{H}^\infty\}$, or the intersection of such cones. Problems of this type appear often in controller design—for example, in $\mathbf{H}^\infty$ control, in robust stabilization under parametric uncertainty, in output error identification, etc. The strong practical aspect has become the main motivation of our work.

The main contribution of the paper is a theoretical development of the duality relation for this cone optimization. An analytical expression of the infinite-dimensional dual problem is obtained. Similarly to the primal problem, it takes the form of quasi-convex optimization. A cornerstone of our approach is convex duality, namely, Ky Fan's min-max theorem [25]. It is shown that there is no duality gap, and the primal-dual method can be used to obtain a suboptimal solution with any predefined level of optimality.

The paper is organized as follows. Section 2 introduces all major notation used throughout the paper. In section 3, the cone optimization problem in $\mathbf{H}^\infty$ is described. Several examples of interesting control problems that can be reduced to such a cone

*Received by the editors March 8, 2000; accepted for publication (in revised form) January 17, 2002; published electronically June 5, 2002.
http://www.siam.org/journals/sicon/41-1/36961.html
†Department of Automatic Control, Lund Institute of Technology, Box 118, 221 00 Lund, Sweden (ghulchak@control.lth.se, rantzer@control.lth.se).

optimization are gathered in section 4. The main result, which is the dual form of the problem, is presented in section 5, followed by a discussion of some particular cases in section 6. Section 7 compares the main result with duality by Megretski and Rantzer [17]. A brief discussion on possible numerical realizations of the primal-dual method is presented in section 8. Finally, a numerical example is given in section 9, where the primal-dual method is applied to a nonstandard $\mathbf{H}^\infty$ optimization problem. All proofs are moved to the appendices.

2. Notation. By $\mathbb{R}$ (or $\mathbb{C}$) we denote the field of real (or complex) numbers. The subset of $\mathbb{R}$ of nonnegative numbers is denoted by $\mathbb{R}_+$. The unit circle and the open unit disc in $\mathbb{C}$ are denoted by $\mathbb{T}$ and $\mathbb{D}$, respectively:

$$\mathbb{T} = \{z \in \mathbb{C} \;\big|\; |z| = 1\}, \quad \mathbb{D} = \{z \in \mathbb{C} \;\big|\; |z| < 1\}.$$

For every measurable $Y \subset \mathbb{C}^n$, the notation $\mathbf{L}^p(Y)$ stands for the standard Lebesgue space of functions $f\colon \mathbb{T} \to Y$ equipped with the norm

$$\|f\|_p = \begin{cases} \left(\int_{\mathbb{T}} |f(z)|^p \, dm(z)\right)^{1/p}, & 1 \le p < +\infty, \\ \operatorname{ess\,sup}_{z\in\mathbb{T}} |f(z)|, & p = +\infty, \end{cases}$$

where, by $|\cdot|$, we denote the usual Hölder 2-norm in $\mathbb{C}^n$

$$|f| = \sqrt{|f_1|^2 + |f_2|^2 + \cdots + |f_n|^2}.$$

$\mathbf{H}^p(Y)$ denotes the Hardy space of functions in $\mathbf{L}^p(Y)$ that have an analytical continuation inside the unit disc. $\mathbf{H}^p_0(Y)$ denotes the shifted $\mathbf{H}^p(Y)$; that is,

$$\mathbf{H}^p_0(Y) = z\mathbf{H}^p(Y) = \{f \in \mathbf{H}^p(Y) \;\big|\; f(0) = 0\}.$$

We use the notation $\mathbf{RH}^\infty(Y)$ for the set of all functions from $\mathbf{H}^\infty(Y)$ that are rational with real coefficients. The space of all continuous functions $f\colon \mathbb{T} \to Y$ is denoted by $C(Y)$. The notation $\mathbf{A}(Y)$ stands for the Banach disc algebra $\mathbf{H}^\infty(Y) \cap C(Y)$.

The short notations $\mathbf{L}^p$, $\mathbf{H}^p$, etc. will be used if $Y = \mathbb{C}^n$ and the dimension n is clear from context or makes no difference for presentation.

The orthogonal projection from L^2 to H^2 is denoted by P_+, and $P_- = I - P_+$. The prefix $\mathcal{B}$ denotes the unit ball in the corresponding space, and $\mathcal{S}$ is the unit sphere. The superscript ^{T} stands for transposition. Re is the real part of a complex number. A bar over a function denotes the complex conjugate. For two sets A and B,

$$A \setminus B = \{a \in A \;\big|\; a \notin B\}.$$

3. Preliminaries. Let $F \in \mathbf{A}(\mathbb{C}^{1\times n})$ and $G \in \mathbf{A}(\mathbb{C}^{m\times n})$. Denote

$$\Phi_\delta := F + \delta^T G, \tag{3.1}$$

$$J_\delta(h, z) := \operatorname{Re} \Phi_\delta(z) h(z). \tag{3.2}$$

A number of control design problems (see examples in section 4) can be stated as a convex specification

$$J_\delta(h, z) > 0 \quad \forall z \in \mathbb{T}, \ \forall \delta \in \Delta, \tag{3.3}$$

where the function $h \in \mathbf{RH}^\infty(\mathbb{C}^{n\times 1})$ is the design parameter (in control applications, it is usually the free parameter from the Youla parameterization of all stabilizing

controllers), and the set $\Delta \subset \mathbb{C}^m$ is the uncertainty region. The cone optimization problem is to find a (δ-independent) function $h \in \mathbf{RH}^\infty$ that satisfies the condition (3.3) for as large a region Δ as possible.

The condition (3.3) is a linear inequality with respect to δ. This implies that, if it holds for some set Δ, then it holds for the convex hull of the set. Therefore, without loss of generality, it is sufficient to consider the convex Δ's only. In addition, we assume Δ to be compact.

Assumption. The region Δ is assumed to be a compact convex set in $\mathbb{C}^m$.

To define how "large" or "small" the region is, we assume that the regions are described by a monotone family $\{\Delta_\nu\}$ so that the "size" ν is assigned to each set Δ_ν of the family. The monotonicity means that the larger set has the larger size; i.e., if $\nu_1 \le \nu_2$, then $\Delta_{\nu_1} \subset \Delta_{\nu_2}$. A common particular case of the monotonic family is the linear homotopy $\Delta_\nu = \nu\Delta$ of a set $\Delta \subset \mathbb{C}^m$ containing the origin.

Thus the problem is to maximize the size ν as follows:

$$\nu_{opt} = \sup\{\nu \mid \exists h \in \mathbf{RH}^\infty \colon J_\delta(h, z) > 0 \quad \forall z \in \mathbb{T},\ \forall \delta \in \Delta_\nu\}. \tag{3.4}$$

It is called the cone optimization problem since the set

$$K_\nu = \{h \in \mathbf{RH}^\infty \mid J_\delta(h, z) > 0 \quad \forall z \in \mathbb{T},\ \forall \delta \in \Delta_\nu\} \tag{3.5}$$

is a convex cone in $\mathbf{H}^\infty$. Thus, for a given ν, we have a cone programming (feasibility) problem which is related to robust linear programming [4] in the following sense: the linear inequality (3.3) should be satisfied *robustly* with respect to Δ.

4. Examples of $\mathbf{H}^\infty$ cone optimization problems. In this section, we gather several problems in control theory that can be reduced to the cone optimization in $\mathbf{H}^\infty$.

In all examples, the homotopy has the form $\Delta_\nu = \nu\Delta$ for a convex compact set $\Delta \subset \mathbb{C}^m$ containing the origin.

The following theorem is useful for reducing a problem to the cone optimization. It is a slightly modified version of [22, Theorem 1].

THEOREM 4.1. *Let $g \in C(\mathbb{C}^{m\times 1})$, and $\Delta \ni 0$ is a convex set in $\mathbb{C}^m$. The following statements are equivalent:*

1. $1 + \delta^T g(z) \ne 0$ *for all* $z \in \mathbb{T}$ *and* $\delta \in \Delta$.
2. *There exists a function* $\alpha \in \mathbf{RH}^\infty$ *such that*

$$\operatorname{Re}(1 + \delta^T g(z))\alpha(z) > 0 \quad \forall z \in \mathbb{T},\ \delta \in \Delta.$$

Example 1: $\mathbf{H}^\infty$ optimization. Given a $(N_z + N_y) \times (N_w + N_u)$ plant P

$$\begin{pmatrix} z \\ y \end{pmatrix} = P \begin{pmatrix} w \\ u \end{pmatrix},$$

the problem is to find a stabilizing controller $u = Ky$ that minimizes the $\mathbf{H}^\infty$-norm of the closed-loop transfer function T_{zw}. If the disturbance w is scalar ($N_w = 1$), then the Youla parameterization of all admissible closed-loop transfer functions (see, for example, [6]) has the form

$$T_{zw} = T_1 + T_2 Q, \tag{4.1}$$

where $T_1 \in \mathbf{A}(\mathbb{C}^{N_z})$ and $T_2 \in \mathbf{A}(\mathbb{C}^{N_z \times N_u})$ are defined by the plant P, and Q is any function in $\mathbf{RH}^\infty(\mathbb{C}^{N_u})$. With this parameterization, the problem can be rewritten as

$$\min_{Q \in \mathbf{RH}^\infty} \|T_1 + T_2 Q\|_\infty.$$

Equivalently, the problem is to maximize ν such that at all frequencies

$$|T_1(z) + T_2(z)Q(z)| < \nu^{-1} \quad \forall z \in \mathbb{T}. \tag{4.2}$$

For real-rational plants P, the well-developed theory [6] can be applied to solve the problem. However, in some important cases where P is not rational (e.g., systems with delays), there is no analytical solution to the problem in general.

The inequality (4.2) can be rewritten as

$$1 + \delta^T(T_1 + T_2 Q) \neq 0 \quad \forall z \in \mathbb{T},\ \delta \in \nu\Delta,$$

where $\Delta = \mathcal{B}\mathbb{C}^{N_z}$, and Theorem 4.1 reduces it to

$$\operatorname{Re}\left(1 + \delta^T(T_1(z) + T_2(z)Q(z))\right)\alpha(z) > 0 \quad \forall z \in \mathbb{T},\ \delta \in \nu\Delta. \tag{4.3}$$

It takes the form (3.3) with

$$\begin{aligned} F &= \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix} \in \mathbb{R}^{N_u+1}, \\ G &= \begin{pmatrix} T_1 & T_2 \end{pmatrix} \in \mathbf{A}, \\ h &= \begin{pmatrix} \alpha \\ Q\alpha \end{pmatrix} \in \mathbf{RH}^\infty. \end{aligned} \tag{4.4}$$

Setting $\delta = 0$ in (4.3) gives $\operatorname{Re}\alpha > 0$. Therefore, $1/\alpha \in \mathbf{RH}^\infty$, and the solution $Q \in \mathbf{RH}^\infty$ can be reconstructed from h as

$$Q = \frac{1}{h_1} \begin{pmatrix} h_2 \\ h_3 \\ \vdots \\ h_{N_u+1} \end{pmatrix}. \tag{4.5}$$

Example 2: Robust stabilization. Given a $(N_z + N_y) \times (N_w + N_u)$ uncertain plant P

$$\begin{pmatrix} y \\ z \end{pmatrix} = P \begin{pmatrix} w \\ u \end{pmatrix},$$
$$w = \delta^T z,$$

where $\delta \in \nu\Delta$ for some convex compact $\Delta \subset \mathbb{R}^{N_z}$ with $0 \in \Delta$, the problem is to find a controller $u = Ky$ that robustly stabilizes the plant for as large a ν as possible.

In the case where $N_w = 1$, the same Youla parameterization (4.1) gives the following equivalent problem: Find a function $Q \in \mathbf{RH}^\infty(\mathbb{C}^{N_u})$ that maximizes ν subject to

$$1 + \delta^T(T_1(s) + T_2(s)Q(s)) \neq 0 \quad \forall s\colon \operatorname{Re} s \geq 0,\ \forall \delta \in \nu\Delta.$$

Again by Theorem 4.1, this problem can be reduced to the form (3.3), with F and G defined in (4.4). The function Q can be reconstructed from h as in (4.5).

Example 3: Adaptive output error identification. Given a set of stable scalar polynomials

$$\mathcal{P}_\nu = \{z^m + p_1 z^{m-1} + \cdots + p_m \mid |p_i - p_i^0| \le \nu\epsilon_i\},$$

find a rational function b such that $1/b \in \mathbf{RH}^\infty$ and

$$\operatorname{Re}\frac{p(z)}{b(z)} > 0 \quad \forall z \in \mathbb{T},\ \forall p \in \mathcal{P}_\nu,$$

for as large a ν as possible.

The problem appears when a gradient algorithm is applied to output error identification of a plant transfer function with denominator polynomial $p(z)$. To ensure exponential convergence of the identification algorithm, certain signals are supposed to be filtered by a transfer function $1/b$, where p/b is strictly positive real [1].

The problem takes the form (3.3) if we denote $h = 1/b$, $F = p^0(z)$, $G = (z^{m-1}, z^{m-2}, \dots, 1)^T$, and

$$\Delta = \{\delta \in \mathbb{R}^m \mid \delta_i \in [-\epsilon_i, \epsilon_i]\}.$$

5. The problem and the main result. The problem (3.4) is a problem of quasi-convex optimization; for a given ν, the cone K_ν from (3.5) is convex. We will refer to the problem of finding $h \in K_\nu$ as the *primal problem*.

Primal problem. Given $F \in \mathbf{A}(\mathbb{C}^{1\times n})$, $G \in \mathbf{A}(\mathbb{C}^{m\times n})$, and a convex compact set $\Delta_\nu \subset \mathbb{C}^m$, find a function $h \in \mathbf{RH}^\infty(\mathbb{C}^{n\times 1})$ such that

$$J_\delta(h,z) := \operatorname{Re}\,(F(z) + \delta^T G(z))h(z) > 0 \quad \forall z \in \mathbb{T},\ \forall \delta \in \Delta_\nu. \tag{5.1}$$

Since Δ_ν and $\mathbb{T}$ are compact and all functions in (5.1) are continuous, the primal problem takes an equivalent form given by the following proposition.

PROPOSITION 5.1. *Given $\nu \in \mathbb{R}_+$, the following statements are equivalent:*

1. *The primal problem has a solution, i.e., $\nu < \nu_{opt}$ with ν_{opt} defined in* (3.4).
2. $$\gamma_{opt}(\nu) := \sup_{h\in\mathcal{B}\mathbf{A}} \inf_{z\in\mathbb{T}} \inf_{\delta\in\Delta_\nu} J_\delta(h,z) > 0. \tag{5.2}$$

Proof. See Appendix A. □

Remark 1. Note that $\gamma_{opt}(\nu) \ge 0$ even if $\nu \ge \nu_{opt}$. To see this, put $h = 0$.

Remark 2. The choice of the unit ball in $\mathbf{A}$ as an optimization set in (5.2) is not essential. Due to linear dependence of J_δ on h, any bounded set with the closure that absorbs $\mathbf{A}$ could be chosen. However, the unit ball has better relations to and easier interpretations from the classical results, such as the Banach duality, for example, than any other set (see Appendix B).

To obtain a dual representation of (5.2), we have to slightly modify the problem. The following lemma states that γ_{opt} can be calculated via optimization over the larger set $\mathcal{B}\mathbf{H}^\infty$. This is a very useful property for our technique, since $\mathbf{H}^\infty$ is dual to a "nice" space. This will play a central role in the dual presentation.

LEMMA 5.2.

$$\gamma_{opt}(\nu) = \sup_{h\in\mathcal{B}\mathbf{H}^\infty} \operatorname*{ess\,inf}_{z\in\mathbb{T}} \inf_{\delta\in\Delta_\nu} J_\delta(h,z). \tag{5.3}$$

Proof. See Appendix A. ∎

Now we are in a position to present the main result on duality.

THEOREM 5.3 (duality). *Let $F \in \mathbf{H}^\infty(\mathbb{C}^{1\times n})$, $G \in \mathbf{H}^\infty(\mathbb{C}^{m\times n})$, and let $\Delta_\nu \subset \mathbb{C}^m$ be a convex compact set. Denote Φ_δ as in* (3.1). *Then the following equality holds:*

(5.4)
$$\sup_{h\in\mathcal{B}\mathbf{H}^\infty(\mathbb{C}^{n\times 1})} \operatorname*{ess\,inf}_{z\in\mathbb{T}} \inf_{\delta\in\Delta_\nu} \operatorname{Re}\Phi_\delta(z)h(z) = \inf_{\delta\in\mathbf{L}^\infty(\Delta_\nu)} \inf_{w\in\mathcal{S}\mathbf{L}^1(\mathbb{R}_+)} \inf_{p\in\mathbf{H}_0^1(\mathbb{C}^{1\times n})} \|\Phi_\delta w - p\|_1.$$

Proof. See Appendix B. ∎

Remark 1. The left-hand side of the equality is $\gamma_{opt}(\nu)$ from (5.3). Theorem 5.3 gives the dual representation of the quantity as the minimization problem. The problem provides an upper bound on $\gamma_{opt}(\nu)$ by which we can determine the case when $\gamma_{opt}(\nu) = 0$.

Remark 2. The dual condition (5.4) can be rewritten in a convex way by setting a new variable $x = \delta w$:

$$\begin{aligned}\gamma_{opt}(\nu) = \inf\{\|Fw + x^TG - p\|_1 \;\big|\; p \in \mathbf{H}_0^1,\ \|w\|_1 = 1,\\ w \geq 0,\ x(z) \in w(z)\Delta_\nu \quad \forall z \in \mathbb{T}\}.\end{aligned} \tag{5.5}$$

Using Theorem 5.3, we state the *dual problem* to (5.1) as follows.

Dual problem. Given $F \in \mathbf{A}(\mathbb{C}^{1\times n})$, $G \in \mathbf{A}(\mathbb{C}^{m\times n})$, and a convex compact set $\Delta_\nu \subset \mathbb{C}^m$, find a sequence of functions $\{(w_i, \delta_i, p_i)\}_{i=0}^{+\infty}$ such that $w_i \in \mathcal{S}\mathbf{L}^1(\mathbb{R}_+)$, $\delta_i \in \mathbf{L}^\infty(\Delta_\nu)$, $p_i \in \mathbf{H}_0^1(\mathbb{C}^{1\times n})$, and

$$\|(F + \delta_i^TG)w_i - p_i\|_1 \to 0, \quad i \to +\infty. \tag{5.6}$$

We can draw two obvious conclusions from Theorem 5.3.

COROLLARY 5.4. *A number ν is an upper bound on ν_{opt} if and only if the dual problem has a solution.*

Proof. By Theorem 5.3, the dual problem has a solution if and only if $\gamma_{opt}(\nu) = 0$, which is equivalent to $\nu \geq \nu_{opt}$ by Proposition 5.1. ∎

COROLLARY 5.5. *If there exist $w \in \mathbf{L}^1(\mathbb{R}_+) \setminus 0$ and $\delta \in \mathbf{L}^\infty(\Delta_\nu)$ such that $\Phi_\delta w \in \mathbf{H}_0^1$, then $\nu \geq \nu_{opt}$.*

Proof. Scale by $\|w\|_1$, and apply Theorem 5.3 to conclude that $\gamma_{opt}(\nu) = 0$. ∎

Let us point out the main idea behind the equality (5.4). We will see that, if $\Phi_\delta w \in \mathbf{H}_0^1$ for some $w \in \mathbf{L}^1(\mathbb{R}_+)\setminus 0$ and $\delta \in \mathbf{L}^\infty(\Delta_\nu)$, then, in fact, $\gamma_{opt}(\nu) = 0$. For a function $h \in \mathbf{H}^\infty$, we have $f := \Phi_\delta wh \in \mathbf{H}_0^1$, so the mean value property for harmonic functions [27] gives

$$\int_{\mathbb{T}} \operatorname{Re}\left[(F(z) + \delta(z)^TG(z))h(z)\right]w(z)\,dm(z) = \operatorname{Re} f(0) = 0.$$

Therefore, since $w \geq 0$ and $w \not\equiv 0$, there must exist $z \in \mathbb{T}$ such that $\operatorname{Re}(F(z) + \delta(z)^TG(z))h(z) \leq 0$, which contradicts the condition (5.1).

The dual problem in the form (5.6) is complicated because it is stated in terms of *sequences* of functions. The problem would be much simpler if it were possible to replace the sequences with their limit points. The main difficulty here is a limit point for $\{w_i\}$. It may not exist as an element of $\mathcal{S}\mathbf{L}^1(R_+)$. However, the limit point for $\{w_i\}$ exists either as a regular function in $\mathcal{S}L^1(\mathbb{R}_+)$ or as Dirac's δ-function. Based on this, the dual problem can be naturally split into two parts—one regular

and one singular, with no functional sequences left in any of them. The regular part has already been covered by Corollary 5.5. The simplification of the dual problem is completed by the singular part in the next theorem.

THEOREM 5.6. *The optimal value ν_{opt} of the cone optimization problem* (3.4) *has the dual representation*

$$\nu_{opt} = \min\{\, \nu_{opt|s},\ \nu_{opt|c}\,\}, \tag{5.7}$$

where

$$\nu_{opt|s} = \inf\{\nu \;\big|\; \exists z \in \mathbb{T},\ \exists \delta \in \Delta_\nu, : \Phi_\delta(z) = 0\}, \tag{5.8}$$

$$\nu_{opt|c} = \inf\{\nu \;\big|\; \exists w \in \mathbf{L}^1(\mathbb{R}_+) \setminus 0,\ \exists \delta \in \mathbf{L}^\infty(\Delta_\nu) \colon \Phi_\delta w \in \mathbf{H}^1_0\}. \tag{5.9}$$

Proof. See Appendix C. □

Remark 1. It can be seen that $\nu_{opt|s}$ is the optimal value of the problem (3.4) *without* an analyticity condition on h, i.e., when we replace the space $\mathbf{H}^\infty$ with $\mathbf{L}^\infty$.

COROLLARY 5.7. *The bound $\nu_{opt|c}$ can be alternatively represented as*

$$\nu_{opt|c} = \inf\{\nu \;\big|\; \exists g \in (\mathbf{H}^2)^\perp \setminus 0,\ \exists \delta \in \mathbf{L}^\infty(\Delta_\nu) \colon P_-(\Phi_\delta g) = 0\}. \tag{5.10}$$

Proof. The proof follows easily from the factorization $w = f^* f$ with the outer factor $f \in \mathbf{H}^2$, dividing both sides of $\Phi_\delta w \in \mathbf{H}^1_0$ by f and setting $g(z) = f^*(z)/z$. □

The number $\nu_{opt|s}$ is relatively easy to evaluate:

$$\nu_{opt|s} = \inf_{z \in \mathbb{T}} \nu_s(z), \tag{5.11}$$

where

$$\nu_s(z) = \inf\{\nu \;\big|\; \delta \in \Delta_\nu,\ F(z) + \delta^T G(z) = 0\}; \tag{5.12}$$

that is, (5.8) is the convex problem at every $z \in \mathbb{T}$.

Dealing with $\nu_{opt|c}$ is a bit more complicated because it requires optimization on one more infinite-dimensional parameter in $\mathbf{H}^1_0$. Denoting $x(z) = w(z)\delta(z)$, we have

$$\delta(z) \in \Delta_\nu \quad \Leftrightarrow \quad x(z) \in w(z)\Delta_\nu;$$

hence (5.9) becomes

$$\nu_{opt|c} = \inf\{\nu \;\big|\; \exists w \in \mathbf{L}^1(\mathbb{R}_+) \setminus 0,\ \exists x \colon x(z) \in w(z)\Delta_\nu,\ wF + x^T G \in \mathbf{H}^1_0\}. \tag{5.13}$$

This is a quasi-convex optimization problem. It can be used to obtain an upper bound on $\nu_{opt|c}$ in the following way.

LEMMA 5.8. *The number $\nu \in \mathbb{R}_+$ is the upper bound on $\nu_{opt|c}$ if and only if there exists a solution $(x, w, p) \in \mathbf{L}^1(\mathbb{C}^m) \times \mathbf{L}^1(\mathbb{R}_+) \times \mathbf{H}^1_0(\mathbb{C}^n)$ to the problem*

$$\begin{array}{rcl} p(z) & = & w(z)F(z) + x(z)^T G(z), \\ x(z) & \in & w(z)\Delta_\nu, \\ \|w\|_1 & > & 0. \end{array} \tag{5.14}$$

Proof. The proof follows immediately from (5.13). □

Remark. The last condition $\|w\|_1 > 0$ is needed only to avoid the trivial solution $(x, w, p) = 0$. It can be replaced with $\|w\|_1 = 1$.

The condition (5.14) is linear in (w, p) and convex in x.

To conclude the section, let us give an explicit explanation of how the optimization in (3.4) can be done by using the dual problem. First, it is relatively easy to estimate $\nu_{opt|s}$ from (5.11) by sweeping the unit circle and solving the convex problem (5.12). When it is done, the search for ν_{opt} can be organized as a bisection of $[0, \nu_{opt|s}]$, using the duality result to choose the right part of the interval; namely, for a given $\nu \in [0, \nu_{opt|s}]$, one and only one problem of (5.1) and (5.14) has a solution.

6. The primal and dual problems for some particular sets $\boldsymbol{\Delta}$. To get better insight into the cone optimization problem, we consider in detail several typical cases. We assume, for simplicity, that the family Δ_ν is the linear homotopy of a convex compact set $\Delta \ni 0$, i.e., $\Delta_\nu = \nu\Delta$.

Case 1. $\Delta = 0$.

The primal problem (5.1) is to find a function $h \in \mathbf{RH}^\infty(\mathbb{C}^{n\times 1})$ such that

$$\operatorname{Re} F(z)h(z) > 0 \quad \forall z \in \mathbb{T},$$

for given $F \in \mathbf{A}(\mathbb{C}^{1\times n})$. The dual representation for γ_{opt} in Theorem 5.3 simplifies to

$$\gamma_{opt} := \sup_{h\in\mathcal{B}\mathbf{H}^\infty} \operatorname*{ess\,inf}_{z\in\mathbb{T}} \operatorname{Re} F(z)h(z) = \inf_{w\in\mathcal{S}\mathbf{L}^1(\mathbb{R}_+)} \inf_{p\in\mathbf{H}^1_0} \|Fw - p\|_1,$$

and the duality result in Theorem 5.6 claims that the primal problem has no solution if and only if

1. $\exists w \in \mathbf{L}^1(\mathbb{R}_+) \setminus 0$ such that $Fw \in \mathbf{H}^1_0$, or
2. $\exists z \in \mathbb{T}$ such that $|F(z)| = 0$.

The second condition is the absence of zeros of F on the unit circle. The following proposition shows that the first one is related to that in the open unit disc.

PROPOSITION 6.1. *Let $F \in \mathbf{A}(\mathbb{C}^{1\times n})$. The following conditions are equivalent:*

1. $\exists\lambda \in \mathbb{D}$ *such that* $|F(\lambda)| = 0$.
2. $\exists w \in \mathbf{L}^1(\mathbb{R}_+) \setminus 0$ *such that* $Fw \in \mathbf{H}^1_0(\mathbb{C}^{1\times n})$.

Proof. See Appendix D. □

Note that the existence of $h \in \mathbf{A}$ such that $\operatorname{Re} F(z)h(z) > 0$ for all $z \in \mathbb{T}$ is equivalent to the existence of $g \in \mathbf{A}$ such that $Fg = 1$ (if we just set $g = h(Fh)^{-1}$). Thus the duality theorem in the particular case when $\Delta = 0$ gives the well-known result [27] concerning Gelfand's theory of maximal ideals in disc algebra $\mathbf{A}$:

$$\exists g \in \mathbf{A}\colon Fg = 1 \quad \Leftrightarrow \quad \inf_{\lambda\in\mathbb{D}} |F(\lambda)| > 0.$$

Let us now give some interpretation to the primal and dual problems in the case when $\Delta = 0$. Since the uncertainty set is zero, the primal problem can be interpreted as a nominal stabilization problem. For example, in the scalar case, a controller $K = \beta/\alpha$ stabilizes a plant $P = b/a$ if and only if the characteristic polynomial of the closed-loop system $\chi(s) = a(s)\alpha(s) - b(s)\beta(s)$ is stable. Equivalently, there exists $h \in \mathbf{A}(\mathbb{C}^{2\times 1})$ such that

$$\operatorname{Re} \begin{pmatrix} a(z) & -b(z) \end{pmatrix} h(z) > 0 \quad \forall z \in \mathbb{T}. \tag{6.1}$$

Indeed, one possible choice of h is $(\alpha\ \beta)^T/\chi$.

The dual problem gives a stabilizability criterion as $|F(\lambda)| \neq 0$ for all $\lambda \in \mathbb{D}$. In our scalar case example (6.1), the stabilizability criterion gives the following well-known condition: a plant $P = b/a$ is stabilizable if and only if the polynomials a and b have no common unstable zeros.

Case 2. $\Delta = \mathcal{B}\mathbb{C}^m$.

We will show that this case is reduced to the standard $\mathbf{H}^\infty$ optimization. The first step in this direction is the following proposition.

PROPOSITION 6.2. *Let $F \in \mathbf{A}(\mathbb{C}^{1\times n})$, $G \in \mathbf{A}(\mathbb{C}^{m\times n})$, and $\Delta = \mathcal{B}C^m$. Then the following statements are equivalent:*

1. *$\exists h \in \mathbf{A}(\mathbb{C}^{n\times 1})$ such that*

$$\operatorname{Re}\,(F(z) + \delta^T G(z))h(z) > 0 \quad \forall z \in \mathbb{T}, \quad \forall \delta \in \nu\Delta. \tag{6.2}$$

2. *$\exists g \in \mathbf{A}(\mathbb{C}^{n\times 1})$ such that $Fg = 1$ and $\|Gg\|_\infty < \nu^{-1}$.*

Proof. See Appendix D. □

Proposition 6.2 reduces the cone optimization problem (3.4) to the convex optimization problem

$$\nu_{opt}^{-1} = \inf_{g\in\mathbf{A}} \{\, \|Gg\|_\infty \;\big|\; Fg = 1\}. \tag{6.3}$$

To obtain the standard $\mathbf{H}^\infty$ optimization problem from (6.3), we need to perform the parameterization of all solutions to $Fg = 1$ (known as the Youla parameterization in control theory). If $g_0 \in \mathbf{A}$ is a particular solution to the equation $Fg = 1$ and $M \in \mathbf{A}(\mathbb{C}^{n\times(n-1)})$ is a basis of the kernel of F (i.e., $FM = 0$), then all solutions can be parameterized as

$$g = g_0 + Mq, \quad q \in \mathbf{A}(\mathbb{C}^{(n-1)\times 1}),$$

which gives the standard $\mathbf{H}^\infty$ setting

$$\nu_{opt}^{-1} = \inf_{q\in\mathbf{A}} \|Gg_0 + GMq\|_\infty = \inf_{q\in\mathbf{RH}^\infty} \|T_1 + T_2 q\|_\infty \tag{6.4}$$

with given $T_1, T_2 \in \mathbf{A}$. To go back to the form (6.2), one can put, for example,

$$F = \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix}, \quad G = \begin{pmatrix} T_1 & T_2 \end{pmatrix}. \tag{6.5}$$

Let us now have a look at the dual problem to (6.2). Since the case $\Delta = \mathcal{B}\mathbb{C}^m$ can be reduced to the standard $\mathbf{H}^\infty$ optimization problem (6.4), and the latter has the well-known dual (see Theorem B.1, for instance), it seems interesting to compare our dual formulation with the standard one. As we will see, there are some similarities between these two problems, although they are different.

We assume, for simplicity, that the functions T_1 and T_2 in (6.4) are scalar. By (6.5), the problem (6.4) takes the primal form (3.4) as

$$\nu_{opt} = \sup\{\nu \;\big|\; \exists h \in \mathbf{RH}^\infty \colon \operatorname{Re}\,[\begin{pmatrix} 1 - \delta T_1(z) & -\delta T_2 \end{pmatrix} h(z)] > 0 \quad \forall z \in \mathbb{T} \quad \forall |\delta| \le \nu\}.$$

Applying Theorem 5.6, we obtain the dual problem as $\nu_{opt} = \min\{\nu_{opt|c}, \nu_{opt|s}\}$, where

$$\begin{aligned} \nu_{opt|c} &= \inf\{\|\delta\|_\infty \;\big|\; \exists w \in \mathbf{L}^1(\mathbb{R}_+) \setminus 0 \colon \begin{pmatrix} 1 - \delta T_1 & -\delta T_2 \end{pmatrix} w \in \mathbf{H}_0^1\}, \\ \nu_{opt|s} &= \inf\{|\delta| \;\big|\; \exists z \in \mathbb{T} \colon \begin{pmatrix} 1 - \delta T_1(z) & -\delta T_2(z) \end{pmatrix} = 0\}. \end{aligned}$$

Let us interpret first the singular part of the dual problem. Obviously, the equations $\delta T_1(z) = 1$, $\delta T_2(z) = 0$ have a solution δ if and only if $T_2(z) = 0$, in which case

$$\nu_{opt|s} = \min_{z\in\mathbb{T}}\{|T_1(z)|^{-1} \mid T_2(z) = 0\}.$$

This upper bound has a trivial interpretation. In fact, for all $q \in \mathbf{RH}^\infty$ (even for all $q \in \mathbf{L}^\infty$; see Remark 1 after Theorem 5.6), it holds that

$$\|T_1 + T_2 q\|_\infty = \sup_{z\in\mathbb{T}} |T_1(z) + T_2(z)q(z)| \geq |T_1(z)| \quad \forall z \in \mathbb{T}\colon T_2(z) = 0,$$

which gives $\nu_{opt}^{-1} \geq \nu_{opt|s}^{-1}$. Thus, if we assume that $T_2(z) \neq 0$ on $\mathbb{T}$, then $\nu_{opt|s} = +\infty$; i.e., there is no singular part. This brings some dual understanding as to why the $\mathbf{H}^\infty$ control problem under assumption $T_2 \neq 0$ on $\mathbb{T}$ (or full rank conditions on matrices T_2, T_3 in the four-block case) is easier to solve: the dual problem has only the regular part.

Let us now interpret the regular part of the dual problem. Because all zeros of $T_2(z)$ on $\mathbb{T}$ are under the responsibility of the singular part, we assume without loss of generality that $T_2(z) \neq 0$ on $\mathbb{T}$ to deal with $\nu_{opt} = \nu_{opt|c}$. In this case, the inner-outer factorization $T_2 = T_{2i}T_{2o}$ satisfies $T_{2o}^{-1} \in \mathbf{A}$, and hence the problem (6.4) can be equivalently represented as [6]

$$\nu_{opt}^{-1} = \inf_{Q\in\mathbf{RH}^\infty} \|R - Q\|_\infty \tag{6.6}$$

with $R = T_{2i}^* T_1 \in C(\mathbb{T})$. This convex optimization problem has the well-known dual problem given by the Nehari theorem [6]

$$\nu_{opt}^{-1} = \sup_{f\in\mathbf{H}^2} \frac{\|P_-(Rf)\|_2}{\|f\|_2} \quad \Leftrightarrow \quad \nu_{opt} = \inf_{f\in\mathbf{H}^2} \frac{\|f\|_2}{\|P_-(Rf)\|_2}. \tag{6.7}$$

Equivalently, the distance from the function $R \in C(\mathbb{T})$ to $\mathbf{H}^\infty$ is equal to the norm of Hankel operator $H_R = P_- R$. We show that our regular part of the dual problem is closely connected to (6.7). By Corollary 5.7, we have

$$\nu_{opt|c} = \inf\{\|\delta\|_\infty \mid \exists g \in (\mathbf{H}^2)^\perp \setminus 0\colon P_- \begin{pmatrix} 1 - \delta T_1 & -\delta T_2 \end{pmatrix} g = 0\}.$$

Put $f := \delta T_{2i} g$. We have $P_-(T_{2o} f) = P_-(\delta T_2 g) = 0$, which gives $T_{2o} f \in \mathbf{H}^2$ and hence $f \in \mathbf{H}^2$. Furthermore, $g = P_-(\delta T_1 g) = P_-(Rf)$. Finally $\delta = T_{2i}^* f/g$, so

$$\|\delta\|_\infty = \left\| \frac{f}{g} \right\|_\infty = \left\| \frac{f}{P_-(Rf)} \right\|_\infty,$$

and we obtain the following formula for ν_{opt}:

$$\nu_{opt} = \nu_{opt|c} = \inf_{f\in\mathbf{H}^2} \left\| \frac{f}{P_-(Rf)} \right\|_\infty. \tag{6.8}$$

This is rather similar to (6.7) yet different. Note that, in general for $f \in \mathbf{H}^2$,

$$\frac{\|f\|_2}{\|P_-(Rf)\|_2} \leq \left\| \frac{f}{P_-(Rf)} \right\|_\infty, \tag{6.9}$$

so the duality condition (6.8) appears to be stronger than (6.7). This is a question of more delicate analysis to show that for $R \in C(\mathbb{T})$ the Hankel operator is compact, and then there exists a maximizing vector $f \in \mathbf{H}^2$; i.e., $\|P_-(Rf)\|_2 = \nu_{opt}^{-1}\|f\|_2$. Moreover, for such a vector f, the function $P_-(Rf)/f$ is inner (or all-pass), i.e.,

$$\frac{P_-(Rf)(z)}{f(z)} \equiv \nu_{opt}^{-1} \quad \forall z \in \mathbb{T},$$

and the optimal function Q_{opt} in (6.6) can be found as $R - Q_{opt} = P_-(Rf)/f$ (see, for instance, [6]). These extra properties of $f/P_-(Rf)$ explain the equality in (6.9) for the optimal f.

Case 3. Δ is a convex polytope in $\mathbb{C}^m$.

In this case, both the primal and the dual problems are reduced to linear programming which can be stated explicitly in terms of vertices of Δ.

Denote the vertices of Δ by $\{\delta_k\}_{k=1}^K$, $\delta_k \in \mathbb{C}^m$. By the Krein–Milman theorem [26], the polytope Δ is the convex hull of its vertices

$$\Delta = \operatorname{co}\{\delta_1, \delta_2, \dots, \delta_K\}.$$

This means that the primal problem (5.1) is reduced to the *finite* number of linear inequalities at each point $z \in \mathbb{T}$

$$\operatorname{Re}(\nu^{-1}F(z) + \delta_k^T G(z))h(z) > 0 \quad \forall z \in \mathbb{T},\ 1 \le k \le K, \tag{6.10}$$

and the primal algorithm can be implemented as linear programming as follows.

Consider the power series decomposition of $h \in \mathbf{RH}^\infty(\mathbb{C}^{n\times 1})$

$$h(z) = \sum_{m=0}^{+\infty} h_m z^m = \begin{pmatrix} 1\cdot I & z\cdot I & z^2\cdot I & \dots \end{pmatrix} \begin{pmatrix} h_0 \\ h_1 \\ \vdots \end{pmatrix}.$$

Denoting

$$\phi(z) = \begin{pmatrix} 1 & z & z^2 & \dots \end{pmatrix},$$
$$H = \begin{pmatrix} h_0 \\ h_1 \\ \vdots \end{pmatrix},$$

the decomposition takes the short form $h(z) = [\phi(z) \otimes I]H$, with Kronecker's product $\otimes$ defined as

$$A \otimes B = \begin{pmatrix} A_{11}B & A_{12}B & \dots \\ A_{21}B & A_{22}B & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}.$$

Substituting this representation for h into the primal inequality (6.10), we get

$$\operatorname{Re}(\nu^{-1}F(z) + \delta_k^T G(z))(\phi(z)\otimes I)H = \operatorname{Re}\begin{pmatrix} \nu^{-1} & \delta_k^T \end{pmatrix}\left[\phi(z) \otimes \begin{pmatrix} F(z) \\ G(z) \end{pmatrix}\right] H > 0.$$

Denote

$$R(z) = \begin{pmatrix} F(z) \\ G(z) \end{pmatrix}, \quad D_\nu = \begin{pmatrix} \nu^{-1} & \nu^{-1} & \dots & \nu^{-1} \\ \delta_1 & \delta_2 & \dots & \delta_K \end{pmatrix} \tag{6.11}$$

to end up with the explicit linear inequality form of (6.10) with respect to H

$$\mathrm{Re}\,[\phi(z) \otimes D_\nu^T R(z)]H \succ 0 \quad \forall z \in \mathbb{T}. \tag{6.12}$$

Now we will show that the dual problem also takes the form of linear programming. Since the linear homotopy $\Delta_\nu = \nu\Delta$ is considered, we have by convexity that

$$\delta \in \Delta_\nu \quad \Leftrightarrow \quad \delta = \sum_{k=1}^{K} \mu_k \delta_k, \quad \mu_k \geq 0, \ \sum_{k=1}^{K} \mu_k = \nu,$$

and the pointwise singular part of the upper bound (5.12) becomes

$$\nu_s(z) = \inf \left\{ \sum_{k=1}^{K} \mu_k \;\middle|\; \mu_k \geq 0, \ F(z) + \sum_{k=1}^{K} \mu_k \delta_k^T G(z) = 0 \right\}. \tag{6.13}$$

Similarly, the condition $x(z) \in w(z)\Delta_\nu$ in (5.14) can be rewritten as

$$x(z) = \sum_{k=1}^{K} \mu_k(z)\delta_k, \quad \mu_k(z) \geq 0, \ \sum_{k=1}^{K} \mu_k(z) = \nu w(z),$$

and the regular part of the dual problem takes the form

$$\begin{aligned} w(z)F(z) + \sum_{k=1}^{K} \mu_k(z)\delta_k^T G(z) &= p(z), \\ \sum_{k=1}^{K} \mu_k(z) &= \nu w(z), \\ \mu_k(z) &\geq 0, \\ \int_{\mathbb{T}} w(z)\, dm(z) &> 0. \end{aligned}$$

The function w can be found from the second equation, and the number of variables is reduced:

$$\begin{aligned} \sum_{k=1}^{K} \mu_k(z)(\nu^{-1}F(z) + \delta_k^T G(z)) &= p(z), \\ \mu_k(z) &\geq 0, \\ \int_{\mathbb{T}} \sum_{k=1}^{K} \mu_k(z)\, dm(z) &> 0. \end{aligned} \tag{6.14}$$

Using the notation above for ϕ, R, and D_ν together with $\mu(z) = (\mu_1(z), \dots, \mu_K(z))^T$,

$$p(z) = \sum_{m=1}^{+\infty} p_m z^m = ([z\phi(z) \otimes I]P)^T,$$

the problem can be written as

$$\begin{aligned} \begin{pmatrix} -[z\phi(z) \otimes I] & R(z)^T D_\nu \end{pmatrix} \begin{pmatrix} P \\ \mu(z) \end{pmatrix} &= 0, \\ \mu(z) &\succeq 0, \\ E_K \int_{\mathbb{T}} \mu(z)\, dm(z) &= 1, \end{aligned} \tag{6.15}$$

where $E_K = \begin{pmatrix} 1 & \dots & 1 \end{pmatrix}$ is the K-dimensional row vector of ones.

Thus both the primal and the dual problems take the form of linear feasibility programming. The feasibility problem (6.12) can be solved in the same manner as (5.2). The condition $\|h\|_\infty \le 1$ cannot be represented as a linear condition on the coefficients H, so calculation of γ_{opt} is not linear programming. However, all we need is to check if $\gamma_{opt} > 0$. Let us choose a polyhedral set instead of the unit ball in the problem (5.2) to add the necessary linearity to the problem. For example, with

$$\mathcal{H} = \left\{ H \;\Big|\; \sum_{m=0}^{+\infty} |H_m| \le 1 \right\},$$

we have the linear programming to solve the primal problem

$$\gamma_{opt}^{LP} = \max_{H \in \mathcal{H}} \{\gamma \;|\; \mathrm{Re}\,[\phi(z) \otimes D_\nu^T R(z)] H \succeq \gamma \quad \forall z \in \mathbb{T}\}. \tag{6.16}$$

According to Remark 2 after Proposition 5.1, the conditions $\gamma_{opt} > 0$ and $\gamma_{opt}^{LP} > 0$ are equivalent.

For the dual feasibility problem (6.15), we can do a similar trick,

(6.17)
$$\epsilon_{opt} = \max_P \left\{ \epsilon \;\Big|\; R(z)^T D_\nu \mu(z) = [z\phi(z) \otimes I]P,\ \mu(z) \succeq \epsilon,\ E_K \int_{\mathbb{T}} \mu(z)\, dm(z) = 1 \right\},$$

in order to move the corresponding δ out of boundary inward Δ_ν to obtain certain "robustness" of the inclusion $\delta \in \Delta_\nu$. So the dual linear feasibility problem has a solution if and only if $\epsilon_{opt} \ge 0$.

7. Comparison with the previous result. In this section, we compare the duality result with that obtained by Megretski and Rantzer. The following generalized uniform interpolation problem was considered in [17].

Given $\Omega(z) \subset \mathbb{C}^k$, find $h \in \mathbf{RH}^\infty$ such that

$$h(z) \in \Omega(z) \quad \forall z \in \mathbb{T}. \tag{7.1}$$

Our primal problem (5.1) is recovered with

$$\Omega(z) = \{\omega \in \mathbb{C}^{n+1} \;|\; \mathrm{Re}\,(F(z) + \delta^T G(z))\omega > 0 \quad \forall \delta \in \Delta_\nu\}.$$

The authors of [17] presented the dual to the interpolation problem, which is restated below. Denote

$$\langle f, g \rangle := \mathrm{Re} \int_{\mathbb{T}} f(z)^T g(z)\, dm(z) = \frac{1}{2\pi} \mathrm{Re} \int_{-\pi}^{\pi} f(e^{i\theta})^T g(e^{i\theta})\, d\theta. \tag{7.2}$$

The dual problem is the following integral interpolation problem:

Find a function $f \in \mathbf{RH}_0^1$ such that

$$\langle f, g \rangle < 0 \tag{7.3}$$

for all measurable and bounded functions g satisfying $g(z) \in \Omega(z)$ for all $z \in \mathbb{T}$.

This duality is related to the fact that $\mathbf{H}_0^1$ is an annihilator of $\mathbf{H}^\infty$; i.e., $\langle f, h \rangle = 0$ for all $f \in \mathbf{H}_0^1$ and all $h \in \mathbf{H}^\infty$.

It was also shown in [17] that, under additional assumptions on Ω, namely, *convexity*, *continuity*, and *compactness*, there is no duality gap between the two problems. However, these additional assumptions are very restrictive and are almost never satisfied. To meet them, one has to perform a certain "regularization" procedure. Essentially one has to remove neighborhoods of zero and infinity from the cone Ω, replace the condition in the primal with $h(z) \in \Omega_\epsilon(z)$, and modify the dual integral interpolation problem with

$$\sup\{\langle f, g\rangle \mid g(z) \in \Omega_\epsilon(z) \quad \forall z \in \mathbb{T}\} < 0, \tag{7.4}$$

where

$$\Omega_\epsilon(z) = \{\omega \in \mathcal{B}C^{n+1} \mid J_\delta(\omega, z) \geq \epsilon \quad \forall \delta \in \Delta_\nu\}.$$

Since the primal problem in (5.1) is a special case of (7.1), there must be a relation between the dual (7.4) and that in Theorem 5.3.

The purpose of this section is to trace this relation. The following is a formal brief description of the idea of how to relate the two problems, using the Lagrange multiplier method (Kuhn–Tucker theorem [25]), to the optimization problem in (7.4).

$$\begin{aligned}
0 > \sup_{g(z)\in\Omega_\epsilon(z)} \langle f, g\rangle &= \sup_{\inf_{\delta\in\Delta_\nu} \operatorname{Re}\Phi_\delta(z)g(z)\geq\epsilon} \langle f, g\rangle \\
&= \inf_{\tau(z)\geq 0} \sup_{g\in\mathcal{B}\mathbf{L}^\infty} \inf_{\delta(z)\in\Delta_\nu} \big(\langle f, g\rangle + \langle \tau, \Phi_\delta g - \epsilon\rangle\big) \\
&= \inf_{\tau(z)\geq 0} \inf_{\delta(z)\in\Delta_\nu} \sup_{g\in\mathcal{B}\mathbf{L}^\infty} \big(\langle f + \tau\Phi_\delta^T, g\rangle - \langle \tau, \epsilon\rangle\big) \\
&= \inf_{\delta(z)\in\Delta_\nu} \inf_{\tau(z)\geq 0} \int_{\mathbb{T}} (|f + \tau\Phi_\delta^T| - \epsilon\tau)\, dm.
\end{aligned}$$

Hence a function $f \in \mathbf{H}_0^1$ is a solution to (7.3) if and only if there exist functions $\delta \in \mathbf{L}^\infty(\Delta_\nu)$ and $\tau \in \mathbf{L}^1(\mathbb{R}_+)$ such that

$$\|f + \tau\Phi_\delta^T\|_1 < \epsilon\|\tau\|_1.$$

Denoting $w = \tau/\|\tau\|_1$, $p = -f^T/\|\tau\|_1$, we get the similar condition to (5.6) because $\epsilon > 0$ can be chosen arbitrarily small in the regularization Ω_ϵ.

8. Toward an implementation of the primal-dual method. Although efficiency of a primal-dual method depends on a numerical implementation, we do not provide any *particular* numerical scheme and do not discuss *details* of corresponding numerical issues for the following reason. Once the dual pair (5.1), (5.14) is obtained, one can approach it in many different ways—either directly or via finite-dimensional approximations. The latter has a large variety of forms and methods: among others, there are Galerkin-type finite-element schemes, the method of ellipsoids and other cutting plane methods, the analytic center cutting plane (ACCP), path-following, potential reduction, and other barrier-based interior-point methods [20], and, in some cases, semidefinite programming and linear matrix inequalities (LMIs), etc. Each of them has certain advantages and disadvantages, and we cannot prescribe any of them as a cure for all cases. What we do briefly discuss in this section is a general idea of how to bridge the gap between the infinite-dimensional convex condition and a possible numerical implementation of it, taking into account some specific features of the problem.

The main difficulty is that the convex problems (5.1) and (5.14), and even their linear counterparts (6.12) and (6.15), have infinitely many decision variables and equalities/inequalities. To overcome this difficulty, an approximation should be done. The analytical property of the functions h and p make a pointwise representation useless, so the best we can do for the approximation is to introduce a Schauder basis of the space $\mathbf{RH}^\infty$ and $\mathbf{H}_0^1$. A right choice of basis is very important for fast convergence yet unclear a priori in general, so an adaptive adjustment of the basis (especially of the pole location of basis elements) may be useful. Some insight for the adjustment can be obtained by solving an auxiliary simple problem. For example, the approximation of Δ by a ball in $\mathbb{C}^m$ gives the equivalent $\mathbf{H}^\infty$ optimization problem that can be solved effectively, and the pole location of the optimal solution may be used for the first choice of the basis in $\mathbf{RH}^\infty$. Another possibility is to choose a basis and run the algorithm to find a solution most probably of a very high dimension. If the choice of basis is bad, the solution will have many close pole-zero cancellations, and what is left may be used for the basis adjustment.

In any case, by choosing any basis of $\mathbf{RH}^\infty$, for example, the polynomial basis $\{1, z, z^2, \dots\}$, we know that the solution *can* be found since polynomials are uniformly dense in $\mathbf{A}$. This means that we can consider an equivalent problem with a *finite* (although unknown a priori) number of decision variables. Since the number of inequalities is still the continuum, we get a semi-infinite programming that can be solved in many ways. One is by taking a grid on the unit circle $\mathbb{T}$ (see discretization methods for semi-infinite programming in [10, 23]), which has been realized and discussed in [8, 9]. Another is to use a cutting plane algorithm and, in particular, the ACCP method [2, 19] that usually performs better than the method of ellipsoids. A variation of the ACCP method for cone feasibility problems (our primal problem (5.1) is conic) appears in [21]. Other barrier function methods applied to similar problems can be found in [13, 14, 15].

9. Numerical example: A nonconvex $\mathbf{H}^\infty$ optimization problem. In this section, a nonconvex $\mathbf{H}^\infty$ optimization problem is solved numerically by the primal-dual method derived above. By the $\mathbf{H}^\infty$ optimization, we mean a problem in the form

$$\inf_{Q \in \mathbf{RH}^\infty} \{\gamma \mid T_1(z) + T_2(z)Q(z) \in \gamma\Omega \quad \forall z \in \mathbb{D}\}, \tag{9.1}$$

where Ω is the neighborhood of the origin in $\mathbb{C}$ and T_1, $T_2 \in \mathbf{A}$. If $\Omega = \mathbb{D}$, we have the standard $\mathbf{H}^\infty$ optimization [6] since

$$T_1(z) + T_2(z)Q(z) \in \gamma\mathbb{D} \quad \forall z \in \mathbb{D} \quad \Leftrightarrow \quad \|T_1 + T_2Q\|_\infty \le \gamma.$$

If Ω is a convex set, the problem (9.1) is convex.

In this section, we consider the $\mathbf{H}^\infty$ optimization problem (9.1) with the *nonconvex* set

$$\Omega = \{re^{i\phi} \in \mathbb{C} \mid 0 \le r < |\cos(\phi)| + |\sin(\phi)|\}. \tag{9.2}$$

The shape of Ω is shown in Figure 9.1. This set appears, for example, in the optimization of the stability radius for a linear system (see Example 2 in section 4) with the "diamond" type of uncertainty in feedback

$$\Delta = \{x + iy \in \mathbb{C} \mid |x| + |y| \le 1\}.$$

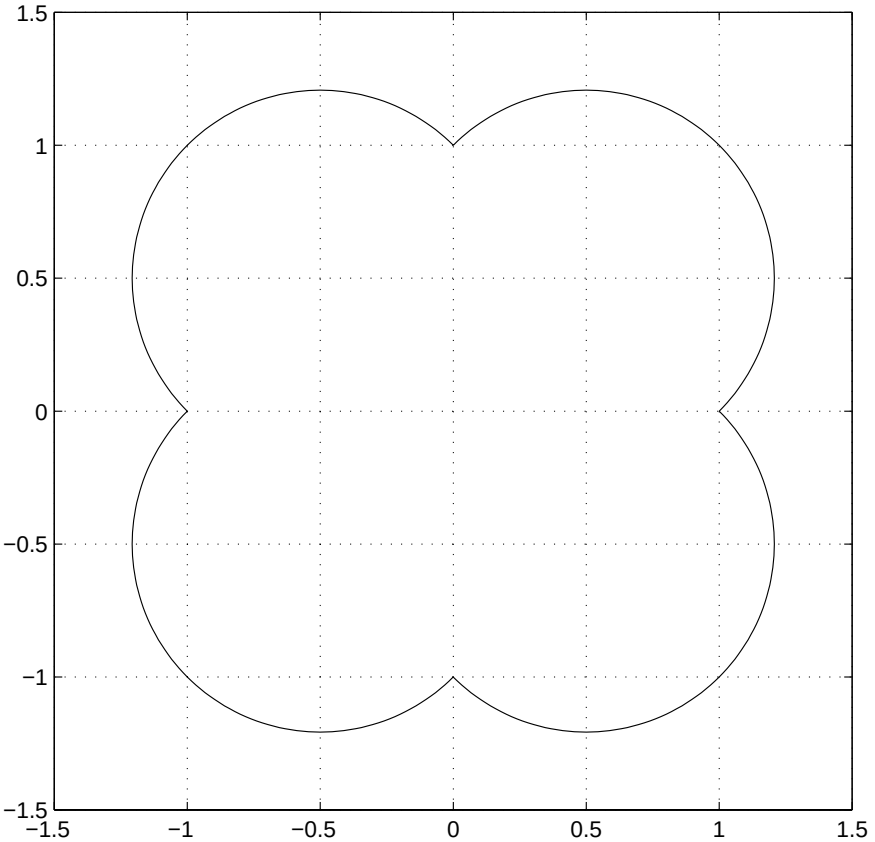


FIG. 9.1. *The set* Ω *from* (9.2).

The problem can be reduced to the cone optimization. First note that $\Omega = \mathbb{C} \setminus \Delta^{-1}$. Then, for $g = T_1 + T_2 Q$, we can rewrite the condition $g(z) \in \nu\Omega$ as

$$\nu^{-1} g(z) \notin \Delta^{-1} \quad \Leftrightarrow \quad 0 \notin 1 + \nu \Delta g(z).$$

Theorem 4.1 implies that there exists a function $\alpha \in \mathbf{RH}^\infty$ such that $\mathrm{Re}\,(1 + \nu\Delta g(z))\alpha(z) > 0$. Thus the problem takes the cone optimization form (5.1) with $F = (1\ 0)$, $G = (T_1\ T_2)$, and $h = \mathrm{col}(\alpha, Q\alpha)$.

The set Δ is a polytope; hence we can use linear programming as described in section 6, Case 3. To approximate it by finite-dimensional problems, we choose a simple idea of discretization. Taking into account only the finite grid $\{z_k\}_{k=1}^K$, representing the upper half of $\mathbb{T}$, and the first N coefficients of the function h, the inequality (6.12) gives the finite-dimensional linear program

$$A_{KN} H_N \succ 0, \tag{9.3}$$

where the matrix A_{KN} is of the size $4K \times 2N$. We solve this feasibility problem as explained in (6.16) except that we choose another set $\mathcal{H}$, namely,

$$\max\{\epsilon \;\big|\; A_{KN} H_N \succeq \epsilon,\ -1 \preceq H_N \preceq 1\}. \tag{9.4}$$

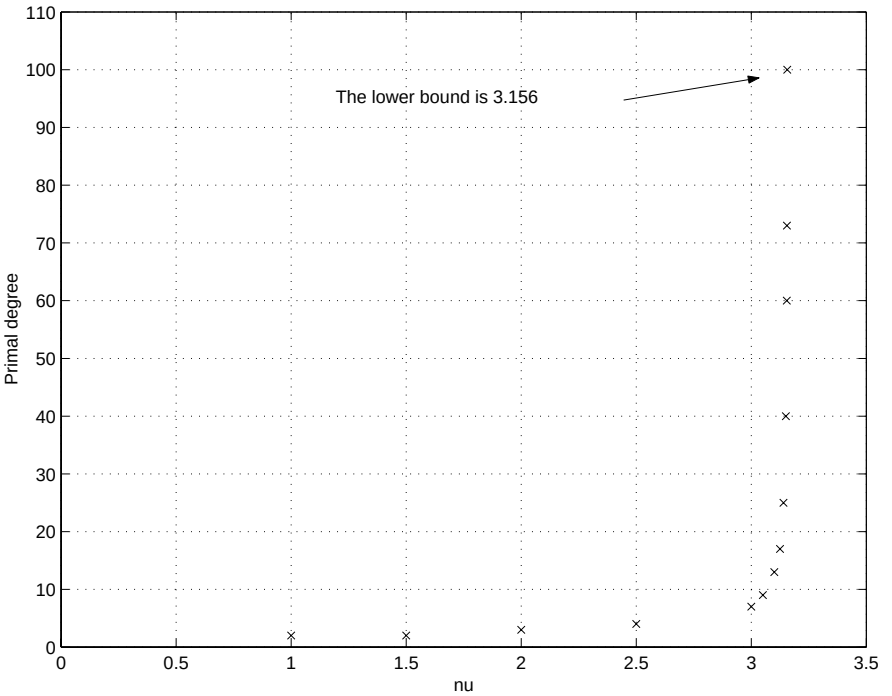


FIG. 9.2. *The primal algorithm produces lower bounds on* ν_{opt}.

The dual problem (6.15) is approximated precisely in the same manner to get the finite-dimensional linear program

$$\begin{array}{rcl} A_{12}X & = & 0, \\ A_{22}X & \succeq & 0, \\ A_{32}X & = & 1, \end{array} \tag{9.5}$$

where the $(2N+4K)$-dimensional vector X absorbs N coefficients of the function $p \in \mathbf{H}_0^1$ and K values of the coordinate functions μ from (6.15). Since the discretization can have difficulties dealing with equalities, we slightly modify the problem (9.5) as

$$\min\{\gamma \mid -\gamma \preceq A_{12}X \preceq \gamma,\ A_{22}X \succeq 0,\ A_{32}X = 1\}. \tag{9.6}$$

This is almost the discretization of (5.6). The difference is that we replace the integral norm with the uniform one (which is approved by Lemma 5.8). The dual algorithm finds a solution if γ is zero. The details of the algorithms can be found in [8, 9].

Take the following functions:

$$T_1(z) = \frac{z^5 + 3z^4 + 2z^3 + 4z^2 + 5z + 3}{z^3 - z^2 - 4z + 12}, \quad T_2(z) = z^2 + \frac{1}{2}.$$

Since $T_1/T_2 \notin \mathbf{H}^\infty$, the solution to the optimization problem is not trivial.

We run the primal algorithm in the linear programming form (9.4) for different values of ν. For small enough ν, it finds a solution, and we increase the value. We stop the optimization at $\nu_{low} = 3.156$ when the degree of approximation N has reached 100. It becomes hard for the linear solver to find an approximation of higher degree because

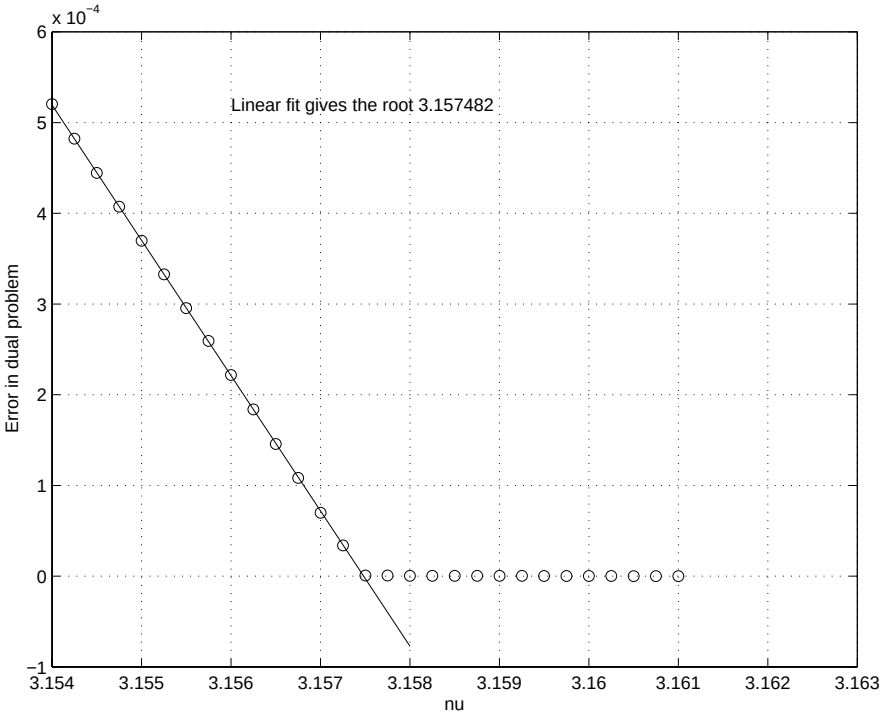


FIG. 9.3. *The dual algorithm gives the opposite side bounds on* ν_{opt}.

the dimension of the linear program (9.4) at this step is already of size 1200×200. The plot of N versus ν is shown in Figure 9.2.

To estimate how far the lower bound ν_{low} is from the optimal value, we use the dual problem in the form (9.6). If $\nu > \nu_{opt}$, the error γ may be minimized to zero. The upper bound ν_{upp} obtained by the dual algorithm is 3.1575. Hence the primal solution for ν_{low} has a good level of suboptimality (about 0.05%). The value of γ as a function of ν is depicted in Figure 9.3.

It happens that the error γ in our example exhibits the linear dependence on ν for $\nu < \nu_{opt}$, which can be also used to find ν_{opt} as a solution to $\gamma(\nu) = 0$.

10. Conclusion. There is a number of design problems in control theory which can be stated as the quasiconvex optimization (3.4). Given a level of suboptimality, the corresponding convex problem can be solved by finite-dimensional approximations. However, these approximations give only one-sided bounds on the optimal value. In this paper, the dual representation (5.7) to this quasiconvex optimization problem has been derived using convex duality arguments in a Banach space setting. Opposite bounds for the optimal value can be obtained by solving the quasiconvex problems (5.8), (5.9). Hence the quasiconvex optimization problem (3.4) can be solved by a primal-dual method followed by a line search on ν.

Appendix A. Proofs of Proposition 5.1 and Lemma 5.2.

PROPOSITION A.1. *Let* $\Delta \subset \mathbb{C}^n$ *be a compact set, and* $g \in C(\mathbb{C}^n)$. *Then*

$$G = \inf_{\delta \in \Delta} \operatorname{Re} \delta^T g \in C.$$

Proof. Let $z_k \to z_0$ as $k \to +\infty$. We have to prove that $G(z_k) \to G(z_0)$. Obviously, for all $\delta \in \Delta$, it holds that $G(z_k) \le \operatorname{Re} \delta^T g(z_k)$, and hence

$$\limsup_{k \to +\infty} G(z_k) \le G(z_0).$$

To prove the opposite inequality, let us consider a subsequence $\{z_{k_n}\}$ of $\{z_k\}$ such that $G(z_{k_n})$ has a limit. The set Δ is compact, and the function $\operatorname{Re} \delta^T g(z_{k_n})$ is continuous (linear) with respect to δ. Hence there exists $\delta_n \in \Delta$ such that $G(z_{k_n}) = \operatorname{Re} \delta_n^T g(z_{k_n})$. The sequence $\{\delta_n\}$ is compact, and then there exists a limit point $\delta_0 \in \Delta$. Moreover,

$$G(z_{k_n}) \to \operatorname{Re} \delta_0^T g(z_0) \ge G(z_0).$$

Since the subsequence $\{z_{k_n}\}$ is arbitrary, we conclude that

$$\liminf_{k \to +\infty} G(z_k) \ge G(z_0).$$

The proposition is proved. ∎

PROPOSITION A.2.

1. *The function*

$$\Gamma(h) := \inf_{z \in \mathbb{T}} \inf_{\delta \in \Delta_\nu} J_\delta(h, z) \tag{A.1}$$

is concave, continuous in the uniform topology of $\mathbf{A}$*, and positively homogeneous:*

$$\Gamma(\lambda h) = \lambda \Gamma(h) \quad \forall h \in \mathbf{A},\ \lambda \in \mathbb{R}_+.$$

Proof. Recall that $J_\delta(h, z) = \operatorname{Re}\,(F(z) + \delta^T G(z))h(z)$. The function Γ is concave and positively homogeneous as an infimum of linear functions.

To prove that the function Γ is continuous with respect to the topology of uniform convergence in $\mathbf{A}$, consider the identity

$$\operatorname{Re} \Phi_\delta h_1 = \operatorname{Re} \Phi_\delta h_2 + \operatorname{Re} \Phi_\delta (h_1 - h_2).$$

Taking the infimum over $\mathbb{T}$ and Δ_ν of both sides yields

$$\Gamma(h_1) \ge \Gamma(h_2) + \inf_{z \in \mathbb{T}} \inf_{\delta \in \Delta_\nu} \operatorname{Re} \Phi_\delta(z)(h_1(z) - h_2(z)).$$

Since h_1 and h_2 are arbitrary functions in $\mathbf{A}$ and can be interchanged, we get

$$\begin{aligned} |\Gamma(h_2) - \Gamma(h_1)| &\le \sup_{z \in \mathbb{T}} \sup_{\delta \in \Delta_\nu} |\operatorname{Re} \Phi_\delta(z)(h_2(z) - h_1(z))| \\ &\le \left(\|F\|_\infty + \sup_{\delta \in \Delta_\nu} |\delta|\, \|G\|_\infty \right) \|h_2 - h_1\|_\infty. \end{aligned}$$

Thus the function Γ is continuous (in fact, even Lipschitz). The proposition is proved. ∎

Proof of Proposition 5.1. (2) $\Rightarrow$ (1) Since the set $\mathbf{RH}^\infty$ is uniformly dense in $\mathbf{A}$ and, by Proposition A.2, Γ is continuous in the uniform topology of $\mathbf{A}$, we have

$$\gamma_{opt} = \sup_{h \in \mathcal{B}\mathbf{RH}^\infty} \inf_{z \in \mathbb{T}} \inf_{\delta \in \Delta_\nu} J_\delta(h, z).$$

Hence $\gamma_{opt} > 0$ obviously implies (5.1).

(1) $\Rightarrow$ (2) Let $h_0 \in \mathbf{RH}^\infty$ be such that (5.1) holds. The function $J_\delta(h_0, z)$ is continuous on (z, δ), and the set $\mathbb{T} \times \Delta$ is compact. Then there exists (z_0, δ_0) such that

$$\Gamma(h_0) = \inf_{z\in\mathbb{T}} \inf_{\delta\in\Delta_\nu} J_\delta(h_0, z) = J_{\delta_0}(h_0, z_0) > 0.$$

Finally,

$$\gamma_{opt} \geq \sup_{h\in\mathcal{B}\mathbf{RH}^\infty} \Gamma(h) \geq \Gamma\left(\frac{h_0}{\|h_0\|_\infty}\right) = \frac{\Gamma(h_0)}{\|h_0\|_\infty} = \frac{J_{\delta_0}(h_0, z_0)}{\|h_0\|_\infty} > 0. \qquad \square$$

Proof of Lemma 5.2. Recall the notation from (3.1) and (3.2),

$$\Phi_\delta = F + \delta^T G, \quad J_\delta(h, z) = \operatorname{Re} \Phi_\delta(z) h(z),$$

and denote

$$\gamma_{\mathbf{H}^\infty} = \sup_{h\in\mathcal{B}\mathbf{H}^\infty} \operatorname*{ess\,inf}_{z\in\mathbb{T}} \inf_{\delta\in\Delta_\nu} J_\delta(h, z).$$

It is obvious that $\gamma_{\mathbf{H}^\infty} \geq \gamma_{opt}$. We have to prove the opposite inequality.

By definition, for any $\epsilon > 0$, there is a function $h \in \mathcal{B}\mathbf{H}^\infty$ such that

$$J_\delta(h, z) \geq \gamma_{\mathbf{H}^\infty} - \epsilon$$

for all $\delta \in \Delta_\nu$ and almost all $z \in \mathbb{T}$. The function J_δ is harmonic as the real part of analytical function. Hence, by the mean value theorem,

$$J_\delta(h, rz) \geq \operatorname*{ess\,inf}_{z\in\mathbb{T}} J_\delta(h, z) \geq \gamma_{\mathbf{H}^\infty} - \epsilon$$

for all $\delta \in \Delta$, $z \in \mathbb{T}$, and $0 \leq r < 1$.

Denote $h_r(z) = h(rz)$ for $0 \leq r < 1$. It holds that $h_r \in \mathbf{A}$ and $\|h_r\|_\infty \leq \|h\|_\infty$; hence $h_r \in \mathcal{B}\mathbf{A}$. Denote similarly $F_r(z) = F(rz)$ and $G_r(z) = G(rz)$. We then obtain

$$\begin{aligned} J_\delta(h_r, z) &= \operatorname{Re} \Phi_\delta(z) h_r(z) = J_\delta(h, rz) + \operatorname{Re} (\Phi_\delta(z) - \Phi(rz)) h_r(z) \\ &\geq \gamma_{\mathbf{H}^\infty} - \epsilon - \sup_{z\in\mathbb{T}} |\Phi_\delta(rz) - \Phi_\delta(z)| \\ &\geq \gamma_{\mathbf{H}^\infty} - \epsilon - \|F - F_r\|_\infty - |\delta| \|G - G_r\|_\infty. \end{aligned}$$

Therefore,

$$\Gamma(h_r) = \inf_{\delta\in\Delta_\nu} \inf_{z\in\mathbb{T}} J_\delta(h_r, z) \geq \gamma_{\mathbf{H}^\infty} - \epsilon - \|F - F_r\|_\infty - \sup_{\delta\in\Delta_\nu} |\delta| \, \|G - G_r\|_\infty.$$

The functions F and G are continuous, and the Fatou theorem [7] implies that $\|F - F_r\|_\infty \to 0$ and $\|G - G_r\|_\infty \to 0$ as $r \to 1$. The constant $\sup_{\delta\in\Delta_\nu} |\delta|$ is bounded. Thus

$$\gamma_{opt} \geq \lim_{r\to 1} \Gamma(h_r) \geq \gamma_{\mathbf{H}^\infty} - \epsilon$$

for all $\epsilon > 0$. Hence $\gamma_{opt} \geq \gamma_{\mathbf{H}^\infty}$. $\square$

Appendix B. Proof of Theorem 5.3. We cite first the classical result which gives the basic Banach duality relation [7, 16]. Recall the notation (7.2)

$$\langle f, g \rangle = \operatorname{Re} \int_{\mathbb{T}} f(z)^T g(z) \, dm(z).$$

THEOREM B.1. *Let $1 \le p \le +\infty$, $1/p + 1/q = 1$, and $f \in \mathbf{L}^p$. Then the distance from the f to $\mathbf{H}^p$ is*

$$\operatorname{dist}_{\mathbf{L}^p}(f, \mathbf{H}^p) = \sup_{h \in \mathcal{B}\mathbf{H}_0^q} \langle f, h \rangle.$$

There exists a function $f_{opt} \in \mathbf{H}^p$ such that $\operatorname{dist}_{\mathbf{L}^p}(f, \mathbf{H}^p) = \|f - f_{opt}\|_p$. The function f_{opt} is unique if $p \neq +\infty$.

COROLLARY B.2. *Under the same conditions as in Theorem* B.1*, it holds that*

$$\operatorname{dist}_{\mathbf{L}^p}(f, \mathbf{H}_0^p) = \sup_{h \in \mathcal{B}\mathbf{H}^q} \langle f, h \rangle.$$

Proof. Consider a function $f(z)/z \in \mathbf{L}^p$, and apply Theorem B.1. ∎

Now we prove the following lemma.

LEMMA B.3. *Let a measurable function $\phi\colon \mathbb{C}^n \times \mathbb{C} \to \mathbb{R}$ be concave and continuous in the first argument. Denote $\phi_h(z) = \phi(h(z), z)$. Then*

$$\sup_{h \in \mathcal{B}\mathbf{H}^\infty(\mathbb{C}^n)} \operatorname*{ess\,inf}_{z \in \mathbb{T}} \phi_h(z) = \inf_{w \in \mathcal{S}\mathbf{L}^1(\mathbb{R}_+)} \sup_{h \in \mathcal{B}\mathbf{H}^\infty(\mathbb{C}^n)} \langle \phi_h, w \rangle. \tag{B.1}$$

Proof. Note that

$$\operatorname*{ess\,inf}_{z \in \mathbb{T}} \phi_h(z) = \inf_{w \in \mathcal{S}\mathbf{L}^1(\mathbb{R}_+)} \langle \phi_h, w \rangle.$$

The set $\mathcal{B}\mathbf{H}^\infty$ is convex and *weak compact, the set $\mathcal{S}\mathbf{L}^1(\mathbb{R}_+)$ is convex, and the function $\langle \phi_h, w \rangle$ is *weak continuous and concave on h and continuous and convex on w (even linear). Hence, by Ky Fan's min-max theorem [5], the order of sup and inf can be interchanged. ∎

Proof of Theorem 5.3. Denote

$$\mathcal{J}_\nu(h, z) = \inf_{\delta \in \Delta_\nu} J_\delta(h, z).$$

This function is concave in h. By Proposition A.1, it is also continuous in h. Using Lemma B.3, the min-max theorem [5], and Theorem B.1, we have

$$\begin{aligned}
&\sup_{h \in \mathcal{B}\mathbf{H}^\infty} \operatorname*{ess\,inf}_{z \in \mathbb{T}} \mathcal{J}_\nu(h, z) = \inf_{w \in \mathcal{S}\mathbf{L}^1(\mathbb{R}_+)} \sup_{h \in \mathcal{B}\mathbf{H}^\infty} \langle \mathcal{J}_\nu(h, \cdot), w \rangle \\
&= \inf_{w \in \mathcal{S}\mathbf{L}^1(\mathbb{R}_+)} \sup_{h \in \mathcal{B}\mathbf{H}^\infty} \inf_{\delta(z) \in \Delta_\nu} \langle \operatorname{Re} \Phi_\delta h, w \rangle = \inf_{w \in \mathcal{S}\mathbf{L}^1(\mathbb{R}_+)} \inf_{\delta(z) \in \Delta_\nu} \sup_{h \in \mathcal{B}\mathbf{H}^\infty} \langle \Phi_\delta^T w, h \rangle \\
&= \operatorname{dist}_{\mathbf{L}^1}([F + \mathbf{L}^\infty(\Delta_\nu)^T G] \mathcal{S}\mathbf{L}^1(\mathbb{R}_+), \mathbf{H}_0^1). \qquad \blacksquare
\end{aligned}$$

Appendix C. Proof of Theorem 5.6. To begin with, we prove the following lemma. All details about Borel measures and Lebesgue decomposition can be found in [27].

LEMMA C.1. *Let $\Phi \in C$, $g \in \mathbf{L}^1$, and w be the density function of a real Borel measure μ on $\mathbb{T}$. Then*

$$\|\Phi w + g\|_1 = \|\Phi w_c + g\|_1 + \int_{\mathbb{T}} |\Phi| \, d\mu_s, \tag{C.1}$$

where $\mu = w_c \, dm + \mu_s$ is the Lebesgue decomposition to absolutely continuous and singular parts.

Proof. Due to the triangular inequality, it is obvious that

$$\|\Phi w + g\|_1 \le \|\Phi w_c + g\|_1 + \int_{\mathbb{T}} |\Phi|\, d\mu_s.$$

Since $m(\operatorname{supp}(\mu_s)) = 0$, for any $\epsilon > 0$, there is an open set $T_\epsilon \subset \mathbb{T}$ such that $\operatorname{supp}(\mu_s) \subset T_\epsilon$ and $m(T_\epsilon) \le \epsilon$. Then the following inequality holds:

$$\begin{aligned}
\|\Phi w + g\|_1 &= \|\Phi w_s + \Phi w_c + g\|_1 = \int_{\mathbb{T}\setminus T_\epsilon} |\Phi w_c + g|\, dm + \int_{T_\epsilon} |\Phi w_s + \Phi w_c + g|\, dm \\
&\ge \int_{\mathbb{T}\setminus T_\epsilon} |\Phi w_c + g|\, dm + \int_{T_\epsilon} |\Phi|\, d\mu_s - \int_{T_\epsilon} |\Phi w_c + g|\, dm \\
&\xrightarrow{\epsilon\to 0} \|\Phi w_c + g\|_1 + \int_{\operatorname{supp}(\mu_s)} |\Phi|\, d\mu_s = \|\Phi w_c + g\|_1 + \int_{\mathbb{T}} |\Phi|\, d\mu_s.
\end{aligned}$$

This proves the lemma. □

Proof of Theorem 5.6. Denote

$$\gamma_{min} = \inf_{\delta(z)\in\Delta_\nu} \inf_{h\in\mathbf{H}_0^1} \inf_{\|w\|_1=1} \|\Phi_\delta w - h\|_1,$$

and let $\delta_i \in \mathbf{L}^\infty(\Delta_\nu)$ and $w_j \in \mathcal{S}\mathbf{L}^1(\mathbb{R}_+)$ be such that

$$\inf_{h\in\mathbf{H}_0^1} \|\Phi_{\delta_i} w_j - h\|_1 \to \gamma_{min}, \quad i, j \to +\infty.$$

Without loss of generality, we assume that $\delta_i \in C$ because the set $\mathbf{L}^\infty(\Delta_\nu) \cap C$ is *weakly dense in $\mathbf{L}^\infty(\Delta_\nu)$. The functions F and G belong to $\mathbf{A}$; therefore,

$$\Phi_{\delta_i} = F + \delta_i^T G \in C.$$

Embed the set $\mathcal{S}\mathbf{L}^1(\mathbb{R}_+)$ in the unit sphere of real Borel measures on $\mathbb{T}$ as $w \hookrightarrow \mu$,

$$\mu(\Omega) = \int_\Omega w(z)\, dm(z).$$

The latter set is *weakly compact, so there exists a *weak limit point μ_{opt} of the sequence $\{\mu_j\}_{j=1}^{+\infty}$. The measure μ_{opt} can be decomposed to the absolutely continuous μ_c and singular parts μ_s as $\mu_{opt} = \mu_c + \mu_s$. This corresponds to the decomposition of a generalized density w_{opt} as $w_{opt} = w_c + w_s$, where w_c is a regular function in $\Re\mathbf{L}_+^1$ and w_s is a generalized function which is equal to zero on $\mathbb{T} \setminus E$ and $m(E) = 0$.

We claim that w_{opt} can be chosen in such a way that either w_c or w_s is zero. Indeed, let $\lambda_0 = \|w_c\|_1$. Then

$$w_{opt} = \lambda_0 \widetilde{w}_c + (1 - \lambda_0)\widetilde{w}_s,$$

where $\widetilde{w}_c$ and $\widetilde{w}_s$ are normalized densities. In the case when $\lambda_0 = 0$, the claim is proved. So assume that $\lambda_0 > 0$. Using Lemma C.1, we have a decomposition

$$\|\Phi w_{opt} - h\|_1 = \lambda_0 \left\| \Phi\widetilde{w}_c - \frac{h}{\lambda_0} \right\|_1 + (1 - \lambda_0) \int_E |\Phi|\, d\widetilde{\mu}_s$$

whenever $h \in \mathbf{H}_0^1$ and $\Phi \in C$. This implies that

$$\begin{aligned}\inf_{h\in\mathbf{H}_0^1} \|\Phi w_{opt} - h\|_1 &= \lambda_0 \inf_{h\in\mathbf{H}_0^1} \|\Phi \widetilde{w}_c - h\|_1 + (1-\lambda_0)\int_E |\Phi|\, d\widetilde{\mu}_s \\ &= \inf_{\lambda\in[0,1]} \left(\lambda \inf_{h\in\mathbf{H}_0^1} \|\Phi \widetilde{w}_c - h\|_1 + (1-\lambda)\int_E |\Phi|\, d\widetilde{\mu}_s\right).\end{aligned}$$

The function is linear with respect to λ. Hence 0 or 1 is always the optimal value. Moreover, if the optimal λ is 0, then the similar argument proves that the singular part $\widetilde{w}_s$ of the optimal density w_{opt} can be chosen as a Dirac δ-function at one of the points $z_0 = \arg\min_{z\in\mathbb{T}} |\Phi(z)|$.

Thus we have proved that, for any $\Phi \in C$, there exists $\widetilde{w}_c \in \mathcal{S}\Re\mathbf{L}_+^1$ such that

$$\inf_{h\in\mathbf{H}_0^1} \inf_{\|w\|_1=1} \|\Phi w - h\|_1 = \min\left\{\inf_{h\in\mathbf{H}_0^1} \|\Phi\widetilde{w}_c - h\|_1,\ \min_{z\in\mathbb{T}} |\Phi(z)|\right\}.$$

Hence, denoting δ_{opt} a *weak limit point of the sequence $\{\delta_i\}_{i=1}^{+\infty}$, we get

$$\begin{aligned}\gamma_{min} &= \inf_{\delta(z)\in\Delta_\nu} \min\left\{\inf_{h\in\mathbf{H}_0^1} \|\Phi_\delta \widetilde{w}_c - h\|_1,\ \min_{z\in\mathbb{T}} |\Phi_\delta(z)|\right\} \\ &= \min\left\{\inf_{h\in\mathbf{H}_0^1} \|\Phi_{\delta_{opt}} \widetilde{w}_c - h\|_1,\ \operatorname*{ess\,inf}_{z\in\mathbb{T}} |\Phi_{\delta_{opt}}(z)|\right\}.\end{aligned}$$

To complete the proof, we use Theorem B.1 to conclude that there exists the optimal function h_{opt} such that

$$\inf_{h\in\mathbf{H}_0^1} \|\Phi_{\delta_{opt}} \widetilde{w}_c - h\|_1 = \|\Phi_{\delta_{opt}} \widetilde{w}_c - h_{opt}\|_1.$$

Now the condition $\gamma_{min} = 0$ can be easily transformed to that stated in Theorem 5.6. The proof is finished. ∎

Appendix D. Proofs of Propositions in section 6.

Proof of Proposition 6.1. (1) ⇒ (2) Denote by w_λ the Poisson kernel at the point $\lambda \in \mathbb{D}$:

$$w_\lambda(z) = \frac{1-|\lambda|^2}{|z-\lambda|^2}.$$

It is clear that $w_\lambda \in \mathbf{L}^1(\mathbb{R}_+)$ and $\|w_\lambda\|_1 = 1$. Since

$$F(\lambda) = \int_{\mathbb{T}} F(z) w_\lambda(z)\, dm(z) = \int_{\mathbb{T}} (F(z)w_\lambda(z) - p(z))\, dm(z)$$

for all $p \in \mathbf{H}_0^1$, we have the inequality

$$|F(\lambda)| \le \inf_{p\in\mathbf{H}_0^1} \|Fw_\lambda - p\|_1. \tag{D.1}$$

The infimum can be calculated by the minimum norm theorem (see Corollary B.2):

$$\begin{aligned}|F(\lambda)| &\le \inf_{p\in\mathbf{H}_0^1} \|Fw_\lambda - p\|_1 = \sup_{h\in\mathcal{B}\mathbf{H}^\infty} \int_{\mathbb{T}} \operatorname{Re}\left(F(z)h(z)\right) w_\lambda(z)\, dm(z) \\ &= \sup_{h\in\mathcal{B}\mathbf{H}^\infty} \operatorname{Re} F(\lambda)h(\lambda) = |F(\lambda)|.\end{aligned}$$

Hence (D.1) is the equality

$$|F(\lambda)| = \inf_{p \in \mathbf{H}_0^1} \|Fw_\lambda - p\|_1 = \|Fw_\lambda - p_0\|_1,$$

where the existence of p_0 is guaranteed by the same minimum norm theorem.

Thus $|F(\lambda)| = 0$ implies that $Fw_\lambda \in \mathbf{H}_0^1$.

(2) $\Rightarrow$ (1) Denote the entries of F by F_k, and suppose that there exists a $w \in \mathbf{L}^1(\mathbb{R}_+) \setminus 0$ such that

$$F_k(z)w(z) = zp_k(z), \quad p_k \in \mathbf{H}^1. \tag{D.2}$$

For all k, the function $zp_k(z)/F_k(z)$ is analytical everywhere in $\mathbb{D}$ except the zeros of F_k. However, these functions are the same function w from (D.2); hence w is the analytical function everywhere in $\mathbb{D}$ except the common zeros of F_k, which are the zeros of $|F|$.

Assume that there are no zeros of $|F|$ in $\mathbb{D}$, and so $w \in \mathbf{H}^1$. The only analytical functions that take real values on $\mathbb{T}$ are constants. Hence w is the constant function. Furthermore, $w(0) = 0$ by (D.2). Then $w = 0$, which contradicts the assumption that $w \neq 0$. Therefore, there must exist a zero of $|F|$ in $\mathbb{D}$. ∎

Proof of Proposition 6.2. (1) $\Rightarrow$ (2) The inequality (6.2) implies that $\operatorname{Re}(F(\lambda) + \delta^T G(\lambda))h(\lambda) > 0$ for all $\lambda \in \mathbb{D}$. Then

$$(Fh + \delta^T Gh)^{-1} \in \mathbf{A} \quad \forall \delta \in \nu\Delta. \tag{D.3}$$

In particular, it gives $(Fh)^{-1} \in \mathbf{A}$. The function $g = h(Fh)^{-1} \in \mathbf{A}$ satisfies $Fg = 1$, and (D.3) becomes

$$(1 + \delta^T Gg)^{-1} = (Fh)^{-1}(Fh + \delta^T Gh)^{-1} \in \mathbf{A} \quad \forall \delta \in \nu\Delta. \tag{D.4}$$

Since, for each $z \in \mathbb{T}$, one can choose

$$\delta = \alpha\overline{G(z)g(z)}/|G(z)g(z)|, \quad \alpha \in [0, \nu],$$

(D.4) implies that $|G(z)g(z)| \notin [\nu^{-1}, +\infty)$. Hence $\|Gg\|_\infty < \nu^{-1}$.

(2) $\Rightarrow$ (1) By Rouché's theorem [27], we get (D.4). Thus

$$1 + \delta^T G(z)g(z) \neq 0 \quad \forall z \in \mathbb{T},\ \delta \in \nu\Delta.$$

By Theorem 4.1, there exists a function $\alpha \in \mathbf{A}$ such that

$$\operatorname{Re}(1 + \delta^T G(z)g(z))\alpha(z) > 0 \quad \forall z \in \mathbb{T},\ \delta \in \nu\Delta.$$

Finally, for $h = g\alpha \in \mathbf{A}$, we have

$$\operatorname{Re}(F(z) + \delta^T G(z))h(z) = \operatorname{Re}(1 + \delta^T G(z)g(z))\alpha(z) > 0. \qquad ∎$$

REFERENCES

[1] B. D. O. Anderson, S. Dasgupta, P. Khargonekar, F. J. Kraus, and M. Mansour, *Robust strict positive realness: Characterization and construction*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 37 (1990), pp. 869–876.

[2] D. S. Atkinson and P. M. Vaidya, *A cutting plane algorithm for convex programming that uses analytic centers*, Math. Programming, 69 (1995), pp. 1–43.

[3] S. P. Boyd and C. H. Barratt, *Linear Controller Design—Limits of Performance*, Prentice-Hall, Englewood Cliffs, NJ, 1991.
[4] S. P. Boyd and L. Vandenberghe, *Convex Optimization*, to be published in 2003. A preliminary version from December 2001 is available at http://www.stanford.edu/class/ee364/reader.ps (1999).
[5] H. Brézis, L. Nirenberg, and G. Stampacchia, *A remark on Ky Fan's mini-max principle*, Boll. Un. Mat. Ital., 6 (1972), pp. 293–300.
[6] B. A. Francis, *A Course in* $\mathbf{H}^\infty$ *Control Theory*, Lecture Notes in Control and Inform. Sci. 88, Springer-Verlag, New York, 1987.
[7] J. Garnett, *Bounded Analytic Functions*, Academic Press, New York, 1981.
[8] A. Ghulchak and A. Rantzer, *Robust controller design via linear programming*, in Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, AZ, 1999, pp. 1815–1820.
[9] A. Ghulchak and A. Rantzer, *Robust control under parametric uncertainties via primal-dual convex analysis*, IEEE Trans. Automat. Control, 47 (2002), pp. 632–636.
[10] R. Hettich and K. Kortanek, *Semi-infinite programming: Theory, methods, and applications*, SIAM Rev., 35 (1993), pp. 380–429.
[11] U. Jönsson, *Duality in multiplier based robustness analysis*, IEEE Trans. Automat. Control, 44 (1999), pp. 2246–2256.
[12] U. Jönsson and A. Rantzer, *Duality bounds in robustness analysis*, Automatica J. IFAC, 33 (1997), pp. 1835–1844.
[13] C. Kao, U. Jonsson, and A. Megretski, *An algorithm for solving optimization problems involving special frequency dependent LMIs*, in Proceedings of the American Control Conference, Chicago, IL, 2000, pp. 307–311.
[14] C. Kao and A. Megretski, *Fast algorithm for solving IQC feasibility and optimization problem*, in Proceedings of the American Control Conference, Arlington, VA, 2001, pp. 3019–3024.
[15] M. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, *Applications of second-order cone programming*, Linear Algebra Appl., 284 (1998), pp. 193–228.
[16] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.
[17] A. Megretski and A. Rantzer, *Robust control synthesis by convex optimization: Projective parametrization and duality*, Proceedings of the IFAC Congress, Sydney, Australia, 1993.
[18] A. Megretski and A. Rantzer, *System analysis via integral quadratic constraints*, IEEE Trans. Automat. Control, 42 (1997), pp. 819–830.
[19] Yu. Nesterov *Cutting plane algorithm from analytic centers: Efficiency estimates*, Math. Programming, 69 (1995), pp. 149–176.
[20] Yu. Nesterov and A. Nemirovsky, *Interior Point Polynomial Algorithms in Convex Programming*, SIAM Stud. Appl. Math. 13, SIAM, Philadelphia, 1994.
[21] Yu. Nesterov and J.-P. Vial, *Homogeneous analytic center cutting plane methods for convex problems and variational inequalities*, SIAM J. Optim., 9 (1999), pp. 707–728.
[22] A. Rantzer and A. Megretski, *A convex parameterization of robustly stabilizing controllers*, IEEE Trans. Automat. Control, 39 (1994), pp. 1802–1808.
[23] R. Reemtsen and S. Görner, *Numerical Methods for Semi-Infinite Programming: A Survey*, in Semi-Infinite Programming, Reemtsen and Rückmann, eds., Kluwer, Dordrecht, The Netherlands, 1998, pp. 195–275.
[24] P. Young and M. Dahleh, *Infinite-dimensional convex optimization in optimal and robust control theory*, IEEE Trans. Automat. Control, 42 (1997), pp. 1370–1381.
[25] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[26] W. Rudin, *Functional Analysis*, McGraw–Hill, New York, 1973.
[27] W. Rudin, *Real and Complex Analysis*, 3rd ed., McGraw–Hill, New York, 1987.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 1, pp. 278–302

WORST CASE DESIGN FOR ROBUST COMPENSATION*

SHAUL GUTMAN† AND EYTAN PALDI‡

Abstract. We consider the problem

$$\operatorname*{Max}_{F} \quad \operatorname*{Min}_{c(p,F)=0} \quad h(p,F),$$

where $F \in \mathcal{R}^r$, $p \in \mathcal{R}^m$, and where $c(\cdot)$ and $h(\cdot)$ are C^1. Let $\phi(F) = \operatorname{Min}_{c(p,F)=0} h(p,F)$. We show, by means of simple examples, that $\phi(\cdot)$ is, in general, discontinuous. We develop in this paper necessary conditions for the case where $\phi(\cdot)$ is continuous (but not necessarily differentiable). In an alternative approach (which is computationally inferior), we treat the discontinuous case as well. We apply the results to robust control in linear systems where p stands for the (structured) real parameter uncertainty vector and F stands for the control parameters vector. We demonstrate the results by means of examples.

Key words. robust control, min-max

AMS subject classifications. 93C05, 93D09

PII. S0363012999363572

1. Introduction.

Robust control has attracted the attention of researchers for more than three decades. Basically, we consider a linear time invariant plant with (constant) real parameters whose values are known to lie in some prescribed (ellipsoidal) uncertainty set. The uncertainty structure, however, is arbitrary. We seek a fixed (dynamic) compensator in closed loop such that the closed loop spectrum lies in a prescribed set in the complex plane (relative stability) for all values in the uncertainty set. The compensator has any desired structure, like reduced order, or zeros in the left half plane (LHP). Among infinitely many possible robust controllers, we identify a controller with a clear geometrical meaning, one that maximizes the stability margin. In other words, we seek a controller so that we can increase the uncertainty radius to the maximum while maintaining (relative) stability. It turns out that such a controller leads to a max-min problem as pointed out in [1]. The cost in such a max-min problem is the (ellipsoidal) distance, and the constraint is the stability boundary described by the critical polynomial [2].

While problems of max-min nature are discussed in the literature [3, 4, 5], we treat the robust control problem independently since it possesses some exclusive properties. As explained in [5], max-min creates a differentiability problem, since, after the first operation (minimization), the cost may no longer be smooth. Moreover, as will be demonstrated in section 7, at a global max-min point, a small variation in the outer variables (max) may cause a jump in $\phi(\cdot)$. It turns out that, while the largest ellipsoid may cause a jump, a robust compensator associated with a fixed uncertainty ellipsoid never creates a jump. In mathematical terms, the fixed uncertain ellipsoid is associated with $\operatorname{Max}_F \operatorname{Min}_{p\in P} h(p,F)$, where the set P depends only on p (and not on F). On the other hand, the maximum uncertain ellipsoid is associated with

*Received by the editors October 13, 1999; accepted for publication (in revised form) January 11, 2002; published electronically June 5, 2002. This research was supported by the Fund for the Promotion of Research at the Technion.

http://www.siam.org/journals/sicon/41-1/36357.html

†Department of Mechanical Engineering, Technion, Technion City, 32000 Haifa, Israel (mergutm@techunix.technion.
ac.il).

‡113 A.H.Silver St., 32000 Haifa, Israel.

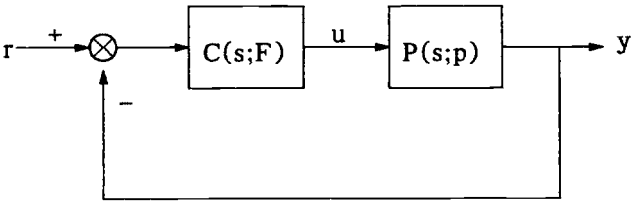


FIG. 1.

$\text{Max}_F \ \text{Min}_{c(p,F)=0} h(p,F)$, where the coupling of p and F in the inner operation is evident. To find necessary conditions for max-min, articles [3, 4, 5] use the directional derivative, which in turn requires that $\phi(\cdot)$ is continuous.

In this paper, we use the Lagrange multiplier method for the inner operation. As for the outer operation, we show that, under mild conditions, at extremum, the number of tangent points needed is related to the (reduced) number of control parameters. It shows, in turn, that the necessary conditions presented in [1] form a special case of our approach. We also present an alternative approach, which is capable of treating jumps. However, computationally this approach is inferior to the first approach. The necessary conditions presented here have the form of a set of polynomial equations. The solutions of these equations are candidates for robust compensation. To find all real solutions, we suggest using a probability one homotopy method like HOMPACK [6]. Finally, we apply to each real solution a robust analysis like [7] and select a robust compensator.

The paper is organized as follows. In section 2, we recall results from root clustering and parameter space, in particular, the critical polynomial, and state our objective. In section 3, we present a max-min formulation associated with the largest ellipsoid contained in the parameter space. Likewise, we present a max-min formulation associated with a fixed uncertainty ellipsoid. In sections 4 and 5, we find necessary conditions for the largest ellipsoid under the assumption that $\phi(\cdot)$ is continuous, while, in section 6, we discuss a more general case using an alternative approach. In section 7, we present necessary conditions for a fixed ellipsoid. Finally, in section 8, we demonstrate some of the results by means of examples.

2. Parameter space and problem statement. We consider a linear time invariant plant P and a linear time invariant controller C connected in a feedback form. We denote the plant's and the controller's transfer functions by $P(s;p)$ and $C(s;F)$, respectively, where $s \in \mathbb{C}$ is the usual complex variable, $p \in \Omega \subset \mathcal{R}^m$ is the plant's uncertainty parameter vector, Ω is the uncertainty set, and $F \in \mathcal{R}^r$ is the controller's design parameter vector. Let $\Delta(s;p,F)$ be the closed loop characteristic polynomial, and let $\sigma(\Delta)$ be the closed loop spectrum. As an illustration, consider the unit feedback configuration as depicted in Figure 1.

Let

$$P(s;p) = D_p^{-1}(s;p)N_p(s;p),$$
$$C(s;F) = N_c(s;F)D_c^{-1}(s;F). \tag{1}$$

Then the characteristic equation $|I + P(s;p)C(s;F)| = 0$ leads to the characteristic polynomial

$$\Delta(s;p,F) = |D_p(s;p)D_c(s;F) + N_p(s;p)N_c(s;F)|. \tag{2}$$

To discuss the matrix version, consider a strictly proper plant with realization

$$\begin{aligned} \dot{x}_p &= A_p(p)x_p + B_p(p)u, \\ y &= C_p(p)x_p. \end{aligned} \tag{3}$$

We seek a stabilizing compensator of fixed order of the form

$$\begin{aligned} \dot{x}_c &= A_c x_c + B_c(r-y), \\ u &= C_c x_c + D_c(r-y), \end{aligned} \tag{4}$$

where r is a reference signal. Combining (3) and (4) in closed loop, we obtain for $r=0$ and for $x=[x_p' \;\; x_c']'$

$$\dot{x} = A(p,F)x, \tag{5}$$

where

$$A(p,F) = A(p) - B(p)FC(p), \tag{6}$$

$$A(p) = \begin{bmatrix} A_p(p) & 0 \\ 0 & 0 \end{bmatrix}, \quad B(p) = \begin{bmatrix} B_p(p) & 0 \\ 0 & I \end{bmatrix},$$

$$C(p) = \begin{bmatrix} C_p(p) & 0 \\ 0 & I \end{bmatrix}, \quad F = \begin{bmatrix} D_c & C_c \\ B_c & A_c \end{bmatrix}. \tag{7}$$

Design objective. Given an uncertain plant P and a region $\aleph \subset \mathbb{C}$ in the complex plane, select F such that $\sigma[\Delta(s;p,F)] \subset \aleph$ or $\sigma[A(p,F] \subset \aleph$ for all $p \in \Omega$. Such a controller is called *robust*.

The root clustering $\aleph$ may represent asymptotic stability, like the left half plane and unit disk, or relative stability like the left hyperbola. Moreover, in some applications, we require a controller of fixed structure (including fixed order) or some other properties like stable controllers. These requirements are included in the above design objective. As mentioned in the introduction, we achieve our objective using the concept of worst case design in the parameter space. To this end, we need an important instrument, known as the *critical polynomial* $c(p,F)$ in the parameter space. This polynomial vanishes at the (relative) stability boundary and is positive in the interior. As early as 1929, it was recognized [8] that the Hurwitz determinant is a critical polynomial for the left half plane, and, in 1963, the unit disk counterpart was constructed [9]. Recently [2, 10], as a part of a general root clustering theory, a critical polynomial for an arbitrary region was constructed. Consider a region $\aleph \subset \mathbb{C}$ in the complex plane

$$\aleph = \left\{ x + jy : f(x,y) = \sum_{i,j} f_{ij} x^i y^j < 0 \right\}, \tag{8}$$

where $f(x,y)$ is a given polynomial in the two real variables x and y. Define the coefficients ϕ_{ik} in the following way:

$$\phi(\alpha,\beta) = f\left(\frac{\alpha+\beta}{2}, \frac{\alpha-\beta}{2j}\right) = \sum \phi_{ik}\alpha^i\beta^k. \tag{9}$$

Then $\Phi = [\phi_{ik}]$ is Hermitian and

$$\aleph = \{s \in \mathbb{C} : \phi(s, \bar{s}) < 0\}. \tag{10}$$

By construction, if $s = x + jy \in \partial\aleph$, the boundary of $\aleph$, then $\phi(s, \bar{s}) = 0$. In 1983, it was recognized [11] that, in parameter space analysis, a region $\aleph$ should be *regular* in the following sense.

DEFINITION 2.1. *$\phi(\alpha, \beta)$, defined above, is* regular *if α, $\beta \in \aleph \Rightarrow \phi(\alpha, \beta) \neq 0$. We say that $\aleph$ is regular if $\phi(\alpha, \beta)$ is regular.*

Note that, in previous articles, we have used the term H-transformability to describe regularity. Following [2, 12], we define the critical polynomial $c(p, F)$ by a double resultant operator as follows:

$$c(p, F) = \mathrm{Res}_\lambda[\Delta(\lambda; p, F), \mathrm{Res}_s[\bar{\Delta}(s; p, F), -\phi(\lambda, s)]], \tag{11}$$

where "Res" is the resultant of two polynomials, $\Delta(\cdot)$ is the (possibly complex) characteristic polynomial, and $\bar{\Delta}(\cdot)$ is the conjugate complex of $\Delta(\cdot)$. Likewise, given a matrix $A(p, F) \in \mathbb{C}^{n \times n}$,

$$c(p, F) = (-1)^n \det\left[\sum \phi_{ik} A^i \otimes \bar{A}^k\right] \tag{12}$$

where $\otimes$ is the Kronecker product of two matrices.

In this matrix version, we will feel free to replace the vector F by an appropriate matrix F. Now consider the image $\hat{\aleph}$ of $\aleph$ in the parameter space.

$$(i) \quad \hat{\aleph} = \{(p, F) \in \mathcal{R}^{m+r} : \sigma(A(p, F)) \subset \aleph\},$$

and in the polynomial setting,

$$\text{(ii)} \quad \hat{\aleph} = \{(p, F) \in \mathcal{R}^{m+r} : \sigma(\Delta(s; p, F)) \subset \aleph\}. \tag{13}$$

It is known [2] that

$$\begin{aligned} &\text{(i)} \quad \hat{\aleph} \subset \{(p, F) \in \mathcal{R}^{m+r} : c(p, F) \geq 0\} \text{ always,} \\ &\text{(ii)} \quad \hat{\aleph} \subset \{(p, F) \in \mathcal{R}^{m+r} : c(p, F) > 0\} \text{ for regular regions,} \\ &\text{(iii)} \quad \partial\hat{\aleph} \subset \{(p, F) \in \mathcal{R}^{m+r} : c(p, F) = 0\} \text{ always.} \end{aligned} \tag{14}$$

3. Robust control: Max-min approach. We first assume that the compensator's parameter vector F is fixed, and we discuss robust *analysis*. In particular, let the system characteristic polynomial be

$$\Delta(\lambda; p) = \sum_{i=0}^{n} a_i(p) \lambda^i, \tag{15}$$

where $p \in \mathcal{R}^m$ is the uncertainty vector associated with the plant and is restricted by

$$p \in \Omega \subset \mathcal{R}^m; \quad \Omega = \left\{ p \in \mathcal{R}^m : \sum_{i=1}^{m} \nu_i^2 p_i^2 - \gamma^2 \leq 0 \right\}. \tag{16}$$

The following theorem is straightforward.

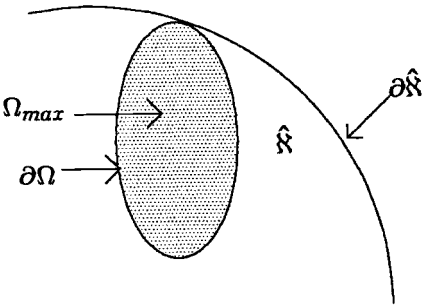


FIG. 2.

THEOREM 3.1. *Consider the uncertain polynomial* (15)*, and suppose that the nominal polynomial* $\Delta(\lambda;0)$ *has all of its roots inside a regular region* $\aleph$*. Then for all* $p \in \text{int}(\Omega)$*,* $\sigma(\Delta(\lambda;p)) \subset \aleph$ *if and only if for all* $p \in \text{int}(\Omega)$*,* $c(p) > 0$*.*

Instead of discussing the positivity $c(p)$ in Ω directly, we pose an optimization problem whose solution identifies the maximum uncertainty radius allowed. Consider

$$\text{Min} \sum_{i=1}^{m} \nu_i^2 p_i^2 \text{ subject to (s.t.) } c(p) = 0. \tag{17}$$

The general meaning of this optimization is depicted in Figure 2.

If we are interested in a fixed radius rather than the maximum radius, we pose the following problem:

$$\text{Min} \sum_{i=1}^{m} \nu_i^2 p_i^2 \text{ s.t. } c(p) - \rho^2 = 0. \tag{18}$$

Here we have to solve (18) for different values of ρ until the corresponding value of $\sum_{i=1}^{m} \nu_i^2 p_i^2$ takes the required value γ^2. However, in light of (14), we can reverse the role of the cost and the constraint to obtain a direct representation for a fixed uncertainty radius as follows:

$$\text{Min } c(p) \text{ s.t.} \sum_{i=1}^{m} \nu_i^2 p_i^2 - \gamma^2 \leq 0. \tag{19}$$

Note that (19) is equivalent in $\hat{\aleph}$ to $\text{Max}[-\ln c(p)]$. Thus we may write

$$\text{Max}[-\ln c(p)] \text{ s.t.} \sum_{i=1}^{m} \nu_i^2 p_i^2 - \gamma^2 \leq 0. \tag{20}$$

The geometry of (19) and (20) is depicted in Figure 3.

We now present a matrix version. Let $A(p) \in \mathcal{R}^{n\times n}$ be a system matrix, where $A(0)$, the nominal matrix, is stable with respect to $\aleph$. With $A(p)$ and $\aleph$ we associate the matrix equation

$$\sum_{i,k} \phi_{ik} A^i(p) P A'^k(p) = -Q, \tag{21}$$

where $(\cdot)'$ is a matrix transposition, $\phi(\cdot)$ defined by (9) is M-transformable [2], and $Q = Q' > 0$.

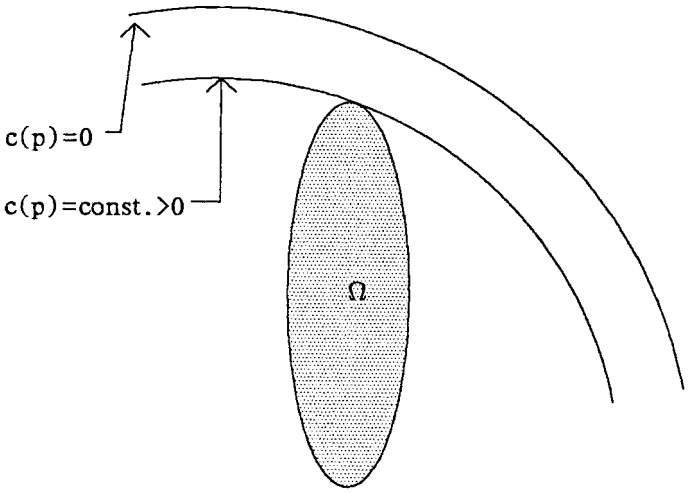


FIG. 3.

According to the theory of composite matrices, the solution to (21) has the form

$$\phi(A(p) \otimes A(p))\hat{p} = -\hat{q}, \tag{22}$$

where $\hat{p}$ and $\hat{q}$ are the stacking operators of matrices P and Q. Recalling the critical polynomial (12), we see that matrix P has a unique solution if and only if $c(p) \neq 0$. We conclude

$$\begin{aligned} &\operatorname{tr}(P) > 0 \quad \forall\, p \in \hat{\aleph}, \\ &\operatorname{tr}(P) \to \infty \text{ as } (p \in \hat{\aleph}) \to \partial\hat{\aleph}, \end{aligned} \tag{23}$$

where $\operatorname{tr}(P)$ is the trace of P. Thus (20) can be replaced by the matrix version

$$\begin{aligned} &\text{Max } \operatorname{tr}(P) \text{ s.t.} \sum_{i,k} \phi_{ik} A^i(p) P A'^k(p) + Q = 0, \\ &\sum_{i=1}^{m} \nu_i^2 p_i^2 - \gamma^2 \leq 0. \end{aligned} \tag{24}$$

So far we have seen that robust stability can be associated with an optimization problem. However, robust stability is the *analysis* part in the design process. Now we discuss the *synthesis*; namely, we construct a compensator such that closed loop relative stability (with respect to $\aleph$) is preserved in the entire uncertainty set Ω. In moving from analysis to synthesis, the optimization problem moves from minimization to a conflict between the uncertainty and the controller. In other words, minimization becomes max-min operation. This operation reflects our deterministic approach to robust compensators. That is, we seek a single compensator F such that

$$\sigma(\Delta(\lambda; p, F)) \subset \aleph \quad \forall\, p \in \Omega, \tag{25}$$

where now $\Delta(\lambda; p, F)$ is the closed loop characteristic polynomial.

To this end, we free the compensator F and seek a value that maximizes the uncertainty radius. Using (17), we obtain the following max-min problem:

$$\underset{F}{\text{Max}}\ \underset{p}{\text{Min}}\ \sum_{i=1}^{m}\nu_i^2 p_i^2 \quad \text{s.t. } c(p,F)=0. \tag{26}$$

Likewise, (18) becomes

$$\underset{F}{\text{Max}}\ \underset{p}{\text{Min}}\ \sum_{i=1}^{m}\nu_i^2 p_i^2 \quad \text{s.t. } c(p,F)-\rho^2=0, \tag{27}$$

while, for fixed uncertainty radius, (19) implies

$$\underset{F}{\text{Max}}\ \underset{p}{\text{Min}}\ c(p,F) \quad \text{s.t.} \sum_{i=1}^{m}\nu_i^2 p_i^2-\gamma^2\le 0. \tag{28}$$

For the inverse critical polynomial (20), we have

$$\underset{F}{\text{Min}}\ \underset{p}{\text{Max}}\ [-\ln c(p,F)] \quad \text{s.t.} \sum_{i=1}^{m}\nu_i^2 p_i^2-\gamma^2\le 0. \tag{29}$$

Finally, the matrix version (24) now becomes

$$\underset{F}{\text{Min}}\ \underset{p}{\text{Max}}\ \ \text{tr}(P) \quad \text{s.t.} \sum_{i,k}\phi_{ik}A^i(p,F)PA'^k(p,f)+Q=0,$$

$$\sum_{i=1}^{m}\nu_i^2 p_i^2-\gamma^2\le 0. \tag{30}$$

It should be emphasized that in a max-min or min-max operation, we first take the inner operation with respect to p and then the outer operation with respect to F. While the first operation always leads to a finite point (Ω is compact), the second operation (with respect to F) may lead to a point at infinity. This is so since F is not restricted to lie in a compact set. Furthermore, the number of solutions to the above problems may be infinity. We leave the discussion of these difficulties to a later section. Finally, note that we do not suppose that a real conflict between the uncertainty and the controller takes place. Rather, we associate with the robust compensator problem a max-min formulation, the solution of which leads to a robust compensator. This compensator is, of course, not unique.

4. Tangency points. Consider the max-min problem (26)

$$\underset{F}{\text{Max}}\ \underset{p}{\text{Min}}\ \sum_{i=1}^{m}\nu_i p_i^2 := p'Dp \quad \text{s.t. } c(p,F)=0. \tag{31}$$

ASSUMPTION 4.1. *The nominal system is stabilizable with respect to* $\aleph$*; that is,* $\{F:(0,F)\in\hat{\aleph}\}\neq\phi$.

We define the robustness radius $\gamma(F)$ by

$$\gamma^2(F):=\underset{c(p,F)=0}{\text{Min}}\ \sum_{i=1}^{m}\nu_i^2 p_i^2 \ \text{ if } \ (0,F)\in\hat{\aleph};\ \ \text{otherwise},\gamma(F):=0. \tag{32}$$

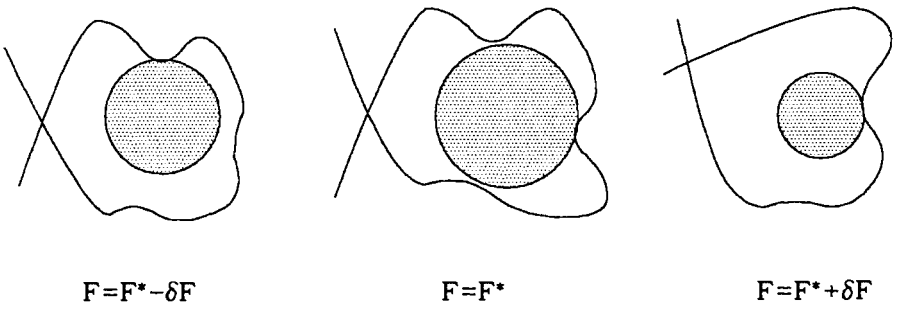


FIG. 4.

In the case of a jump in $\gamma(F)$, we can add the notion of a *practical robustness radius.* Thus the optimal robust controller F^* in the max-min sense is given by

$$\gamma^* := \operatorname*{Sup}_F \gamma(F) = \gamma^*(F^*) \geq \gamma(F^*), \tag{33}$$

where the inequality in (33) becomes an equality if and only if F^* is not a jump point. The minimization in (32) takes place at *contact points* of the critical constraint and the uncertain ellipsoid. Such points are depicted in Figure 4.

Suppose that, at each F with $\gamma(F) > 0$, the contact takes place at $p = p^{(k)}(F)$. These are the local maxima of problem (32). We say that points $p^{(k)}(F)$ are *regular* points of the critical surface $\{p : c(p, F) = 0\}$ if the gradient of $c(p, F)$ at the surface with respect to p does not vanish at these points. In other words,

$$c(p^{(k)}, F) = 0, \quad \gamma^2(F) = \sum_{i=1}^{m} \nu_i^2 p^{(k)2} \Rightarrow c_p(p^{(k)}, F) \neq 0. \tag{34}$$

At this point, it is worth noting that the critical polynomial $c(p, F)$ must be in a reduced form. In particular, let $\{\lambda_i\}$ be the roots of $\Delta(\cdot; p, F) = 0$. Then [2]

$$c(p, F) = \prod_{i,j=1}^{n} (-\phi(\lambda_i, \bar{\lambda}_j)) = (-1)^n \tilde{c}^2(p, F) \prod_{i=1}^{n} \phi(\lambda_i, \bar{\lambda}_i), \tag{35}$$

where

$$\tilde{c}(p, F) = \prod_{i<j} |\phi(\lambda_i, \bar{\lambda}_j)|. \tag{36}$$

Thus $c_p(p, F) \equiv 0$ for every real (p, F) on the critical surface. To overcome this difficulty, we have to replace $c(p, F)$ by the reduced form

$$\hat{c}(p, F) = (-1)^n \tilde{c}(p, F) \prod_{i=1}^{n} \phi(\lambda_i, \bar{\lambda}_i). \tag{37}$$

In the case of complex polynomials, this factorization is impossible. However, for the *real* case, we have the following result:

(i) Given a real characteristic polynomial $\Delta(s;p,F)$,

$$c(p,F) = \mathrm{Res}[\Delta(s;p,F), -\phi(s,s)] \left(\frac{\mathrm{Res}[\Delta(\lambda;p,F), \mathrm{Res}[\Delta(s;p,F), -\phi(\lambda,s)]]}{\mathrm{Res}[\Delta(s;p,F), -\phi(s,s)]} \right)^{\frac{1}{2}}. \tag{38}$$

(ii) Given a real matrix $A(p,F)$,

$$c(p,F) = \det\left[\sum_i f_{i0} A^i(p,F)\right] \det[-\phi(A(p,F) \odot A(p,F))], \tag{39}$$

where $\odot$ is the bialternate produce [2]. In what follows, by $c(p,F)$ we mean the reduced form. It should be noted that, for computational simplicity, we remove from $c(\cdot)$ factors using $\hat{c} = c/gcd(c, \frac{\partial c}{\partial p_1}, \ldots, \frac{\partial c}{\partial p_m})$.

Let us first assume that a max-min point is regular. Let $p^{(i)}, i = 1,2,\ldots,l$, be points of tangency of the critical constraint and the ellipsoid. Using the Lagrange multiplier method, we may write

$$\begin{aligned} &\text{(i)} \quad c_p(p^{(i)},F) = \lambda^{(i)} p^{(i)\prime} D, \quad D = \mathrm{diag}[\nu_k^2]_{k=1}^m, \\ &\text{(ii)} \quad c(p^{(i)},F) = 0, \qquad i-1,2,\ldots,l. \end{aligned} \tag{40}$$

Here

$$c_p(p,F) = \left[\frac{\partial c}{\partial p_1} \cdots \frac{\partial c}{\partial p_m}\right], \tag{41}$$

and λ is the (inverse) Lagrange multiplier.

DEFINITION 4.1. *Given a compensator F satisfying $(0,F) \in \hat{\aleph}$, we say that $p \in \mathcal{R}^m$ is a* contact point *for F if the minimum* (32) *is attained. Given F, we say that $p \in \mathcal{R}^m$ is a* tangency point *if, for some real λ,*

$$c(p,F) = 0, \quad c_p(p,F) = \lambda p' D \neq 0. \tag{42}$$

Note that every contact point for F for which $c_p \neq 0$ is a tangency point (but not necessarily the opposite).

DEFINITION 4.2. *If, for some F^*, there exist a tangency point p^* and a corresponding λ^*, we say that p^* has a* regular multiplicity *k if there exist a neighborhood N_{F^*} of F^* and $C^1(N_{F^*})$ functions $p^{(i)}(F)$, $\lambda^{(i)}(F), i = 1,2,\ldots,k$, satisfying* (42) *for every $F \in N_{F^*}$ and $p^{(i)}(F^*) = p^*$, $\lambda^{(i)}(F^*) = \lambda^*$, such that k is the maximal possible number of such $\{p^{(i)}(F), \lambda^{(i)}(F)\}$ pairs. If $k = 1$, we say that p^* has a* simple regular multiplicity. *If, in addition, $p^{(1)}(F)$ and $\lambda^{(1)}(F)$ are the unique solutions in the neighborhoods of p^*, λ^*, for every $F \in N_{F^*}$, we say that p^* is* simple regular.

We open by characterizing simple regular points. If, for some F^*, there exist a simple regular tangency point p^* and a corresponding λ^*, then there exist a neighborhood N_{F^*} of F^* and functions $p(F)$, $\lambda(F) \in C^1(N_{F^*})$ satisfying (42) for every $F \in N_{F^*}$ with initial conditions $p(F^*) = p^*$, $\lambda(F^*) = \lambda^*$. Thus, taking $p = p(F)$, $\lambda = \lambda(F)$, we obtain from (42)

$$\begin{bmatrix} c_{pp} - \lambda D & -Dp \\ c_p & 0 \end{bmatrix} \begin{bmatrix} dp \\ d\lambda \end{bmatrix} = \begin{bmatrix} -c_{pF} \\ -c_F \end{bmatrix} dF, \quad \text{where } c_{pF} := \left[\frac{\partial^2 c}{\partial p_i \partial F_j}\right]. \tag{43}$$

DEFINITION 4.3. *We define the two matrices*

$$J = \begin{bmatrix} c_{pp} - \lambda D & -Dp \\ c_p & 0 \end{bmatrix}, \qquad J_1 = \begin{bmatrix} c_{pp} - \lambda D & -Dp & -c_{pF} \\ c_p & 0 & -c_F \end{bmatrix}, \tag{44}$$

where $p \in \mathcal{R}^m$, $F \in \mathcal{R}^r$, J *is* $(m+1) \times (m+1)$, *and* J_1 *is* $(m+1) \times (m+r+1)$.

ASSUMPTION 4.2. *If* $p \in \mathcal{R}^m$, $\lambda \in \mathcal{R}$, $F \in \mathcal{R}^r$ *satisfy* (42), *then* J_1 *is of full rank; that is,*

$$c(p,F) = 0, \quad c_p(p,F) = \lambda p' D \Rightarrow \operatorname{rank}(J_1) = m+1. \tag{45}$$

Suppose that, contrary to the assumption, $\operatorname{rank}(J_1) = m$. Then choose m independent rows in J_1, and express the remaining one as a linear combination of these m rows, using m combination coefficients. In that case, we obtain (together with (42)) $2m + r + 2$ equations in $2m + r + 1$ unknowns. If these equations are independent, then the extra equation restricts the nominal plant. Thus the assumption holds for almost every nominal plant.

THEOREM 4.1. *Let* p^* *be a tangent point for* F^* *with a corresponding* λ^*. *Suppose that Assumptions* 4.1 *and* 4.2 *are satisfied. Then* p^* *is simple regular if and only if*

$$|J(F^*, p^*, \lambda^*)| \neq 0. \tag{46}$$

Proof. Let Assumption 4.1 be satisfied at (p^*, F^*, λ^*). Then the implicit function theorem implies that (46) is sufficient for (42) to have, in some neighborhood N_{F^*} of F^*, a unique solution $p(F) \in N_{p^*}$, $\lambda(F) \in N_{\lambda^*}$. Moreover, $p(F), \lambda(F) \in C^1(N_{F^*})$ and satisfy the differential equation

$$\frac{d}{dF}\begin{bmatrix} p \\ \lambda \end{bmatrix} = J^{-1}(p, F, \lambda) \begin{bmatrix} -c_{pF} \\ -c_F \end{bmatrix}, \quad \text{where } \frac{dp}{dF} := \left[\frac{dp_i}{dF_j}\right]. \tag{47}$$

The uniqueness implies the initial conditions $p(F^*) = p^*$, $\lambda(F^*) = \lambda^*$, and, according to Definition 4.2, p^* is simple regular.

To prove necessity, suppose that the above assumptions are satisfied and that p^* is simple regular. Then, in some neighborhood N_{F^*}, system (42) has a solution $p(F)$, $\lambda(F) \in C^1(N_{F^*})$ for which $p(F^*) = p^*$, $\lambda(F^*) = \lambda^*$. For sufficiently small $\|\Delta F\|$, let $F = F^* + \Delta F$, $\Delta p = p(F) - p^*$, $\Delta\lambda = \lambda(F) - \lambda^*$. Then, from (42), we obtain

$$J(p^*, F^*, \lambda^*)\begin{bmatrix} \Delta p \\ \Delta\lambda \end{bmatrix} + O(\|\Delta p\|^2 + |\Delta\lambda|^2) = \begin{bmatrix} -c_{pF} \\ -c_F \end{bmatrix}\Delta F + O(\|\Delta F\|^2). \tag{48}$$

Since $p(F)$, $\lambda(F) \in C^1(N_{F^*})$, it follows that, for sufficiently small $\|\Delta F\|$,

$$(\|\Delta p\|, |\Delta\lambda|) = O(\|\Delta F\|). \tag{49}$$

If (46) does not hold, then there exists $\mu \in \mathcal{R}^{m+1}$ with $\|\mu\| = 1$ such that

$$\mu' J(p^*, F^*, \lambda^*) = 0. \tag{50}$$

Denote $J_2 := \left[\begin{smallmatrix} -c_{pF} \\ -c_F \end{smallmatrix}\right]$. Then, multiplying (48) by μ' and using (49)–(50), we obtain

$$\mu' J_2 \Delta F = O(\|\Delta F\|^2). \tag{51}$$

However, by Assumption 4.2, we have $0 \neq \mu' J_1 = [\mu' J \ \mu' J_2] = [0 \ \mu' J_2]$. Thus $\nu' := \mu' J_2 \neq 0$.

Choosing ΔF in the direction of ν, we obtain from (51) $\|\nu\| \, \|\Delta F\| = |\nu' \Delta F| = O(\|\Delta F\|^2)$ or $\|\Delta F\|^{-1} = O(1)$, which is impossible for sufficiently small ΔF. This contradiction implies that (46) is satisfied. □

Remark. Using the above proof, we find that, under the same assumptions, if (46) does not hold, every solution (p, F, λ) in some neighborhood of (p^*, F^*, λ^*) satisfies $\|\Delta F\| = O(\|\Delta p\|^2 + |\Delta\lambda|^2)$. Since $p^* \neq 0$, it follows that $|\Delta\lambda| = O(\|\Delta p\|)$ and that there exists a positive constant c such that $\|\Delta p\| \geq c\|\Delta F\|^{\frac{1}{2}}$. In particular, there exists no Lipschitzian solution $p(F)$ in the neighborhood of (p^*, F^*). In fact, under Assumptions 4.1 and 4.2, there are only two possibilities. Either a Lipschitzian solution passes through the point and is unique, or every solution that passes through the point is not Lipschitzian. We conclude that every regular point must be simple regular.

COROLLARY 4.1. *Suppose that Assumptions* 4.1 *and* 4.2 *are satisfied. Then every jump point F^* is a member of the closed algebraic set S_1 given by*

$$S_1 := \{F \in \mathcal{R}^r : (\exists p \in \mathcal{R}^m, \lambda \in \mathcal{R}), c(p, F) = 0, c_p = \lambda p' D, |J(p, F, \lambda)| = 0\}.$$

ASSUMPTION 4.3. *The set S_1 does not span the entire space; that is, $S_1 \neq \mathcal{R}^r$.*

This assumption is equivalent to the claim that the elimination of p, λ from the $m+2$ equations defining S_1 results in a nonidentity polynomial equation. As a result, $\dim(S_1) < r$, and S_1 has zero Lebesgue measure in $\mathcal{R}^r$.

5. Necessary conditions for max-min: The largest uncertainty possible. Next we turn to the necessary conditions for max-min points. Let F^* be optimal in the sense of (33) with l tangent points, all simple regular (see Definition 4.2), denoted by $p^{(i)}(F^*), i = 1, 2, \ldots, l$, and Lagrange multipliers $\lambda^{(i)}(F^*)$. Suppose Assumption 4.1 is satisfied. Then, at each tangent point, the functions

$$\gamma^{(i)2}(F) := p^{(i)'}(F) D p^{(i)}(F) \text{ are in } C^1(N_{F^*}). \tag{52}$$

Define the Jacobian

$$J_3(F) = \frac{1}{2}\left[\frac{\partial \gamma^{(i)2}}{\partial F_k}\right] \in \mathcal{R}^{l \times r}, \tag{53}$$

and denote $r(J_3) = \text{rank} J_3(F^*)$.

Using $\gamma^{(i)2} = p' D p$, we have $d(\gamma^{(i)2}) = 2p' D dp$. Multiplying by λ, we have $\lambda d(\gamma^{(i)2}) = 2\lambda p' D dp$. But by (42), $c_p = \lambda p' D$. Thus $\lambda d(\gamma^{(i)2}) = 2c_p dp$. On the other hand, $c(p, F) = 0$ implies $c_p dp + c_F dF = 0$. Thus $\lambda d(\gamma^{(i)2}) = -2c_F dF$ or $\frac{\partial(\gamma^{(i)2})}{\partial F} = -2\lambda^{-1} c_F$.

Thus

$$J_3(F) = \frac{1}{2}\left[\frac{\partial \gamma^{(i)2}}{\partial F_k}\right]_{i,k} = \begin{bmatrix} \cdot \\ -(\lambda^{(i)})^{-1} c_F(p^{(i)}, F) \\ \cdot \end{bmatrix} \overset{\text{rank}}{=} \begin{bmatrix} \cdot \\ c_F(p^{(i)}, F) \\ \cdot \end{bmatrix}$$

$$\overset{\text{rank}}{=} \left[\frac{\partial c(p^{(i)}, F)}{\partial F_k}\right]_{i,k}.$$

Clearly, at $F = F^*$,

$$\gamma^{(i)2} = p^{(i)'} D p^{(i)}, \quad i = 1, \ldots, l. \tag{54}$$

Here we have l equations, and (42),

$$c(p^{(i)}, F^*) = 0, \quad c_p(p^{(i)}, F^*) = \lambda^{(i)} p^{(i)'} D, \quad i = 1, \dots, l, \tag{55}$$

yield another $l(m+1)$ equations for l tangent points. Thus we have $l(m+2)$ equations in $l(m+1)+r+1$ unknowns. We conclude that, if the equations are independent, we obtain a finite number of solutions, provided $l \geq r+1$. To cover all possibilities, we present the following cases.

Case 1. $l \geq r+1$. In this case, it suffices to consider $l = r+1$. Note that, in (54), the vectors $p^{(i)}, i = 1, \dots, r+1$, must be distinct. For more details, see the example in section 8. The set of equations takes the following form:

$$\boxed{\begin{aligned} &\gamma^{*2} = p^{(i)'} D p^{(i)}, \\ &c(p^{(i)}, F^*) = 0, \\ &c_p(p^{(i)}, F^*) = \lambda^{(i)} p^{(i)'} D, \quad i = 1, \dots, r+1. \end{aligned}} \tag{56}$$

Case 2. $r(J_3) = 0$, $1 \leq l \leq r$. In this case, $J_3(F^*) = 0$; that is, $c_F(p^{(i)}, F^*) = 0$. Note that $J_3(F^*) = 0$ defines lr equations, (42) defines $l(m+1)$ equations, and (54) defines another l equations. Thus we have $l(m+r+2)$ equations in $l(m+1)+r+1$ unknowns. We conclude that, for $l = 1$, we obtain $m+r+2$ equations in $m+r+2$ unknowns. For $l \geq 2$, we obtain more equations than unknowns, and, provided these equations are independent, we conclude that the solution set is empty for almost every nominal plant. In other words, we have a single simple regular tangent point ($l = 1$) for almost all nominal plants. We conclude that $c_F = 0$ and obtain

$$\left.\begin{array}{|l|l|}\hline c(p, F^*) = 0 & (1 \text{ equation}) \\ c_p(p, F^*) = \lambda p' D & (m \text{ equations}) \\ c_F(p, F^*) = 0 & (r \text{ equations}) \\ \hline \end{array}\right\} \begin{array}{l} (m+r+1) \text{ equations in} \\ (m+r+1) \text{ unknowns.} \end{array} \tag{57}$$

Remark. This is the classical case in which $\gamma(F) \in C^1(N_{F^*})$.

Case 3. $1 \leq r(J_3) \leq l \leq r$. In this case, J_3 has $r(J_3)$ independent rows. Without loss of generality, we may assume that rows $i = 1, \dots, r(J_3)$, are independent and that the rest are their linear combinations. Using $\Delta F := F - F^*$, we have

$$\begin{bmatrix} \gamma^{(1)2}(F) \\ \vdots \\ \gamma^{(l)2}(F) \end{bmatrix} - \begin{bmatrix} \gamma^{(1)2}(F^*) \\ \vdots \\ \gamma^{(l)2}(F^*) \end{bmatrix} = J_3(F^*)\Delta F + o(\|\Delta F\|), \quad F \in N_{F^*}. \tag{58}$$

We first prove that $r(J_3) = l$ is impossible. If $r(J_3) = l$, $J_3(F^*)$ has l independent columns. Thus there exists $v \in \mathcal{R}^r$ such that $J_3(F^*)v = \mathrm{col}[1 \cdots 1]$. For sufficiently small $\varepsilon > 0$, we choose $\Delta F = \varepsilon v$ and obtain

$$\begin{bmatrix} \gamma^{(1)2}(F) \\ \vdots \\ \gamma^{(l)2}(F) \end{bmatrix} - \begin{bmatrix} \gamma^{(1)2}(F^*) \\ \vdots \\ \gamma^{(l)2}(F^*) \end{bmatrix} = \begin{bmatrix} \varepsilon \\ \vdots \\ \varepsilon \end{bmatrix} + 0(\varepsilon) > 0. \tag{59}$$

That is $\gamma^{(i)}(F) > \gamma^{(i)}(F^*) := \gamma(F^*), i = 1, \dots, l$. Consequently (see the appendix for discussion), $\gamma(F) = \min_{1 \leq i \leq l} \gamma^{(i)}(F) > \gamma(F^*)$, contradicting the optimality of $\gamma(F^*)$.

We conclude that $1 \le r(J_3) < l \le r$, which also implies $r \ge 2$. Now we express the rows $i = r(J_3)+1, \dots, l$, in $J_3(F^*)$ as $[l - r(J_3)]$ linear combinations of the first $r(J_3)$ rows. This generates $(l-r(J_3))r$ equations in $(l-r(J_3))r(J_3)$ coefficients. In addition, we have (55). Thus we have a total of $(l-r(J_3))r + l + l(m+1) = l(m+r+2) - rr(J_3)$ equations in $(l-r(J_3))r(J_3)+l(m+1)+r+1$ unknowns. A simple calculation shows that the number of equations is bigger than the number of unknowns by $(l - r(J_3) - 1)(r - r(J_3) + 1)$. Thus the number of equations equals the number of unknowns only if $l = r(J_3) + 1$. Note that, in all other cases, $r(J_3) + 2 \le l \le r$, in which $l - r(J_3) - 1 \ge 1$ and $r - r(J_3) + 1 \ge 3$. In these cases, we find at least three extra equations. We conclude that, if the equations are independent, if $l \ge r(J_3) + 2$, the solution set is empty for almost every nominal plant, and it suffices to consider $l = r(J_3) + 1$, in which $l \le r(J_3) \le r - 1$ and $r \ge 2$. Thus

$$\boxed{\begin{array}{l} \text{for each } r(J_3) = 1, \dots, r-1, \\ \left.\begin{array}{l} \gamma^{*2} = p^{(i)'} D p^{(i)}, \\ c(p^{(i)}, F^*) = 0, \\ c_p(p^{(i)}, F^*) = \lambda^{(i)} p^{(i)'} D, \end{array}\right\} i = 1, \dots, r(J_3)+1, \\ c_F(p^{(i)}, F^*)|_{i=r(J_3)+1} = \displaystyle\sum_{i=1}^{r(J_3)} \mu_i c_F(p^{(i)}, F^*). \end{array}} \tag{60}$$

That is, for each $r(J_3) = 1, \dots, r-1$, we obtain $r + (m+2)(r(J_3)+1)$ equations in the same number of unknowns. We summarize our results as follows.

THEOREM 5.1. *Consider max-min problem* (26) *or* (31)*, with Assumptions* 4.1*. Suppose that, at a max-min point (optimum), all tangent points are simple regular. Then an optimal compensator satisfies at least one of the following sets of equations:* (56), (57), (60).

Remark. Theorem 4.1 characterizes simple regular points. However, we do not use it in Theorem 5.1.

COROLLARY 5.1. *The results in Theorem* 5.1 *remain unchanged if the constraint in* (31) *is a* C^1 *function not necessarily a polynomial. If the cost* $\gamma^2 = p'Dp$ *is replaced by a more general* C^1 *function, say,* $h(p, F)$*, we replace, in* (57) *and in* (60)*,* $[c_F]$ *by* $[c_F - \lambda h_F]$*. Likewise,* $c_p = \lambda p'D$ *is replaced by* $c_p = \lambda h_p$*.*

6. An alternative approach. Consider the max-min problem (26) or (31)

$$\operatorname*{Max}_{F} \operatorname*{Min}_{p} \gamma^2 = \sum_{i=1}^{m} \nu_i^2 p_i^2 := p'Dp \quad \text{s.t. } c(p, F) = 0. \tag{61}$$

Assuming that the critical polynomial $c(\cdot)$ does not vanish with its gradient, we first perform the minimization part. Define the Lagrangian

$$\mathcal{L} = p'Dp + \lambda c(p, F). \tag{62}$$

Then necessary conditions for the minimum are

$$\begin{aligned} &\mathcal{L}_p := \nabla_p \mathcal{L}(p, F) = 2p'D + \lambda c_p = 0, \\ &\mathcal{L}_\lambda = c(p, F) = 0. \end{aligned} \tag{63}$$

In order to perform the second operation, namely, the maximization with respect to F, we must, in principle, solve (63) for (p, λ), substitute p in γ^2, and maximize with respect to F. To do so, we set the system of polynomial equations

$$\begin{aligned} &\gamma^2 - p'Dp = 0, \\ &2p'D + \lambda c_p(p, F) = 0, \\ &c(p, F) = 0. \end{aligned} \tag{64}$$

Using the Grobner bases [13] to eliminate (p, λ), we obtain the polynomial equation

$$\varphi(\gamma^2, F) = 0. \tag{65}$$

The implicit function $\varphi(\cdot)$ is a polynomial in γ^2 whose coefficients are polynomials in F. To solve

$$\operatorname*{Max}_{F} \ \gamma^2 \quad \text{s.t. } \varphi(\gamma^2, F) = 0, \tag{66}$$

we recall that, while the zeros γ^2 of $\varphi(\cdot, F)$ are continuous functions of F, the function $\gamma^2(F)$ obtained from $\varphi(\gamma^2, F) = 0$ is not single-valued. Thus the Lagrange multiplier method does not apply to (66) directly. To see this, define the Lagrangian (with a "new" multiplier λ) as

$$\mathcal{L} = \gamma^2 + \lambda\varphi(\gamma^2, F). \tag{67}$$

Denote $\varphi' := \frac{\partial\varphi}{\partial(\gamma^2)}$; we have the following necessary conditions for maximum:

$$\begin{aligned} &\frac{\partial\mathcal{L}}{\partial(\gamma^2)} = 1 + \lambda\varphi'(\gamma^2, F) = 0, \text{provided } \varphi' \neq 0, \\ &\frac{\partial\mathcal{L}}{\partial F_i} = \lambda\frac{\partial\varphi}{\partial F_i} = 0. \end{aligned}$$

In other words

$$\boxed{\begin{aligned} &\varphi(\gamma^2, F) = 0, \\ &\frac{\partial\varphi}{\partial F_i} = 0, \quad i = 1, \ldots, r. \end{aligned}} \tag{68}$$

However, at F^*, a max-min point, $\varphi = 0$ and $\varphi' = 0$ may hold simultaneously. Thus we consider the following set of possibilities. Denote $\varphi^{(i)} := (\frac{\partial}{\partial(\gamma^2)})^i\varphi(\gamma^2, F)$, and consider, for $k = 1, \ldots, r-1$,

$$\operatorname*{Max}_{F} \ \gamma^2 \quad \text{s.t. } \varphi^{(i)} = 0, \quad i = 0, 1, \ldots, k. \tag{69}$$

We assume here that, at F^*, $\varphi^{(k+1)} \neq 0$. Define the Lagrangian

$$\mathcal{L} = \gamma^2 + \sum_{i=0}^{k} \lambda_i\varphi^{(i)}. \tag{70}$$

Then

(71)
$$\boxed{\begin{aligned}&\text{For each } k = 1, \dots, r-1,\\ &\frac{\partial \mathcal{L}}{\partial \lambda_i} = \varphi^{(i)} = 0, \qquad i = 0, 1, \dots, k,\\ &\frac{\partial \mathcal{L}}{\partial F_j} = \sum_{i=0}^{k} \lambda_i \frac{\partial \varphi^{(i)}}{\partial F_j} = 0, \quad j = 1, \dots, r.\end{aligned}}$$

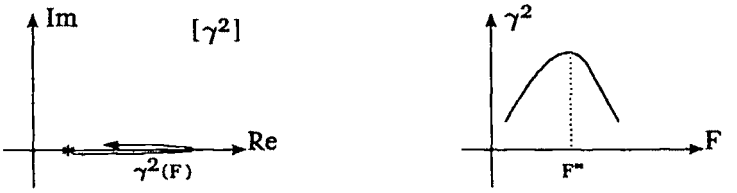
In the case when $k = r$ in (69), the constraints define F^*. In that case, we have the following conditions:

(72)
$$\boxed{\varphi^{(i)} = 0, \quad i = 0, 1, \dots, r.}$$

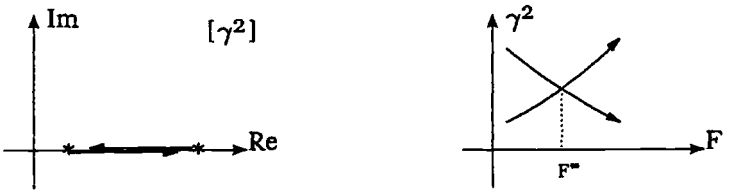
THEOREM 6.1. *Consider max-min problem* (61). *Suppose that Assumption* 4.1 *is satisfied and that* F *is in reduced form. Then necessary conditions for max-min are either* (68) *or* (71) *or* (72).

At this point, we wish to use the so-called root locus to illustrate both regular and singular max-min points. Recall that $\phi(\gamma^2, F)$ is a polynomial in γ^2 whose coefficients are functions of F. Denote by $\gamma^2(F)$ the roots γ^2 of $\varphi(\gamma^2, F) = 0$ as a function of F. Consider three basic cases.

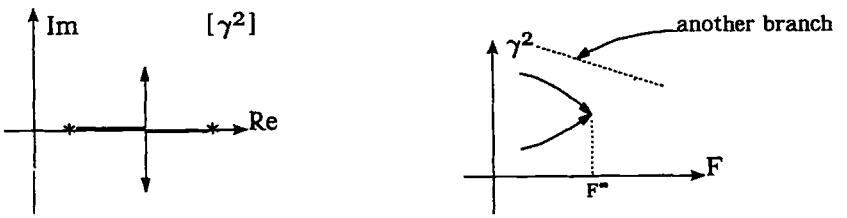
Case 1. A single root moves (as a function of F) in the positive direction, and backward.



Case 2. A positive real pair of roots move "head on" and pass each other.



Case 3. A positive real pair of roots move "head on" and split into a complex pair.



7. Necessary conditions for max-min: The uncertainty region is given. In the previous section, we have discussed the largest ellipsoid contained in the stability region. To discuss a given uncertainty region, consider max-min formulation (27):

$$\underset{F}{\text{Max}}\ \underset{p}{\text{Min}}\ \sum_{i=1}^{m} \nu_i^2 p_i^2 \quad \text{s.t. } c(p,F) - \rho^2 = 0. \tag{73}$$

We now solve a series of max-min problems (73) by increasing ρ^2 from zero. At each step, we calculate the measure $\gamma^{*2} = \sum_{i=1}^{m} \nu_i^2 p_i^2$. The process is stopped at the first time when γ^* becomes the given uncertainty radius.

To present a direct approach, we assume that there exists a robust controller F and recall (28):

$$\underset{F}{\text{Max}}\ \underset{p}{\text{Min}}\ c(p,F) \ \text{ s.t. } \sum_{i=1}^{m} \nu_i^2 p_i^2 - \gamma^2 \le 0. \tag{74}$$

Note that this problem can be handled using the results presented in [3, 4]. For additional comments, see section 9.

To solve (74) using our previous results, we first address the inner operation; namely, we fix F and solve

$$\underset{p}{\text{Min}}\ c(p;F) \ \text{ s.t. } \sum_{i=1}^{m} \nu_i^2 p_i^2 - \gamma^2 \le 0. \tag{75}$$

Since the *constraint qualification* holds for this problem, we can use Kuhn–Tucker conditions as necessary conditions. In particular, there exists a Lagrange multiplier λ such that

$$\begin{aligned} &\text{(i)} \quad c_p(p,F) + 2\lambda p' D = 0, \\ &\text{(ii)} \quad \lambda(p' D p - \gamma^2) = 0. \end{aligned} \tag{76}$$

We remind the reader that the inner operation in problem (31), namely $\gamma(F)$, may result in a discontinuous function. In the present formulation, the inner operation is always continuous. The reason lies in the facts that Ω is a connected set and $c(p,F)$ is continuous. In contrast to the present situation, in (31), we search for the largest ellipsoid contained in the stability region. Since the stability region may be disconnected, it follows that $\gamma(F)$ may be discontinuous. We conclude that our present formulation is *regular*. We now proceed as in section 4. In particular, we replace (56) by

$$\boxed{\begin{aligned} &c_p(p^{(i)},F) + 2\lambda^{(i)} p^{(i)\prime} D = 0, \qquad i = 1,2,\dots r+1, \\ &\lambda^{(i)}(p^{(i)\prime} D p^{(i)} - \gamma^2) = 0, \qquad i = 1,2,\dots,r+1, \\ &c(p^{(i)},F) = c(p^{(1)},F), \qquad i = 2,3,\dots,r+1. \end{aligned}} \tag{77}$$

Likewise, we replace (57) and (60) by proper sets of equations. Note that some $\lambda^{(i)}$ may vanish at optimum (interior point). Before closing this section, we wish to comment on the construction of a fixed order compensator for a nominal plant (with

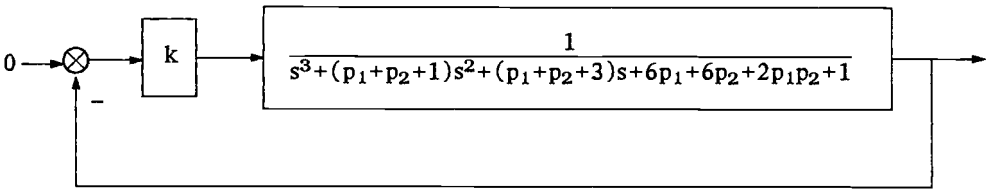


FIG. 5.

no uncertainty). Suppose our objective is to find a compensator F_0 so as to maximize the "distance" to the stability boundary. Then we may write

$$\max_{F_0} \min_{F} \ (F - F_0)' D (F - F_0) \quad \text{s.t. } c(F) = 0. \tag{78}$$

Remarks.

1. Since $c(\cdot)$ is a function of F alone, a jump never occurs.
2. We may interpret the formulation (78) as the "maximum stability margin" possible.
3. Case 2 given by (57) is impossible for (78) since a single contact point ($l = 1$) is impossible.
4. From a practical point of view, we require F to be bounded. In particular, suppose that we require $c_1(F) := F' D_1 F - \rho^2 \leq 0$. Then, in order to achieve bounded control parameters (especially in the case when $\hat{\aleph}$ is unbounded), we replace the constraint $c(F) = 0$ by $c(F)c_1(F) = 0$.

8. Illustrative examples.

Example 1. Consider the unit feedback system (a feedback version of [14]) pictured in Figure 5.

The closed loop characteristic polynomial is

$$\Delta_c(s; p, F) = s^3 + (p_1 + p_2 + 1)s^2 + (p_1 + p_2 + 3)s + (6p_1 + 6p_2 + 2p_1p_2 + 1 + k).$$

Recall that the critical polynomial can be generated from the Hurwitz matrix as follows:

$$c(p, F) = |H_n(p, F)| = a_0(p, F)|H_{n-1}(p, F)|,$$

where $c(p, F) = 0$ contains the stability boundary and, in particular,

(i) $a_0(p, F) = 0$ contains the real root boundary (r.r.b.),

(ii) $|H_{n-1}(p, F)| = 0$ contains the complex root boundary (c.r.b.).

In the present example,

(i) $a_0(p, F) = 6p_1 + 6p_2 + 2p_1p_2 + 1 + k$,

(ii) $|H_2(p, F)| = (p_1 - 1)^2 + (p_2 - 1)^2 - k$.

The stability region is defined by $a_0 > 0$, $a_2 > 0$, and $|H_2| = a_1a_2 - a_0 > 0$ and is depicted in Figure 6.

Our objective is to select the control gain k so as to maximize the robustness radius with respect to the origin. First, note that, as k increases, the circle approaches the origin, while the hyperbola diverges from the origin. Thus max-min takes place at equal distances. Note that, since the circle radius is $\sqrt{k}$, the circle is feasible only for $0 < k < 2$. Also note that the stability interval for the nominal system is $-1 < k < 2$.

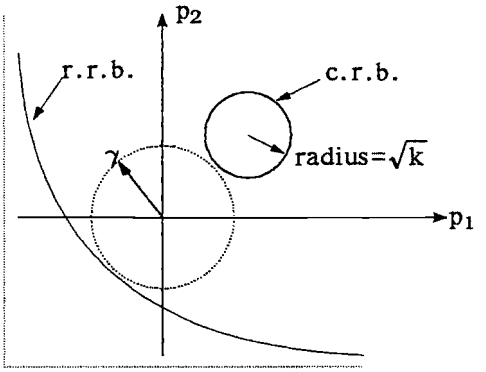


FIG. 6.

From symmetry, it is clear that the robustness circle is tangent to the hyperbola at $p_1 = p_2 < 0$. Indeed, $c_p + \lambda p' = 0$ yields, after solving for λ, $p_2 \frac{\partial c}{\partial p_1} - p_1 \frac{\partial c}{\partial p_2} = 0$. A detailed calculation yields $(p_2 - p_1)(p_1 + p_2 + 3) = 0$, which means $p_1 = p_2 := p$. The hyperbola now yields $p^2 + 6p + \frac{1+k}{2} = 0$. The solution becomes $p = -3 + \sqrt{9 - \frac{k+1}{2}}$. Equating the distances from the origin to the circle and to the hyperbola, we obtain $-(-3 + \sqrt{9 - \frac{1+k}{2}})\sqrt{2} = \sqrt{2} - \sqrt{k}$. The two sides of this equation are depicted in Figure 7. Simple algebra yields $4k^2 - 68k + 81 = 0$. The optimal gain becomes $k^* = 1.29$, and the robustness radius is $\gamma^* = \sqrt{2} - \sqrt{k^*} = 0.27$. Finally, note that, for our example, a complete set of necessary conditions is given by (56). In particular, since $r = 1$, we have $l = 2$. Denote $p^{(1)} = (x, y)$, $p^{(2)} = (v, w)$, and $g = \gamma^2$. Then, eliminating λ, set (56) takes the form

(1) $g - x^2 - y^2 = 0$,
(2) $g - v^2 - w^2 = 0$,
(3) $(6x + 6y + 2xy + 1 + k)((x-1)^2 + (y-1)^2 - k) = 0$,
(4) $(6v + 6w + 2vw + 1 + k)((v-2)^2 + (w-1)^2 - k) = 0$,
(5) $y[(6 + 2y)((x-1)^2 + (y-1)^2 - k) + 2(x-1)(6x + 6y + 2xy + 1 + k)]$
$-x[(6 + 2x)((x-1)^2 + (y-1)^2 - k) + 2(y-1)(6x + 6y + 2xy + 1 + k)] = 0$,
(6) $w[(6 + 2w)((v-1)^2 + (w-1)^2 - k) + 2(v-1)(6v + 6w + 2vw + 1 + k)]$
$-v[(6 + 2v)((v-1)^2 + (w-1)^2 - k) + 2(w-1)(6v + 6w + 2vw + 1 + k)] = 0$.

Using the Grobner package in Maple V, we find that a Grobner basis is formed of 64 (yes, sixty four!) reduced sets of polynomial equations. In principle, we have to test each of them. To save space, we present the two relevant sets

(a) $x + w = 0$, $v - w = 0$, $2g - 16w + 3 = 0$, $2k - 8w - 1 = 0$, $y + w = 0$, $4w^2 - 16w + 3 = 0$,
(b) $x + w = 0$, $v - w = 0$, $2g + 16w + 3 = 0$, $2k + 8w - 1 = 0$, $y + w = 0$, $4w^2 + 16w + 3 = 0$.

Each of these sets yields $k^* = 1.3$.

Finally, we wish to present a solution based on the alternative approach of section 5. Once more, using the Grobner package in Maple V, we obtain from (64)–(65)

$$\varphi(\gamma^2, F) = \prod_{i=1}^{4} \varphi_i(\gamma^2, F),$$

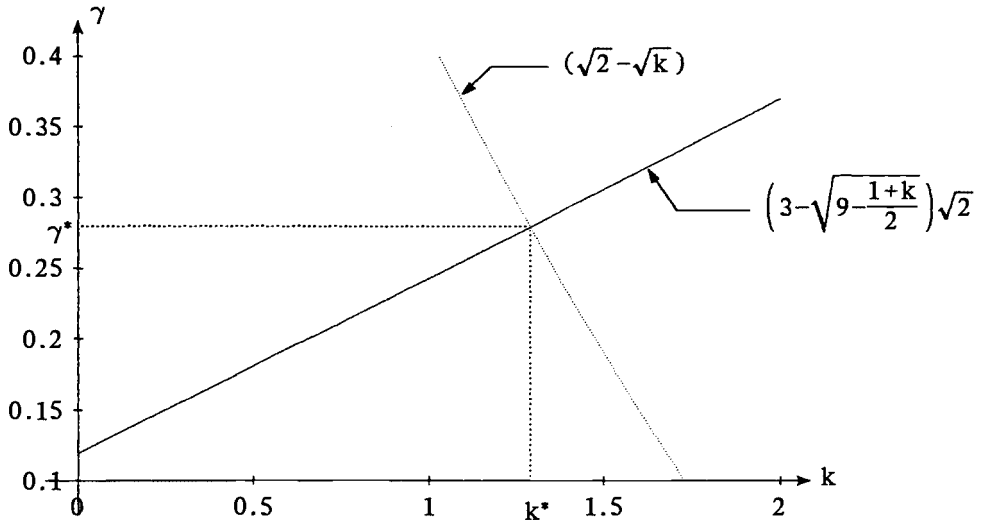


FIG. 7.

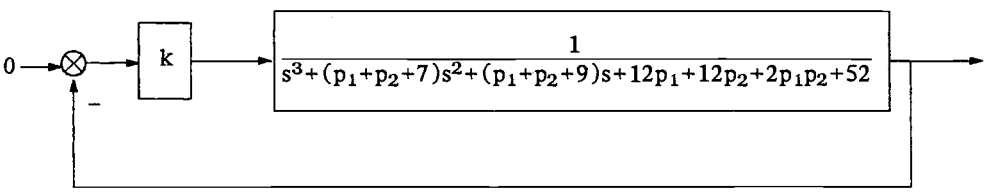


FIG. 8.

where

$$\varphi_1(\gamma^2, F) = (\gamma^2 + k + 1)^2 - 72\gamma^2, \quad \varphi_2(\gamma^2, F) = (\gamma^2 - k + 2)^2 - 8\gamma^2,$$
$$\varphi_3(\gamma^2, F) = \gamma^2 - k + 4, \quad \varphi_4(\gamma^2, F) = \gamma^2 - k + 8.$$

Using (72), we find that $\varphi(\gamma^2, F) = 0$ and $\varphi'(\gamma^2, F) = 0$ are satisfied (among other points) at $\varphi_1 = \varphi_2 = 0$. This yields $k^2 - 17k + 20.25 = 0$ or $k^* = 1.3$.

Example 2. Consider the unit feedback system described by Figure 8. Note that this example is obtained from Example 1 by a simple shift on the parameters:

$$(p_i)_o = (p_i)_n + 3, (k)_o = (k)_n + 3, \text{where } \ o = \text{old and } n = \text{new}.$$

The closed loop characteristic polynomial is

$$\Delta_c(s; p, F) = s^3 + (p_1 + p_2 + 7)s^2 + (p_1 + p_2 + 9)s + (12p_1 + 12p_2 + 2p_1p_2 + 52 + k).$$

The critical polynomial becomes

$$c(p, F) = (12p_1 + 12p_2 + 2p_1p_2 + 52 + k)((p_1 + 2)^2 + (p_2 + 2)^2 + 3 - k).$$

As a result, the stability boundary is formed of two branches—a circle and a hyperbola; see Figure 9. The circle (c.r.b.) disappears as k reaches the value 3. If we calculate the distance from each branch to the origin, we find the description of Figure 10.

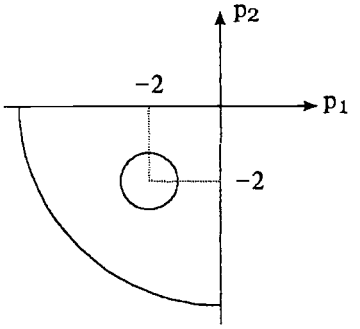


FIG. 9.

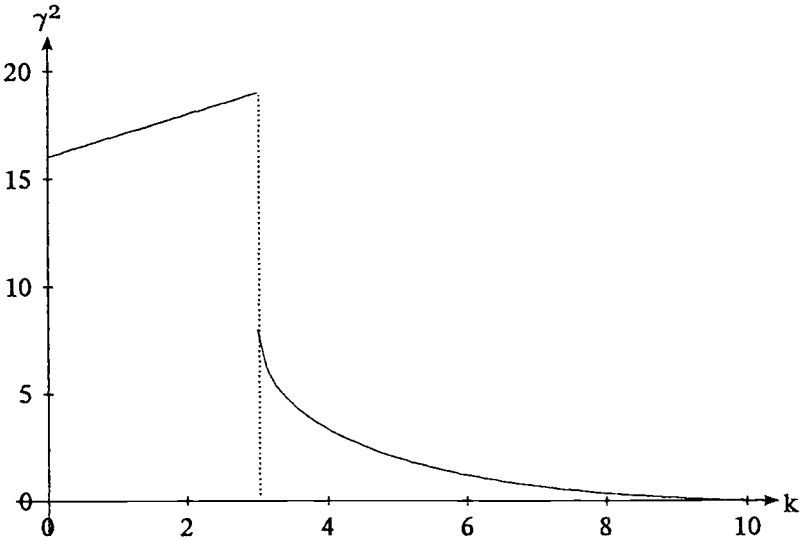


FIG. 10.

The graph illustrates that the robustness radius has a jump at $k = 3$. From an algebraic point of view, using the Grobner package in Maple 6, we obtain from (64)–(65)

$$\varphi(\gamma^2, F) = \prod_{i=1}^{5} \varphi_i(\gamma^2, F), \quad \text{where}$$

$$\varphi_1 = (\gamma^2)^2 + (2k - 184)\gamma^2 + k^2 + 2704 + 104k,$$

$$\varphi_2 = (\gamma^2)^2 - (2k + 10)\gamma^2 + k^2 - 22k + 121,$$

$$\varphi_3 = k - \gamma^2 + 16, \varphi_4 = k - \gamma^2 + 17, \varphi_5 = k - \gamma^2 + 25.$$

Using (72), we find that $\varphi(\gamma^2, F) = 0$ and $\varphi'(\gamma^2, F) = 0$ are satisfied (among other points) at $\varphi_2 = \varphi_2' = 0$. This yields $k = 3$ as expected. Clearly, this example belongs to Case 3, a singular case. It is singular in the sense that, at a max-min point, the stability radius γ has a jump, as shown in Figure 10. In fact, a closer look at

$\varphi_2 = \varphi_2' = 0$ reveals that the roots $\gamma^2(F)$ described by Figure 10 satisfy

$$\gamma^2(F) = \begin{cases} k + 16 & \text{for } -16 < k < 3, \\ k + 5 - 4\sqrt{2k-6} & \text{for } 3 < k < 11. \end{cases}$$

In closing this example, we note that, while the detection of the jump is simple, it illustrates a possible numerical difficulty. The circle tangent to the hyperbola is very close to the hyperbola along a large portion of the hyperbola.

Example 3. According to Theorem 5.1, one has to find all the solutions of the sets of equations (56), (57), and (60). In previous examples, the optimal solution belongs to set (56). The geometric reason is $l = r + 1$, where $l = 2$ is the number of tangent points and $r = 1$ is the number of control parameters. In the present example, $l = r = 2$. In particular, consider the uncertain system $\dot{x} = A(\alpha, \beta)x + B(\alpha, \beta)u$,

$$A = \begin{bmatrix} 0 & \alpha - 1 \\ \beta & 0 \end{bmatrix}, \quad B = \begin{bmatrix} \alpha \\ 1 - \beta \end{bmatrix},$$

where the uncertain parameters α and β lie in the set

$$\{(\alpha, \beta) : (\alpha - 0.5)^2 + (\beta - 0.5)^2 \leq \rho^2\}.$$

First, we transform (α, β) into (p_1, p_2) so that the nominal values $\alpha = 0.5$, $\beta = 0.5$ correspond to the origin $p_1 = 0$, $p_2 = 0$. To this end, let $p_1 = \alpha - 0.5$, $p_2 = \beta - 0.5$. Then

$$A = \begin{bmatrix} 0 & p_1 - 0.5 \\ p_2 + 0.5 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} p_1 + 0.5 \\ 0.5 - p_2 \end{bmatrix},$$

where the uncertainty set takes the form

$$\Omega = \{(p_1, p_2) : p_1^2 + p_2^2 \leq \gamma^2\}.$$

Our objective is to stabilize the above uncertain system using state feedback $u = Fx$, $F = [f_1 \ f_2]$ for all $(p_1, p_2) \in \Omega$ so as to maximize the robustness radius γ. To apply Theorem 5.1, we calculate the closed loop characteristic polynomial

$$\begin{aligned} \Delta_c(s; p, F) = s^2 &- [(p_1 + 0.5)f_1 + (0.5 - p_2)f_2]s \\ &- [(p_1 - 0.5)(p_2 + 0.5) + (p_1 - 0.5)(0.5 - p_2)f_1 \\ &+ (p_1 + 0.5)(p_2 + 0.5)f_2]. \end{aligned}$$

The stability constraints are

$$\begin{aligned} c_1 &= -[(p_1 + 0.5)f_1 + (0.5 - p_2)f_2] > 0 && \text{(r.r.b.)}, \\ c_2 &= -[(p_1 - 0.5)(p_2 + 0.5) + (p_1 - 0.5)(0.5 - p_2)f_1 \\ &\quad + (p_1 + 0.5)(p_2 + 0.5)f_2] > 0 && \text{(c.r.b.)}. \end{aligned}$$

Note that the open loop satisfies $s^2 = (p_1 - 0.5)(p_2 + 0.5)$. Thus, in the region $\{(p_1, p_2) : |p_i| \leq 0.5\}$, the roots of the characteristic polynomial are all located in the interval $[-1, 1]$ on the imaginary axis. Now we wish to study the effect of feedback control on the stability. To simplify the discussion, let $F = [0 \ \ f_2]$. In this case,

$$\begin{aligned} 2c_1 &= -(1 - 2p_2)f_2 > 0, \\ 4c_2 &= -(1 + 2p_2)(-1 + 2p_1 + f_2 + 2p_1 f_2) > 0. \end{aligned}$$

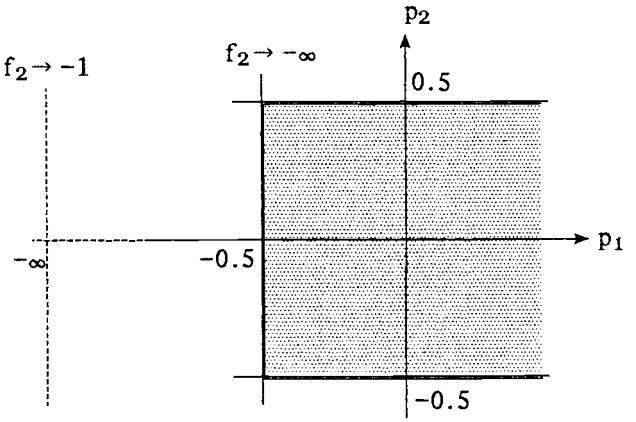


FIG. 11.

Thus the critical polynomial $c = c_1 c_2$ defines the following robust stability boundaries:

$$\{(p_1, p_2) : p_2 = 0.5\}, \quad \{(p_1, p_2) : p_2 = -0.5\}, \quad \{(p_1, p_2) : p_1 = 0.5(1 - f_2)/(1 + f_2)\}.$$

The stability region is depicted in Figure 11. The stabilizing gain is $f_2 \in (-\infty, -1)$. In other words, the use of f_2 in this interval assures stability in the shaded region.

Since the robustness radius ($\gamma^* = 0.5$) is invariant with respect to $f_2 \in (-\infty, -1)$, the optimization process does not detect any optimal point. Since the open loop is marginally stable for all $|p_i| \leq 0.5$, we modify our original objective as follows: Find $u = Fx$ such that the closed loop is stable relative to $\mathrm{Re}(s) < -0.1$. We continue with the control structure $F = [0 \;\; f_2]$. The modified characteristic polynomial becomes

$$\begin{aligned}\Delta(s - 0.1) = s^2 &- [0.2 + (0.5 - p_2)f_2]s \\ &+ [0.26 - 0.2f_2 - 0.6p_2 f_2 - p_1 p_2 - 0.5p_1 + 0.5p_2 - p_1 p_2 f_2 - 0.5p_1 f_2].\end{aligned}$$

The (relative) stability constraints are now

$$\begin{aligned}c_1 &= -0.2 - (0.5 - p_2)f_2 > 0, \\ c_2 &= 0.26 - 0.2f_2 - 0.6p_2 f_2 - p_1 p_2 - 0.5p_1 + 0.5p_2 - p_1 p_2 f_2 - 0.5p_1 f_2 > 0.\end{aligned}$$

Clearly, given f_2, $c_1 = 0$ describes a straight line, while $c_2 = 0$ describes a hyperbola. We suppose that, at the maximum robustness radius, there are two tangent points—one on the line and one on the hyperbola. The distance from the origin to the line is $0.5 + 0.2/f_2$. Let $(p_1, p_2) = (x, y)$ be the tangent point with the hyperbola. Then

$$(1) \quad x^2 + y^2 = (0.5 + 0.2/f_2)^2.$$

The tangency condition at (x, y) is

$$(2) \quad x(0.5 - x - xf_2 - 0.6f_2) - y(-0.5 - y - yf_2 - 0.5f_2) = 0,$$

and the hyperbola satisfies

$$(3) \quad c_2(x, y, f_2) = 0.$$

Solving these equations, we obtain $x = -0.02$ and $y = -0.39$, and the optimal gain is $f_2 = -1.83$. The maximum robustness radius is $\gamma^* = 0.39$.

Now we proceed to the general case where $F = [f_1 \quad f_2]$. As before, we are interested in stability with respect to $\mathrm{Re}(s) < -0.1$. Since we do not have a priori information about the number of tangent points, we assume, as before, two tangent points—one on the line and one on the hyperbola. However, in order to use the simple set of equations (56), we assume three tangent points so that $l = r + 1$ and solve (56) numerically. If our hypothesis concerning two tangent points is correct, two points must be identical. Let the point on the line be (x_1, x_2) and on the hyperbola (x_3, x_4), (x_5, x_6). Then (56) becomes

(1) $-0.8 - 4x_1f_1 - 2f_1 - 2f_2 + 4x_2f_2 = 0$,

(2) $1.04 - 1.6x_3f_1 + 1.2f_1 - 0.8f_2 - 2.4x_4f_2 - 4x_3x_4 - 2x_3 + 2x_4 + 4x_3x_4f_1 - 2x_4f_1 - 4x_3x_4f_2 - 2x_3f_2 = 0$,

(3) $1.04 - 1.6x_5f_1 + 1.2f_1 - 0.8f_2 - 2.4x_6f_2 - 4x_5x_6 - 2x_5 + 2x_6 + 4x_5x_6f_1 - 2x_6f_1 - 4x_5x_6f_2 - 2x_5f_2 = 0$,

(4) $x_1f_2 + x_2f_1 = 0$,

(5) $x_3(-2.4f_2 - 4x_3 + 2 + 4x_3f_1 - 2f_1 - 4x_3f_2) - x_4(-1.6f_1 - 4x_4 - 2 + 4x_4f_1 - 4x_4f_2 - 2f_2) = 0$,

(6) $x_5(-2.4f_2 - 4x_5 + 2 + 4x_5f_1 - 2f_1 - 4x_5f_2) - x_6(-1.6f_1 - 4x_6 - 2 + 4x_6f_1 - 4x_6f_2 - 2f_2) = 0$,

(7) $x_1^2 + x_2^2 - x_3^2 - x_4^2 = 0$,

(8) $x_1^2 + x_2^2 - x_5^2 - x_6^2 = 0$.

Solving numerically this set of equations using Maple 6, we find three real solutions:

$$
\begin{array}{lll}
(x_1, x_2) = (0.052, 0.433), & (0.044, 0.426), & (0.027, 0.412), \\
(x_3, x_4) = (0.027, -0.435), & (0.013, -0.l428), & (-0.006, -0.413), \\
(x_5, x_6) = (0.027, -0.435), & (0.013, -0.428), & (-0.006, -0.413), \\
(f_1, f_2) = (23.91, -199.9), & (1.102, -10.76), & (0.247, -3.749).
\end{array}
$$

Note, of course, that there are more real solutions to this set of equations. The first solution $(f_1, f_2) = (24, -200)$ induces a robustness radius $\gamma^* = 0.44$. This is a 12.5% increase with respect to the previous $(f_1, f_2) = (0, -1.83)$ with $\gamma^* = 0.39$. The fact that there are two identical tangent points implies that neither (56) nor (60) can be used directly. Indeed, these sets of equations are based on the assumption (which may not be necessary) of simple roots. Moreover, Maple solves the above set of equations using a gradient method, starting at some initial point. In other words, the solution containing a multiple root is approached approximately. On the other hand, the use of a homotopy method cannot detect a multiple root since homotopy curves are not allowed to intersect. Still, it is possible to use the homotopy approach indirectly by first approximating the above set of equations and then decreasing the approximation parameter.

9. Conclusions. The most important consequence of this paper is the fact that max-min completely characterizes the largest (or a fixed) ellipsoid contained in the (relative) stability region in the parameter space. As a result, we have a nonconservative approach to feedback compensation in the presence of a real parameter uncertainty. Since in max-min the inner operation results in a nondifferential function, we cannot use for the outer operation the Lagrange multiplier method. Instead, we have found that, under mild conditions, the number of extreme points is related to the (reduced) number of the controller parameters. Moreover, the largest uncertain

ellipsoid might result in a jump. In such a case, the performance of the "robust" controller is sensitive to (small) changes in control parameters. This *singular* situation can be handled using our alternative approach; however, due to elimination, it is numerically expensive. On the other hand, in the case of a fixed uncertain ellipsoid, a jump never occurs. It is worth noting that the largest uncertain ellipsoid and the fixed uncertain ellipsoid are related via limiting processes. In particular, solving (73) and increasing ρ^2 from zero, we can approach the required fixed radius. On the other hand, solving $\mathrm{Max}_F\ \mathrm{Min}_{p'Dp=r^2} c(p,F)$ by increasing r from zero, we can approach the maximum radius. The limiting process, however, requires, in principle, the solution of an infinite number of max-min problems. To achieve performance, we use relative rather than asymptotic stability. Our approach can handle both polynomial and matrix versions—the latter can be transformed into the former. The resulting necessary conditions have the form of a set of nonlinear equations. We suggest the use of a probability one homotopy method (HOMPACK [6]) for solving these equations (see also [15, 16, 17]). In the application to control theory, the functions $h(\cdot)$ and $c(\cdot)$ are usually polynomials. In this important case, the necessary conditions have the form of a set of polynomial equations, the total degree is simple to calculate, and the use of HOMPACK is simple. Moreover, in the polynomial case, the use of Grobner bases simplifies these equations and enables us to solve nontrivial control problems. Finally, the singular case needs more study. This is left for future research.

Appendix. In section 5, Case 3, we use the following result.

LEMMA A.1. *The following is satisfied in some neighborhood of F^*:*

$$\gamma(F) = \min_{1 \le i \le l} \gamma^{(i)}(F).$$

Proof. We have to show that at least one of the points $p^{(i)}(F)$ is indeed a contact point and not just one satisfying the tangency condition (42). Let

$$\|p\|_D := (p'Dp)^{1/2} = \|D^{1/2}p\|.$$

Since D is a positive definite (p.d.) matrix, it follows that $\|p\|_D$ can be used as a norm of p. Note that, if p, F, λ satisfy (42) and $p \neq 0$, then λ is uniquely determined by

$$\lambda(p,F) = \frac{c_p p}{p'Dp} = \frac{c_p p}{\|p\|_D^2}.$$

From Assumption 4.1 it follows that $\gamma^* := \gamma(F^*) = \mathrm{Max}_F \gamma(F)$. Thus F^* is a stabilizing controller for the nominal system; that is, $(0,F^*) \in \hat{\aleph}$, and, in particular, $c(0,F^*) > 0$.

Denote $S(F) := \{p^{(i)}(F)\colon\ i = 1,\dots,l\}$ in some neighborhood of F^*. By construction, $S(F^*)$ is the set of all contact points for F^*, all simple regular. Thus there exists an open set U containing $S(F^*)$ such that, for each F in some neighborhood of F^*, if p, F, λ, satisfy (42) with $p \in U$, then $p \in S(F)$. Denote $B := \{p \in \mathcal{R}^m : \|p\|_D \le \gamma^*\}$. By construction, $S(F^*) \subset U$. Thus $p \in B\backslash U$ implies $p \notin S(F^*)$, and $p \in B\backslash U \Longrightarrow c(p,F^*) \neq 0$. We conclude that the compactness of $B\backslash U$ and the continuity of $c(p,F)$ imply that there exists an open set V containing $B\backslash U$ such that, for each F in some neighborhood of F^*, $p \in V \Rightarrow c(p,F) \neq 0$.

Now denote $W = U \cup V$. Then W^c, the complement of W, is a closed set (nonempty, since we may assume that W is bounded). Thus we may denote $\gamma_+ :=$

$\mathrm{Min}_{p\in W^c}\|p\|_D$. Since $B\backslash U \subset V$, it follows that $B \subset W$; thus $p \in W^c \Rightarrow p \notin B \Rightarrow \|p\|_D > \gamma^*$, and $\gamma_+ > \gamma^*$.

Since F^* stabilizes the nominal system, and since the zeros of $\Delta(\cdot; 0, F)$ vary continuously with F, it follows that, for each F in some neighborhood of F^*, $(0, F) \in \hat{\aleph}$. Since $\|p^{(1)}(F^*)\|_D = \gamma^* < \gamma_+$, it follows from the continuity of $p^{(1)}(F)$ at F^* that there exists a neighborhood U_{F^*} of F^* for which $\gamma(F) = \mathrm{Min}_{c(p,F)=0}\|p\|_D \leq \|p^{(1)}(F)\|_D < \gamma_+$, $F \in U_{F^*}$. Thus, if $F \in U_{F^*}$ with a contact point $p(F)$, it follows that $\|p(F)\|_D = \gamma(F) < \gamma_+ = \mathrm{Min}_{p\in W^c}\|p\|_D$; thus $p(F) \notin W^c$; that is, $p(F) \in W = U \cup V$.

Now, if $p(F) \in V$, it follows from the definition of V (with U_{F^*} sufficiently small) that $c(p(F), F) \neq 0$, contradicting the fact that $p(F)$ is a contact point for F. We thus conclude that $p(F) \in U\backslash V$. Since $p(F)$ satisfies (42), it follows from the definition of U (with U_{F^*} sufficiently small) that $p(F) \in S(F)$. Thus $\gamma(F) = \|p(F)\|_D > \mathrm{Min}_{p\in S(F)}\|p\|_D$. On the other hand, $\gamma(F) = \mathrm{Min}_{c(p,F)=0}\|p\|_D \leq \mathrm{Min}_{p\in S(F)}\|p\|_D$. Thus we finally conclude:

$$\gamma(F) = \min_{p\,\in\, S(F)} \|p\|_D = \min_{1\,\leq\, i\,\leq\, l} \|p^{(i)}(F)\|_D, F \in U_{F^*}. \qquad \square$$

REFERENCES

[1] M. Fischer and S. Gutman, *The synthesis of fixed order compensators*, J. Optim. Theory and Appl., 70 (1991), pp. 331–352.

[2] S. Gutman, *Root Clustering in Parameter Space*, Lecture Notes in Control and Inform. Sci. 141, M. Thoma and W. Wyner, eds., Springer-Verlag, New York, 1990.

[3] V. F. Demyanov and V. N. Malozemov, *On the theory of nonlinear minimax problems*, Uspekhi Mat. Nauk, 26 (1971), pp. 53–104.

[4] V. F. Demyanov and V. N. Malozemov, *Introduction to Minimax*, Dover, New York, 1990.

[5] J. M. Danskin, *The Theory of Max-Min and Its Application to Weapons Allocation Problems*, Econometrics and Operations Research V, Springer-Verlag, New York, 1967.

[6] L. T. Watson, S. C. Billups, and A. P. Morgan, *Algorithm* 652: *HOMPACK: A suite of codes for globally convergent homotopy algorithms*, ACM Trans. Math. Software, 13 (1987), pp. 281–310.

[7] A. Vicino, A. Tesi, and M. Milanese, *Computation of nonconservative stability perturbation bounds for systems with nonlinearly correlated uncertainties*, IEEE Trans. Automat. Control, 35 (1990), pp. 835–841.

[8] R. A. Frazer and W. J. Duncan, *On the criteria for the stability of small motion*, Proceedings of the Royal Society A, 124 (1929), p. 642.

[9] E. I. Jury and T. Pavlidis, *Stability and aperiodicity constraints for system design*, IEEE Trans. Circuit Theory, 10 (1963), pp. 137–141.

[10] S. Gutman and F. Chojnowski, *Fixed and minimal compensators*, IMA J. Math. Control Inform., 7 (1991), pp. 361–373.

[11] S. Gutman, *Output feedback root clustering in parameter space*, Israel J. Technology, 21 (1983), pp. 81–84.

[12] H. Taub and S. Gutman, *Roots of composite polynomials—an application to root clustering*, Linear Algebra Appl., 87 (1987), pp. 181–188.

[13] W. W. Adams and P. Loustaunau, *An Introduction to Grobner Bases*, AMS, Providence, RI, 1994.

[14] J. Ackermann, H. Z. Hu, and D. Kaesbauer, *Robustness analysis: A case study*, IEEE Trans. Automat. Control, 35 (1990), pp. 352–356.

[15] S. N. Chow, J. Mallet-Paret, and J. A. Yorke, *Finding zeroes of maps: Homotopy methods that are constructive with probability one*, Math. Comp., 32 (1978), pp. 887–899.

[16] L. T. Watson, *A globally convergent algorithm for computing fixed points of C^2 maps*, Appl. Math. Comput., 5 (1979), pp. 297–311.

[17] L. T. Watson, *Engineering applications of the Chow–York algorithm*, Appl. Math. Comput., 9 (1981), pp. 111–133.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 1, pp. 303–312

POLE PLACEMENT BY STATIC OUTPUT FEEDBACK FOR GENERIC LINEAR SYSTEMS*

A. EREMENKO† AND A. GABRIELOV†

Abstract. We consider linear systems with m inputs, p outputs, and McMillan degree n such that $n = mp$. If both m and p are even, we show that there is a nonempty open (in the usual topology) subset U of such systems, where the real pole placement map is not surjective. It follows that, for each system in U, there exists an open set of pole configurations, symmetric with respect to the real line, which cannot be assigned by any real static output feedback.

Key words. linear systems, static output control feedback, pole placement

AMS subject classifications. 14N10, 14P99, 14M15, 30C99, 26C15

PII. S0363012901391913

1. Introduction.

We consider *linear systems* $S = (A, B, C)$ described by the equations

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ y &= Cx. \end{aligned} \tag{1.1}$$

Here the state x, the input u, and the output y are functions of a real variable t (time), with values in $\mathbf{R}^n$, $\mathbf{R}^m$, and $\mathbf{R}^p$, respectively, the dot denotes the derivative with respect to t, and A, B, C are real matrices of sizes $n \times n$, $n \times m$, and $p \times n$, respectively.

Assuming zero initial conditions and applying the Laplace transform, we obtain

$$Y(s) = C(sI - A)^{-1} B\, U(s),$$

so the behavior of our linear system is described by a rational matrix-function $C(sI - A)^{-1}B$ of size $p \times m$ of a complex variable s, which is called the (open loop) *transfer function* of S. It is clear that $G(\infty) = 0$. The poles of the transfer function are the eigenvalues of the matrix A.

For a given $p \times m$ matrix function G with the property $G(\infty) = 0$, there exist infinitely many representations of G in the form $G(s) = C(sI - A)^{-1}B$. The smallest integer n over all such representations is called the *McMillan degree* of G.

We consider the possibility of controlling a given system S by attaching a feedback. This means that the output is sent to the input after a preliminary linear transformation, called a *compensator*. The compensator may be another system of the form (1.1) (dynamic output feedback) or just a constant matrix (static output feedback). In this paper, we consider only static output feedback, referring for the recent results on dynamic output feedback to [14, 11].

A static output feedback is described by the equation

$$u = Ky, \tag{1.2}$$

*Received by the editors July 5, 2001; accepted for publication (in revised form) January 11, 2002; published electronically June 5, 2002.

http://www.siam.org/journals/sicon/41-1/39191.html

†Department of Mathematics, Purdue University, West Lafayette, IN 47907 (eremenko@math.edu, www.math.purdue.edu/∼eremenko, agabriel@math.purdue.edu, www.math.purdue.edu/∼agabriel).
The first author was supported by NSF grant DMS-0100512 and the Humboldt Foundation. The second author was supported by NSF grant DMS-0070666 and the James S. McDonnell Foundation.

where K is an $m \times p$ matrix which is usually called a *gain* matrix. Eliminating u and y gives

$$\dot{x} = (A + BKC)x,$$

whose characteristic polynomial is

$$\varphi_K(s) = \det(sI - A - BKC). \tag{1.3}$$

It is called the *closed loop characteristic polynomial.*

The pole placement problem is formulated as follows.

Given a system $S = (A, B, C)$ and a set of points $\{s_1, \dots, s_n\}$ in $\mathbf{C}$ (listed with multiplicities) symmetric with respect to the real axis, find a real matrix K such that the zeros of φ_K are exactly $s_1, \dots, s_n$.

For a fixed system S, we define the (real) *pole placement map*

$$\chi_S : \mathrm{Mat}_{\mathbf{R}}(m \times p) \to \mathrm{Poly}_{\mathbf{R}}(n), \quad \chi_S(K) = \varphi_K, \tag{1.4}$$

where $\mathrm{Mat}_{\mathbf{R}}(m \times p)$ is the set of all real matrices of size $m \times p$, $\mathrm{Poly}_{\mathbf{R}}(n)$ is the set of all real monic polynomials of degree n, and the polynomial φ_K is defined in (1.3). Thus to say that, for a system S, an arbitrary symmetric set of poles can be assigned by a real gain matrix is the same as saying that the real pole placement map χ_S is surjective. Extending the domain to complex matrices K and the range to complex monic polynomials gives the *complex pole placement map*

$$\mathrm{Mat}_{\mathbf{C}}(m \times p) \to \mathrm{Poly}_{\mathbf{C}}(n),$$

defined by the same formula as the real one.

It is easy to see that, for every m, n, p, there are systems for which the pole placement map is not surjective. For example, one can take $B = 0$ or $C = 0$. A necessary condition of surjectivity proved in [13] is that S is observable and controllable. This is equivalent to saying that the McMillan degree of the transfer function is equal to n, the dimension of the state space. Notice that this property is *generic:* it holds for an open dense subset of the set

$$\mathfrak{A} = \mathrm{Mat}_{\mathbf{R}}(n \times n) \times \mathrm{Mat}_{\mathbf{R}}(n \times m) \times \mathrm{Mat}_{\mathbf{R}}(p \times n)$$

of all triples (A, B, C). All topological terms in this paper refer to the usual topology.

In this paper, we consider the following problem: *for a given triple of integers (m, n, p), does there exist an open dense subset $V \subset \mathfrak{A}$ such that the real pole placement map χ_S is surjective for $S \in V$?* If this is the case, we say that the real pole placement map is *generically surjective* for these m, n, and p.

We briefly recall the history of the problem, referring to a comprehensive survey [2]. The pole placement map defined by (1.3) and (1.4) is a regular map of affine algebraic varieties. Comparing the dimensions of its domain and range, we conclude that $n \leq mp$ is a necessary condition for generic surjectivity of the pole placement map, real or complex. In the complex case, this condition is also sufficient [7]. To show this, one extends the pole placement map to a regular map between compact algebraic manifolds and verifies that its Jacobi matrix is of full rank. In the case when $n = mp$, we have the following precise result.

THEOREM A (see [1]). *For $n = mp$, the complex pole placement map is generically surjective. Moreover, it extends to a finite regular map between projective varieties and has degree*

$$d(m,p) = \frac{1!2!\dots(p-1)!\,(mp)!}{m!(m+1)!\dots(m+p-1)!}.$$

It follows that, for a generic system (A,B,C) with $n = mp$ and a generic monic complex polynomial φ of degree mp, there are $d(m,p)$ complex matrices K such that $\varphi_K = \varphi$.

The numbers $d(m,p)$ occur as the solution of the following problem of enumerative geometry: how many m-subspaces intersect mp given p-subspaces in $\mathbf{C}^{m+p}$ in a general position? The answer $d(m,p)$ was obtained by Schubert in 1886 (see, for example, [9]).

The real pole placement map is harder to study. For a survey of early results, we refer to [2, 12]. Wang [16] proved that $n < mp$ is sufficient for generic surjectivity of a real (or complex) pole placement map. A simplified proof of this result can be found in [17, 12].

From now on, we discuss only the so-called critical case; that is, we assume that

$$n = mp$$

in the rest of the paper. In addition, we may assume, without loss of generality, that $p \le m$, in view of the symmetry of our problem with respect to the interchange of m and p (see, for example, [15, Theorem 3.3]).

One corollary from Theorem A is that the real pole placement map is generically surjective if $d(m,p)$ is odd. This number is odd if and only if one of the following conditions is satisfied [2]: (a) $\min\{m,p\} = 1$ or (b) $\min\{m,p\} = 2$, and $\max\{m,p\}+1$ is an integral power of 2.

In the opposite direction, Willems and Hesselink [18] found by explicit computation that the real pole placement map is not generically surjective for $(m,p) = (2,2)$. A closely related fact, that the problem of enumerative geometry mentioned above may have no real solutions for the case $(m,p) = (2,2)$, even when the given 2-subspaces are real, is mentioned in [8].

In [13], Rosenthal and Sottile found with a rigorous computer-assisted proof that the real pole placement map is not generically surjective in the case $(m,p) = (4,2)$, thus disproving a conjecture of Kim that $(2,2)$ is the only exceptional case.

In [6], we showed that the real pole placement map is not generically surjective when $p = 2$ and m is even, thus extending the negative results for the cases $(2,2)$ and $(4,2)$ stated above.

In the present paper, we extend this result to all cases when both m and p are even.

THEOREM 1.1. *If $n = mp$ and m and p are both even, then the real pole placement map is not generically surjective.*

Our proof of Theorem 1.1 explicitly gives a system $S_0 \in \mathfrak{A}$ and a polynomial element $u(s) = s(s^2+1)^{mp/2-1}$ such that, for any S' in a neighborhood of S_0, the real pole placement map $\chi_{S'}$ omits a neighborhood of u.

Our proofs in [6] depend on a hard analytic result from [5], related to the so-called B. and M. Shapiro conjecture, which is stated below in section 2. The proofs in the present paper are new, even in the case $\min\{m,p\} = 2$, and they are elementary.

We conclude the introduction with an unsolved problem.

A system S is called stabilizable (by real static output feedback) if there exists a gain matrix $K \in \mathrm{Mat}_{\mathbf{R}}(m \times p)$ such that all zeros of the closed loop characteristic polynomial φ_K belong to the left half-plane. From the positive results on pole placement stated above, it follows that generic systems with m inputs, p outputs, and state of dimension n are stabilizable if $n < mp$ or if $n = mp$ and $m + p$ is odd. We ask whether generic systems with $n = mp$ and even m and p are stabilizable. The answer is known to be negative in the case $(m, p) = (2, 2)$ [3]. For complex output feedback, with static or dynamic compensators, the problem of generic stabilizability was solved in [10].

2. A class of linear systems. We begin with a well-known transformation of the closed loop characteristic polynomial (1.3). The open loop transfer function of a system of McMillan degree n, equal to the dimension of the state space, can be factorized as

$$C(sI - A)^{-1}B = D(s)^{-1}N(s), \quad \det D(s) = \det(sI - A), \tag{2.1}$$

where D and N are polynomial matrix-functions of sizes $p \times p$ and $p \times m$, respectively. For the possibility of such factorization for systems (1.1) of McMillan degree n, we refer to [4, Assertion 22.6]. Using (2.1) and the identity $\det(I - PQ) = \det(I - QP)$, which is true for all rectangular matrices of appropriate dimensions, we write

$$\begin{aligned}\varphi_K(s) &= \det(sI - A - BKC) = \det(sI - A)\det(I - (sI - A)^{-1}BKC)\\ &= \det(sI - A)\det(I - C(sI - A)^{-1}BK)\\ &= \det D(s)\det(I - D(s)^{-1}N(s)K) = \det(D(s) - N(s)K).\end{aligned}$$

This can be rewritten as

$$\varphi_K(s) = \det\left([D(s), N(s)]\begin{bmatrix} I \\ -K \end{bmatrix}\right). \tag{2.2}$$

Now we extend $\chi_S : K \mapsto \varphi_K$ to a map between compact manifolds. For this purpose, we allow an arbitrary $(m + p) \times p$ complex matrix L of rank p in (2.2) instead of

$$\begin{bmatrix} I \\ -K \end{bmatrix}, \tag{2.3}$$

and we define

$$\varphi_L(s) = \det\left([D(s), N(s)]\ L\right). \tag{2.4}$$

A system S represented by $[D(s), N(s)]$ is called *nondegenerate* if $\varphi_L \not\equiv 0$ for every $(m + p) \times p$ matrix L of rank p. Such matrices are called equivalent; $L_1 \sim L_2$ if $L_1 = L_2U$, where $U \in GL_p(\mathbf{C})$. The set of equivalence classes is the Grassmannian $G_{\mathbf{C}}(p, m + p)$, which is a compact algebraic manifold of dimension mp. If $L_1 \sim L_2$, we have $\varphi_{L_1} = c\varphi_{L_2}$, where $c \neq 0$ is a constant. The space of all nonzero polynomials of degree at most mp, modulo proportionality, is identified with the projective space $\mathbf{CP}^{mp}$, coefficients of the polynomials serving as homogeneous coordinates. Monic polynomials represent the points of an open dense subset of $\mathbf{CP}^{mp}$, a so-called *big cell,* which consists of polynomials of degree mp. This construction extends the complex pole placement map of a nondegenerate system to a regular map of compact algebraic manifolds

$$\chi_S : G_{\mathbf{C}}(p, m + p) \to \mathbf{CP}^{mp}, \tag{2.5}$$

where $\chi_S(L)$ is the proportionality class of the polynomial φ_L in (2.4), and L is a matrix of rank p representing a point in $G_{\mathbf{C}}(p, m+p)$. The set $\mathfrak{B}$ of all nondegenerate systems is open and dense in the set $\mathfrak{A}$ of all systems, and the map

$$X \times G_{\mathbf{C}}(p, m+p) \to \mathbf{CP}^{mp}, \quad (S, L) \mapsto \chi_S(L) \tag{2.6}$$

is continuous. Notice that the subset of $G_{\mathbf{R}}(p, m+p)$ consisting of points which can be represented by matrices L of the form (2.3) is open and dense. It corresponds via χ_S to the big cell in $\mathbf{CP}^{mp}$ consisting of polynomials of degree mp.

We consider a system $S_0 = (A_0, B_0, C_0)$ represented by the polynomial matrix $[D(s), N(s)]$

$$= \begin{bmatrix} 1 & s & \dots & s^{m+p-2} & s^{m+p-1} \\ 0 & 1 & \dots & (m+p-2)s^{m+p-3} & (m+p-1)s^{m+p-2} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \dots & (m+1)\dots(m+p-1)s^m \end{bmatrix}. \tag{2.7}$$

The first row of $[D(s), N(s)]$ consists of monic monomials, and the kth row is the $(k-1)$st derivative of the first for $2 \le k \le p$. This system S_0 has McMillan degree mp, and the matrices A_0, B_0, C_0 can be recovered from $[D, N]$ by [4, Theorem 22.18]. Let $L = (a_{i,j})$. Introducing polynomials

$$f_j(s) = a_{1,j} + a_{2,j}s + \dots + a_{m+p-1,j}s^{m+p-2} + a_{m+p,j}s^{m+p-1} \tag{2.8}$$

for $1 \le j \le p$, we can write (2.4) as

$$\varphi_L = W(f_1, \dots, f_p) = \begin{vmatrix} f_1 & \dots & f_p \\ f_1' & \dots & f_p' \\ \dots & \dots & \dots \\ f_1^{(p-1)} & \dots & f_p^{(p-1)} \end{vmatrix}.$$

Thus, for our system (A_0, B_0, C_0), the pole placement map becomes the *Wronski map,* which sends a p-vector of polynomials into their Wronski determinant. We say that two p-vectors of polynomials are equivalent, $(f_1, \dots, f_p) \sim (g_1, \dots, g_p)$, if $(g_1, \dots, g_p) = (f_1, \dots, f_p)U$, where $U \in GL_p(\mathbf{C})$. Equivalent p-vectors have proportional Wronski determinants. Equivalence classes of p-vectors of linearly independent polynomials of degree at most $m+p-1$ parametrize the Grassmannian $G_{\mathbf{C}}(p, m+p)$. A p-vector of complex polynomials will be called *real* if it is equivalent to a p-vector of real polynomials. The system represented by (2.7) is nondegenerate. This is a consequence of the well-known fact that the Wronski determinant of p polynomials is zero if and only if the polynomials are linearly dependent.

To prove Theorem 1.1, we use the following general result (compare [13, Theorem 3.1]).

PROPOSITION 2.1. *If, for some (m, n, p), there exists a real nondegenerate system $S_0 = (A_0, B_0, C_0)$ such that the real pole placement map χ_{S_0} in (2.5) is not surjective, then, for these (m, n, p), the real pole placement map is not generically surjective.*

Indeed, if χ_{S_0} omits one point u, it omits a neighborhood of u, because the image of a compact space under a continuous map is compact. Using continuity of the map (2.6), we conclude that, for all S in a neighborhood of S_0, the maps χ_S omit a neighborhood of u. □

In view of Proposition 2.1, to prove Theorem 1.1, it is enough to find a nonzero real polynomial of degree at most mp which cannot be represented as the Wronski

determinant of p real polynomials of degree at most $m+p-1$. Thus Theorem 1.1 follows from Proposition 2.1 and the following proposition.

PROPOSITION 2.2. *If $m \geq p \geq 2$ are even integers, then the polynomial $u(s) = s(s^2+1)^{mp/2-1}$ is not proportional to the Wronski determinant of any p real polynomials of degree at most $m+p-1$.*

Proposition 2.2 is motivated by a conjecture of B. and M. Shapiro (see, for example, [15]), which says: *If the Wronskian determinant of a polynomial p-vector has only real roots, then this p-vector is real.* In [5], we proved this conjecture for $p=2$ and used this result in [6] to derive the case $p=2$ of Theorem 1.1. In the present paper, we prove a result, Proposition 3.1 in section 3, which is a very special case of the B. and M. Shapiro conjecture, but it still permits us to derive Proposition 2.2.

3. The Wronski map. A p-vector of linearly independent polynomials of degree at most $m+p-1$ can be represented by an $(m+p)\times p$ matrix L of rank p, whose columns are composed of the coefficients of the polynomials as in (2.8).

The group $GL_p(\mathbf{C})$ acts on such matrices by multiplication from the right. This action is equivalent to the usual column operations on matrices: interchanging two columns, multiplying a column by a nonzero constant, and adding to a column a multiple of another column. For each column j of L, we introduce two integers $1 \leq e_j \leq d_j \leq m+p$, which are the positions of the first and last nonzero elements of this column, counted from above. Thus $\deg f_j = d_j - 1$, and the order of a root of f_j at zero is $e_j - 1$. It is easy to see that, by column operations, every $(m+p)\times p$ matrix $L = (a_{i,j})$ of rank p can be reduced to the following *canonical form:*

(i) $d_1 > d_2 > \cdots > d_p$,

(ii) $a_{e_j,j} = 1$ for every $j \in [1,p]$,

(iii) $a_{e_k,j} = 0$ for $1 \leq j < k \leq p$.

The elements $a_{e_j,j} = 1$, $1 \leq j \leq p$, of the canonical form will be called the *pivot* elements. It follows from (iii) that all numbers e_j are distinct.

PROPOSITION 3.1. *Suppose that mp is even. Then every polynomial p-vector $(f_1,\dots,f_p)$ of degree at most $m+p-1$ in canonical form, which satisfies*

$$W(f_1,\dots,f_p) = \lambda w, \quad \text{where} \quad w(s) = s^{mp/2+1} - s^{mp/2-1}, \quad \lambda \in \mathbf{C}^*, \tag{3.1}$$

has only real entries.

COROLLARY. *All polynomial p-vectors of degree at most $m+p-1$ satisfying* (3.1) *are real.*

This corollary confirms a special case of the B. and M. Shapiro conjecture, when the Wronskian determinant of a polynomial p-vector is $w(s) = s^{mp/2+1} - s^{mp/2-1}$, which is a polynomial with real roots $0, \pm 1$.

The properties of the Wronskian determinants used here are well known and easy to prove.

LEMMA. *The Wronski map $(f_1,\dots,f_p) \mapsto W(f_1,\dots,f_p)$ is linear with respect to each f_j, and*

$$W(s^{n_1},\dots,s^{n_p}) = V(n_1,\dots,n_p)s^{n_1+\cdots+n_p-p(p-1)/2},$$

where

$$V(n_1,\dots,n_p) = \prod_{k<j}(n_j - n_k)$$

is the Vandermonde determinant. □

Using this lemma, we compute the Wronskian determinant of a polynomial p-vector in canonical form and conclude that

$$\deg W(f_1, \dots, f_p) = d_1 + \cdots + d_p - p(p+1)/2 \tag{3.2}$$

and

$$\operatorname{ord} W(f_1, \dots, f_p) = e_1 + \cdots + e_p - p(p+1)/2, \tag{3.3}$$

where ord denotes the multiplicity of a root at zero.

Proof of Proposition 3.1. According to (3.1), $\deg w = mp/2 + 1$, and $\operatorname{ord} w = mp/2 - 1$. So (3.2) and (3.3) imply

$$\begin{aligned} d_1 + \cdots + d_p &= p(p+1)/2 + mp/2 + 1, \\ e_1 + \cdots + e_p &= p(p+1)/2 + mp/2 - 1. \end{aligned}$$

Subtracting the second equation from the first, we get

$$\sum_{j=1}^{p} (d_j - e_j) = 2.$$

As all the summands are nonnegative, there are two possibilities.

Case 1. In all columns but one, all elements, except the pivot elements, are equal to zero, and, for the exceptional column j, $d_j - e_j = 2$. Computing the Wronskian and comparing it with (3.1), we obtain

$$\begin{aligned} & V(\dots, e_j - 1, \dots) s^{mp/2-1} \\ + \quad & V(\dots, e_j, \dots) a_{e_j+1,j} s^{mp/2} \\ + \quad & V(\dots, e_j + 1, \dots) a_{e_j+2,j} s^{mp/2+1} \\ = \quad & -\lambda s^{mp/2-1} + \lambda s^{mp/2+1}. \end{aligned}$$

Here and in what follows, the notation $V(\dots, e_j + m, \dots)$ means the Vandermonde determinant of p arguments, whose kth argument is $e_k - 1$ for $k \neq j$ and whose jth argument is $e_j + m$.

Comparing the terms with $s^{mp/2-1}$, we conclude that λ is real. Comparing the terms with $s^{mp/2+1}$, we conclude that $V(\dots, e_j + 1, \dots) \neq 0$, and thus $a_{e_j+2,j}$ is real. Now we consider the middle term in the expansion of the Wronskian determinant. If $V(\dots, e_j, \dots) = 0$, then $e_k = e_j + 1$ for some k. As $d_k = e_k$ and $d_j = e_j + 2$, we conclude that $d_k = d_j - 1$, so $k > j$ by (i) in the definition of the canonical form. Now (iii) from the definition of the canonical form implies that $a_{e_j+1,j} = 0$. If $V(\dots, e_j, \dots) \neq 0$, we also conclude that $a_{e_j+1,j} = 0$. Thus all entries of L are real.

Case 2. In all columns but two, all nonpivot elements are equal to zero, and the two exceptional columns contain one extra nonzero element each. Let $j < k$ be the positions of the exceptional columns, and let $a = a_{e_j+1,j}$ and $b = a_{e_k+1,k}$ be the nonzero, nonpivot elements of these columns. Computing the Wronskian and comparing it with (3.1), we obtain

$$\begin{aligned} & V(\dots, e_j - 1, \dots) s^{mp/2-1} \\ + \quad & \left(aV(\dots, e_j, \dots) + bV(\dots, e_k, \dots)\right) s^{mp/2} \\ + \quad & abV(\dots, e_j, \dots, e_k, \dots) s^{mp/2+1} \\ = \quad & -\lambda s^{mp/2-1} + \lambda s^{mp/2+1}, \end{aligned} \tag{3.4}$$

where $V(\dots, e_j, \dots, e_k \dots)$ denotes the Vandermonde determinant of p arguments, whose jth argument is e_j and whose kth argument is e_k, and, for all other indices $l \notin \{j, k\}$, the lth argument is $e_l - 1$.

Our first conclusions are

$$V(\dots, e_j - 1, \dots) = -\lambda \tag{3.5}$$

and

$$V(\dots, e_j, \dots, e_k, \dots) \neq 0. \tag{3.6}$$

It follows from (3.5) that λ is real. If exactly one of the numbers $V(\dots, e_j, \dots)$ and $V(\dots, e_k, \dots)$ is zero, then (3.4) implies that at least one of the numbers a or b is zero. Then the third term in the expansion of the Wronskian is zero, which contradicts (3.4). If both $V(\dots, e_j, \dots)$ and $V(\dots, e_k, \dots)$ are zero, then $V(\dots, e_j, \dots, e_k, \dots) = 0$, and this contradicts (3.6). So both $V(\dots, e_j, \dots)$ and $V(\dots, e_k, \dots)$ are nonzero. This means that there are no pivot elements in the rows $e_j + 1$ and $e_k + 1$. Using (3.6), we conclude that $V(\dots, e_j - 1, \dots)$, $V(\dots, e_j, \dots)$, $V(\dots, e_k, \dots)$, and $V(\dots, e_j, \dots, e_k, \dots)$ have the same sign, and, by (3.5), all these numbers have the sign of $-\lambda$. As $V(\dots, e_j, \dots)$ and $V(\dots, e_k, \dots)$ are of the same sign, (3.4) implies that $a = -cb$, where $c > 0$, and from the equations

$$V(\dots, e_j, \dots, e_k, \dots)ab = \lambda$$

and (3.5) we conclude that a and b are real. □

The group $\mathrm{Aut}(\mathbf{CP}^1)$ of fractional-linear transformations acts on the space $\mathbf{CP}^k$ of proportionality classes of nonzero polynomials of degree at most k by the following rule: Let

$$\ell(s) = \frac{as + b}{cs + d}, \quad ad - bc \neq 0,$$

represent a fractional-linear transformation. For a polynomial $r(s)$, we put

$$\ell r(s) = (-cs + a)^k r \circ \ell^{-1}(s).$$

That this is indeed a group action can be verified as follows. The space of proportionality classes of nonzero polynomials of degree at most k can be canonically identified with the symmetric power $\mathrm{Sym}^k(\mathbf{CP}^1)$, which is the set of unordered k-tuples of points in $\mathbf{CP}^1$. To each polynomial r, one puts into correspondence its roots, counted with multiplicity, and the point ∞ with multiplicity $k - \deg r$. Then the action of $\ell \in \mathrm{Aut}(\mathbf{CP}^1)$ on such a k-tuple is simply

$$(s_1, \dots, s_k) \mapsto (\ell(s_1), \dots, \ell(s_k)).$$

It is easy to verify that this action of $\mathrm{Aut}(\mathbf{CP}^1)$ extends to the space $G_{\mathbf{C}}(p, m+p)$ of equivalence classes of polynomial p-vectors of degree at most $m+p-1$. Furthermore, this extended action is respected by the Wronski map:

$$W(\ell g_1, \dots, \ell g_p) = \ell W(g_1, \dots, g_p). \tag{3.7}$$

Of course, in the left-hand side of this equality, the group $\mathrm{Aut}(\mathbf{CP}^1)$ acts on $\mathrm{Sym}^{m+p-1}(\mathbf{CP}^1)$, while, in the right-hand side, it acts on $\mathrm{Sym}^{mp}(\mathbf{CP}^1)$. Equation (3.7) permits us to simplify the polynomial equation

$$W(g_1, \dots, g_p) = v, \quad v(s) \sim s(s^2 - 1)^{mp/2-1}, \tag{3.8}$$

which will be used to prove Proposition 2.2.

Consider the fractional-linear transformation

$$\ell(s) = \ell^{-1}(s) = \frac{1-s}{1+s}. \tag{3.9}$$

We have $\ell : (0, 1, \infty, -1) \mapsto (1, 0, -1, \infty)$, and $\ell(\overline{\mathbf{R}}) = \overline{\mathbf{R}}$.

Using (3.8) and (3.9), we obtain

$$\ell v(s) = (s+1)^{mp} v \circ \ell^{-1}(s) \sim s^{mp/2+1} - s^{mp/2-1} = w(s),$$

where "$\sim$" means "proportional." Thus, with $f_j = \ell g_j$, (3.8) is equivalent to the equation

$$W(f_1, \dots, f_p) = w, \quad w(s) \sim s^{mp/2+1} - s^{mp/2-1}, \tag{3.10}$$

which we solved in Proposition 3.1. The conclusion is that

(3.11) all solutions of (3.8) in canonical form have real coefficients.

Proof of Proposition 2.2. Suppose that $(f_1, \dots, f_p)$ is a real polynomial p-vector in canonical form satisfying

$$W(f_1, \dots, f_p) = u, \quad u(s) = \lambda s(s^2+1)^{mp/2-1}, \quad \lambda \neq 0. \tag{3.12}$$

Then (3.3) implies

$$e_1 + \cdots + e_p = 1 + p(p+1)/2.$$

As $(e_j)_{j=1}^p$ are distinct positive integers, the only possibility is that

$$\{e_1, \dots, e_p\} = \{1, 2 \dots, p-1, p+1\}. \tag{3.13}$$

Similarly, (3.2) implies

$$d_1 + \cdots + d_p = mp + p(p+1)/2 - 1.$$

As $(d_j)_{j=1}^p$ are distinct integers in the interval $[1, m+p]$, the only possibility is that

$$\{d_1, \dots, d_p\} = \{m, m+2, m+3, \dots, m+p\}. \tag{3.14}$$

Notice that the sequence (3.13) contains $p/2+1$ odd numbers and $p/2-1$ even numbers. On the other hand, the sequence (3.14) contains $p/2-1$ odd numbers and $p/2+1$ even numbers. This implies that, at least for one j,

$$d_j - e_j \quad \text{is odd.} \tag{3.15}$$

This means that the polynomial f_j contains both even and odd powers of s with nonzero coefficients. So the polynomial $g_j(s) = f_j(is)$, $i = \sqrt{-1}$, is not proportional to any polynomial with real coefficients. On the other hand, the polynomial p-tuple $(g_1, \dots, g_p)$, where $g_j(s) = \epsilon_j f_j(is)$ with appropriate $\epsilon_j \in \{\pm 1, \pm i\}$, is a solution of (3.8) in canonical form, and we know from (3.11) that all such solutions have real coefficients. This contradiction completes the proof of Proposition 2.2. ∎

Acknowledgment. We thank the referees of this paper for their helpful comments.

REFERENCES

[1] R. Brockett and C. Byrnes, *Multivariable Nyquist criteria, root loci and pole placement: A geometric viewpoint*, IEEE Trans. Automat. Control, 26 (1981), pp. 271–284.

[2] C. Byrnes, *Pole assignment by output feedback*, in Three Decades of Mathematical System Theory, Lecture Notes in Control and Inform. Sci. 135, H. Nijmeijer and J. M. Schumacher, eds., Springer-Verlag, New York, 1989, pp. 31–78.

[3] C. I. Byrnes and B. D. O. Anderson, *Output feedback and generic stabilizability*, SIAM J. Control Optim., 22 (1984), pp. 362–380.

[4] D. Delchamps, *State Space and Input-Output Linear Systems*, Springer-Verlag, New York, 1988.

[5] A. Eremenko and A. Gabrielov, *Rational functions with real critical points and the B. and M. Shapiro conjecture in real enumerative geometry*, Ann. of Math. (2), 155 (2002).

[6] A. Eremenko and A. Gabrielov, *Counterexamples to pole placement by static output feedback*, Linear Algebra Appl., to appear.

[7] R. Hermann and C. Martin, *Applications of algebraic geometry to systems control theory*, I, IEEE Trans. Automat. Control, 22 (1977), pp. 19–25.

[8] S. Kleiman, *Problem* 15: *Rigorous foundation of Schubert's enumerative calculus*, in Mathematical Developments Arising from Hilbert Problems, F. Browder, ed., AMS, Providence, RI, 1976, pp. 445–482.

[9] S. Kleiman and D. Laksov, *Schubert calculus*, Amer. Math. Monthly, 79 (1972), pp. 1061–1082.

[10] M. S. Ravi, J. Rosenthal, and X. Wang, *On generic stabilizability and pole assignability*, Systems Control Lett., 23 (1994), pp. 79–84.

[11] M. S. Ravi, J. Rosenthal, and X. Wang, *Dynamic pole assignment and Schubert calculus*, SIAM J. Control Optim., 34 (1996), pp. 813–832.

[12] J. Rosenthal, J. Schumacher, and J. Willems, *Generic eigenvalue assignment by memoryless output feedback*, Systems Control Lett., 26 (1995), pp. 253–260.

[13] J. Rosenthal and F. Sottile, *Some remarks on real and complex output feedback*, Systems Control Lett., 33 (1998), pp. 73–80.

[14] J. Rosenthal and X. Wang, *Output feedback pole placement with dynamic compensators*, IEEE Trans. Automat. Control, 41 (1996), pp. 830–843.

[15] F. Sottile, *Real Schubert calculus: Polynomial systems and a conjecture of Shapiro and Shapiro*, Experiment. Math., 9 (2000), pp. 161–182.

[16] X. Wang, *Pole placement by static output feedback*, J. Math. Systems Estim. Control, 2 (1992), pp. 205–218.

[17] X. Wang, *Grassmannian, central projection and output feedback pole assignment of linear systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 786–794.

[18] J. Willems and W. Hesselink, *Generic properties of the pole placement problem*, in Proceedings of the 7th IFAC Congress, IFAC, New York, 1978, pp. 1725–1728.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 1, pp. 313–330

FILTERING FOR LINEAR SYSTEMS DRIVEN BY FRACTIONAL BROWNIAN MOTION*

N. U. AHMED† AND C. D. CHARALAMBOUS‡

Abstract. In this paper we study continuous time filtering for linear multidimensional systems driven by fractional Brownian motion processes. We present the derivation of the optimum linear filter equations which involve a pair of functional-differential equations giving the optimum error covariance (matrix-valued) functions and the optimum filter. These equations are the appropriate substitutes of the matrix-Riccati differential equation arising in classical Kalman filtering. However, the optimum filter has the classical appearance, and, as usual, it is driven by the increments of the observed process.

Key words. linear filtering, fractional Brownian motion

AMS subject classifications. 49J55, 60G35, 60H10, 93E11

PII. S0363012900368715

1. Introduction.

In recent years, there has been renewed interest in fractional Brownian motion, originally introduced by Mandelbrot and Van Ness [1], to model phenomena that exhibit so-called self-similarity, which is a form of time invariance of the fundamental structure of the process (fractal), and long range dependence. The long range dependence, which is absent in regular Brownian motion or, more precisely, its (distributional) derivative, the white noise, is historically observed in the study of water accumulation in hydrology [1], ethernet and ATM traffic in telecommunication systems [7, 8], and stock prices in mathematical finance.

A natural question which arises regards the filtering of stochastic dynamical systems driven by fractional Brownian motion, when the measurements are linear in the unobservable variable in additive fractional Brownian motion. Many problems in engineering and science are of this nature, and therefore it is natural to generalize the linear filtering theory to cover the case in which the standard Brownian motion is replaced by fractional Brownian motion. However, unlike the classical Kalman filtering problem, the optimal filter for linear systems perturbed by fractional Brownian motion is not straightforward. This is easily verified by consulting the results described in [14, 11, 13, 16], where the authors treat various versions of filtering and prediction for fractional processes. In particular, it is easily seen in [14] that even the linear prediction problem of fractional Brownian motion based on its past history leads to complicated expressions, and similarly for the rest of these references, which treat the scalar case.

In this paper, we derive the best linear filter for the multidimensional case, when the state is described by a linear stochastic differential equation driven by fractional Brownian motion and the observations are linear in the state process and subject to additive fractional Brownian motion noise. We show that the methodology of variational methods, extensively discussed in [4, 5, 6] for deriving the best linear filter

*Received by the editors February 19, 2000; accepted for publication (in revised form) December 17, 2001; published electronically June 5, 2002.
http://www.siam.org/journals/sicon/41-1/36871.html
†School of Information Technology and Department of Mathematics, University of Ottawa, Ottawa, ON, Canada (ahmed@site.uottawa.ca).
‡School of Information Technology, University of Ottawa, Ottawa, ON, Canada (chadcha@site.uottawa.ca).

when the noises are standard Brownian motions, also apply to the current set-up, although the resulting equations are more complicated. In particular, the gain of the filter is shown to satisfy two coupled functional integro-differential matrix equations. In the limit as the fractional Brownian motion in the dynamics or observations, which have different Hurst parameters, converges to the standard Brownian motion (through the convergence of either of the Hurst parameters to 1/2), we also obtain special forms of the coupled functional equations, which include the well-known Riccati equation of the error covariance of the Kalman filter, when both Hurst parameters converge to 1/2. Thus, our results give as a special case the solution to the linear filtering problem when one of the noises is a fractional Brownian motion while the other is a standard Brownian motion. We also propose a numerical technique for computing the solution of the coupled functional equations, which merits further investigation.

The key properties of fractional Brownian motion are by now well known and documented [9, 10, 11, 12, 13, 14, 15, 16, 17; 2, 3]. We shall attempt to summarize only those results which are important in subsequent developments, while referring to these references for the proofs.

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $H \in (0,1)$. A parameterized family of random process $\{B_{H_1}(t), t \geq 0\}$ based on this probability space is said to be a fractional Brownian motion if

(i) $P\{B_H(0) = 0\} = 1$;

(ii) for each $t \in R_+ \equiv [0, \infty)$, $B_H(t)$ is an $\mathcal{F}$-measurable random variable having Gaussian distribution with $E\{B_H(t)\} = 0$;

(iii) for $t, s \in R_+$, $E\{B_H(t)B_H(s)\} = (1/2)\{t^{2H} + s^{2H} - |t-s|^{2H}\}$.

It follows from (iii) and the well-known Kolmogorov's criterion for continuity that, for $H > (1/2)$,

(iv) the sample paths of B_H are continuous with probability one but nowhere differentiable.

Further, it follows from (iii) again that the variance of $B_H(t)$ is t^{2H}, and for $H = 1/2$, $E\{B_{1/2}(t)B_{1/2}(s)\} = t \wedge s$. That is, $B_{1/2}$ is the standard Brownian motion.

In fact, fractional Brownian motion can be constructed from classical Brownian motion by a linear transformation of the form

$$B_H(t) \equiv \int_0^t K_H(t,s)dB(s), \tag{1.1}$$

where the process $\{B(t), t \geq 0\}$ is the classical Brownian motion and K_H is a kernel dependent on the parameter H, known as the Hurst parameter. Assuming (i), one may choose (see [10])

$$K_H(t,s) = \frac{(t-s)^{H-\frac{1}{2}}}{\Gamma(H+\frac{1}{2})} F\left(\frac{1}{2} - H, H - \frac{1}{2}, H + \frac{1}{2}, 1 - \frac{t}{s}\right) \mathbf{1}_{(0,t)}(s) \tag{1.2}$$

(F is the hypergeometric function), which implies that for $H = (1/2)$, $B_{(1/2)}(t) = B(t)$ is the standard Brownian motion.

It follows from this construction that

(v) B_H is self-similar in the sense that, for any $\alpha > 0$, the probability laws of $\{B_H(\alpha t)\}$ and $\{\alpha^H B_H(t)\}$ coincide.

For other choices and more general fractional Brownian motions and their properties, see Mandlebrot [1], as well as [9, 10, 11, 12, 13, 14, 15] and [2, 3]. It is reported in these papers that random processes arising from hydrological and economic time

series exhibit long range interdependence and self-similarity. Since fractional Brownian motion does have these properties, it is reasonable to use fractional Brownian motion to model such processes.

A function that plays an important role in the construction of stochastic integrals based on fractional Brownian motion is given by

$$\varphi_H(t) \equiv H(2H-1)|t|^{2H-2}, \qquad t \in R. \tag{1.3}$$

It can easily be shown (see, for example, [13, 14, 3]) that, for all $t, s \in R_+$, we have

$$E\{B_H(t)B_H(s)\} = \int_0^t \int_0^s \varphi_H(\tau - \theta)d\theta d\tau. \tag{1.4}$$

One can introduce (see [12, 14, 3]) the class of functions $L^2_\varphi(R_+)$ which consist of all Borel measurable real-valued functions $\{f\}$ defined on R_+ satisfying

$$\| f \|^2_\varphi \equiv \int_0^\infty \int_0^\infty \varphi_H(t-s)f(s)f(t)dsdt < \infty. \tag{1.5}$$

With respect to the scalar product

$$(f,g)_\varphi \equiv \int_0^\infty \int_0^\infty \varphi_H(t-s)f(s)g(t)dsdt, \tag{1.6}$$

$L^2_\varphi(R_+)$ is a Hilbert space. Stochastic integrals, with respect to the fractional Brownian motion B_H, of deterministic integrands from the class L^2_φ are well defined. More precisely, for each $f \in L^2_\varphi$, the element X given by

$$X \equiv \int_0^\infty f(t)dB_H(t) \tag{1.7}$$

is a well-defined random variable (real-valued $\mathcal{F}$-measurable). Since f is deterministic and B_H is Gaussian, the random variable X is also Gaussian, and it is easy to see that

$$\text{(a) } E\{X\} = 0 \qquad \text{and} \qquad \text{(b) } E|X|^2 = \| f \|^2_\varphi . \tag{1.8}$$

Since we are interested in the filtering problem for multidimensional processes, we modify the preceding results to suit this requirement. Again we use $(\Omega, \mathcal{F}, P)$ to denote the basic probability space on which all the random processes to be defined below are supported. For any integer n one may construct the fractional Brownian by the following expression:

$$B_H(t) \equiv \int_0^t K_H(t,\theta)dB(\theta), \tag{1.9}$$

where B is an n dimensional Brownian motion with covariance, say $Q \in M_s^+(n \times n)$. Here $M_s^+(n \times n)$ denotes the class of real symmetric positive definite matrices, and K_H is the scalar kernel as introduced above. In view of the previous results, B_H is an R^n-valued Gaussian random process having mean and covariance given by (see, for example, [3])

(B1) $E\{B_H(t)\} = 0$,

(B2) $E\{(B_H(t),\xi)(B_H(s),\eta)\} = \int_0^t \int_0^s \varphi_H(\tau - \theta)(Q\xi,\eta)d\tau d\theta$ for all $\xi, \eta \in R^n$.

Clearly it follows from (B2) that

$$E\{(B_H(t),\xi)^2\} = t^{2H}(Q\xi,\xi), \qquad \xi \in R^n, t \in R_+. \tag{1.10}$$

Now we can define stochastic Wiener integrals with respect to the FBM (fractional Brownian motion). For simplicity we consider finite intervals $I \equiv [0,T]$, $T < \infty$. Let $M(k \times n)$ denote the vector space of $k \times n$ matrices with real entries. For any $H \in (0,1)$, let $L^2_H(I, M(k \times n))$ denote the Hilbert space with the scalar product defined by

$$(\sigma,\beta)_H \equiv \int_I \int_I \varphi_H(u-v) Tr\{\sigma(u) Q \beta^{'}(v)\} du dv \qquad \text{for } \sigma,\beta \in L^2_H(I, M(k\times n)) \tag{1.11}$$

and the norm by

$$\| \sigma \|_H \equiv \left(\int_I \int_I \varphi_H(u-v) Tr\{\sigma(u) Q \sigma^{'}(v)\} du dv \right)^{1/2} \qquad \text{for } \sigma \in L^2_H(I, M(k\times n)). \tag{1.12}$$

Clearly this Hilbert space is related to the FBM B_H.

For $\sigma \in L^2_H(I, M(k \times n))$, define

$$Z \equiv \int_I \sigma(t) dB_H(t) \equiv L(\sigma). \tag{1.13}$$

This is a well-defined random variable with values in R^k. The following result is useful in what follows.

LEMMA 1.1. *For each $\sigma \in L^2_H(I, M(k \times n))$, the element Z given by* (1.13) *is a well-defined Gaussian random variable with values in R^k satisfying the following properties:*

(p1) $EZ = 0$;

(p2) $E(Z,\xi)^2 = \int_I \int_I \varphi_H(u-v)(Q\sigma^{'}(u)\xi, \sigma^{'}(v)\xi) du dv$ *for each* $\xi \in R^k$;

(p3) $E\{\| Z \|^2\} = \int_I \int_I \varphi_H(u-v) Tr\{\sigma(u) Q \sigma^{'}(v)\} du dv$.

Further,

(p4) *for* $H \geq 1/2$, $L^2(I, M(k \times n)) \hookrightarrow L^2_H(I, M(k \times n))$.

Proof. The first statement is obvious. It follows from the fact that a linear transformation of a Gaussian random process is Gaussian. Since σ is deterministic, (p1) follows from the property of B_H given by (B1), and (p2) follows from the property (B2). Let $\{e_i, i = 1,2,\dots,k\}$ be any basis of the space R^k. Replacing ξ by e_i in (p2) and summing over the indices, one obtains (p3). For (p4), let $H \geq 1/2$ and take any $\sigma \in L^2(I, M(k \times n))$. We must verify that it belongs to $L^2_H(I, M(k \times n))$. By definition (1.12),

$$\| \sigma \|^2_H = \int_I \int_I \varphi_H(u-v) Tr\{\sigma(u) Q \sigma^{'}(v)\} du dv.$$

Since φ_H is symmetric positive and Q is a positive symmetric matrix, we have

$$\begin{aligned} \| \sigma \|^2_H &\leq Tr(Q) \int_I \int_I \varphi_H(u-v) \| \sigma(u) \| \| \sigma(v) \| du dv \\ &\leq Tr(Q) \sqrt{\int_{I\times I} \varphi_H(u-v) \| \sigma(u) \|^2 du dv} \sqrt{\int_{I\times I} \varphi_H(u-v) \| \sigma(v) \|^2 du dv} \\ &\leq Tr(Q) \int_I \| \sigma(u) \|^2 \left(\int_I \varphi_H(u-v) dv \right) du. \end{aligned} \tag{1.14}$$

Now we note that

$$\int_I \varphi_H(u-v)dv = H\{u^{2H-1} + (T-u)^{2H-1}\}. \tag{1.15}$$

It follows from (1.14) and (1.15) that, for $H = 1/2$, we have

$$\| \sigma \|_H^2 \;\leq\; Tr(Q) \;\| \sigma \|^2_{L^2(I,M(k\times n))}, \tag{1.16}$$

and for $H > 1/2$ we have

$$\| \sigma \|_H^2 \leq 2HT^{2H-1}\; Tr(Q) \;\| \sigma \|^2_{L^2(I,M(k\times n))}. \tag{1.17}$$

This proves that for $H \geq 1/2$ the Hilbert space $L^2(I, M(k \times n))$ is continuously embedded in $L^2_H(I, M(k \times n))$, proving (p4). This completes the proof. □

It is clear from the above result that the operator L defined by (1.13) is a bounded linear transformation from $L^2_H(I, M(k \times n))$ to $L^2(\Omega, \mathcal{F}, P; R^k)$.

Remark. It is not clear to us at this time whether the following embedding is true: for $0 < H < 1/2$

$$L^2_H(I, M(k \times n)) \hookrightarrow L^2(I, M(k \times n)).$$

Remark. For $T = \infty$, the embedding (p4) does not hold.

2. System and measurement dynamics and the filtering problem. Since long term interdependence is encountered only for Hurst parameters greater than 1/2, from now on we consider only this case. In what follows we shall introduce two FBMs $\{B_{H_1}(t), V_{H_2}(t), t \geq 0\}$, with $H_1 > \frac{1}{2}, H_2 > \frac{1}{2}$, to represent the noise in the system's dynamics and observations. The system is governed by the following linear stochastic differential equation:

$$\begin{aligned} dx(t) &= A(t)x(t)dt + \sigma(t)dB_{H_1}(t), \\ x(0) &= x_0. \end{aligned} \tag{2.1}$$

Simplified versions of model (2.1) have been used in the literature to describe traffic in Internet applications, such as packetized video data.

The measurement dynamics is given by

$$\begin{aligned} dy(t) &= H(t)x(t)dt + \sigma_0(t)dV_{H_2}(t), \qquad t \geq 0, \\ y(0) &= 0. \end{aligned} \tag{2.2}$$

Often, measurements of traffic data consist of an aggregation of various sources from other users in addition to a linear combination of traffic modeled by x. Under such a scenario, one can justify the use of FBM in the measurements, in view of the experimental evidence suggesting that aggregated traffic in Internet applications shows evidence of self-similarity and long range dependence. We caution the reader that the matrix $H(t)$ appearing in (2.2) is part of the sensor model and is not to be confused with the Hurst parameter discussed in the previous section.

Throughout, we consider the processes $\{x, y\}$ taking values from R^n and R^m, respectively. The noise processes $\{B_{H_1}(t), V_{H_2}(t), t \geq 0\}$ are FBMs taking values from R^d and R^m, respectively. For compatibility, it is clear that the matrices $\{A, \sigma, H, \sigma_0\}$, which are deterministic, must take values from $M(n\times n), M(n\times d), M(m\times n), M(m\times m)$, respectively. Let $\mathcal{F}^y_t,\ t \geq 0$, be an increasing family of subsigma algebras of the sigma algebra $\mathcal{F}$ induced by the random process $\{y(t), t \geq 0\}$. In other words, this is the filtration associated with the process y. The basic filtering problem is to find a process z so that for each $t \geq 0$, $z(t)$ is $\mathcal{F}^y_t$-adapted (measurable) satisfying

(1) $E\{z(t)\} = E\{x(t)\}, t \geq 0$,
(2) $E \parallel x(t) - z(t) \parallel^2 \longrightarrow$ is minimum for $t \geq 0$.

That is, we want an unbiased minimum variance filter. This is given by

$$\hat{x}(t) \equiv E\{x(t)|\mathcal{F}_t^y\}. \tag{2.3}$$

However, this is a very difficult problem; therefore our objective here is to find the best (unbiased-minimum variance = UMV) linear filter driven by the observed process y, as described by the following stochastic differential equation:

$$\begin{aligned} dz(t) &= B(t)z(t)dt + \Gamma(t)dy(t), \qquad t \geq 0, \\ z(0) &= \hat{x}_0 \equiv Ex_0, \end{aligned} \tag{2.4}$$

where B and Γ are suitable matrix-valued functions to be determined. Clearly, by substituting the observed processes into (2.4), the FBM appears in the observations, and conditions similar to the previous discussion on Γ should be introduced for the stochastic integral to be well-defined. These conditions are made explicit in subsequent sections where the function spaces associated with B, Γ are identified. Specifically, we shall require that B be locally integrable and $\Gamma \in L^\infty(R_+, M(n \times m))$, and therefore $\Gamma\sigma_0 \in L^2_{H_2}(I, M(n \times m))$, implying that the stochastic integral is well-defined, and consistent with the definitions of the previous section.

3. Reformulation of the filtering problem as a control problem. We introduce the following basic assumptions:

(A1) There exist matrices $Q \in M_s^+(d \times d)$ and $Q_0 \in M_s^+(m \times m)$ such that

$$E\{(B_{H_1}(t), \xi)(B_{H_1}(s), \eta)\} = \int_0^t \int_0^s \varphi_{H_1}(\theta - \tau)(Q\xi, \eta)d\tau d\theta, \quad \xi, \eta \in R^d,$$
$$E\{(V_{H_2}(t), \xi)(V_{H_2}(s), \eta)\} = \int_0^t \int_0^s \varphi_{H_2}(\theta - \tau)(Q_0\xi, \eta)d\tau d\theta, \quad \xi, \eta \in R^m,$$

where $\varphi_{H_i}, i = 1, 2$, are given by (1.3).

(A2) The matrices A and H are locally integrable, while $\sigma \in L^2_{H_1}(R_+, M(n \times d))$ and $\sigma_0 \in L^2_{H_2}(R_+, M(m \times m))$.

(A3) The random elements $\{x_0, B_{H_1}(t), V_{H_2}(t), t \geq 0\}$ are mutually statistically independent.

Recall that we want our filter to have the structure given by (2.4). Define

$$e(t) \equiv x(t) - z(t), \qquad t \geq 0, \tag{3.1}$$

where x is the solution of (2.1), and z is the solution of (2.4) corresponding to any choice of B which is locally integrable and $\Gamma \in L^\infty(R_+, M(n \times m))$. It follows from these equations that e must satisfy the stochastic differential equation

$$\begin{aligned} de &= (A - \Gamma H)e(t)dt + (A - \Gamma H - B)zdt + \sigma dB_{H_1} - \Gamma\sigma_0 dV_{H_2}, \\ e(0) &= e_0 \equiv x_0 - \hat{x}_0, \end{aligned} \tag{3.2}$$

where, for compactness of notation, we have suppressed the time variable. In fact, all the matrices appearing in (3.2) are functions of time. For an unbiased estimate, it follows from this that B must satisfy the identity $B = A - \Gamma H$. Clearly, for this choice of B, the filter equation (2.4) becomes

$$\begin{aligned} dz(t) &= (A(t) - \Gamma(t)H(t))z(t)dt + \Gamma(t)dy(t), \qquad t \geq 0, \\ z(0) &= \hat{x}_0, \end{aligned} \tag{3.3}$$

and the error equation (3.2) reduces to

$$\begin{aligned} de &= (A - \Gamma H)e(t)dt + \sigma dB_{H_1} - \Gamma\sigma_0 dV_{H_2}, \\ e(0) &= e_0. \end{aligned} \tag{3.4}$$

For any $\Gamma \in L^\infty(R_+, M(n \times m))$, let $\Phi_\Gamma(t,s), 0 \le s \le t < \infty$, denote the transition operator corresponding to $A_\Gamma(t) \equiv A(t) - \Gamma(t)H(t)$. Using this transition operator, we can write the solution of (3.4) as

$$e(t) = \Phi_\Gamma(t,0)e_0 + \int_0^t \Phi_\Gamma(t,\theta)\sigma(\theta)dB_{H_1}(\theta) - \int_0^t \Phi_\Gamma(t,\theta)\Gamma(\theta)\sigma_0(\theta)dV_{H_2}(\theta). \tag{3.5}$$

The error covariance K is defined by

$$(K(t)\xi,\eta) \equiv E\{(e(t),\xi)(e(t),\eta)\}, \qquad \xi,\eta \in R^n,\ t \ge 0. \tag{3.6}$$

From here on, we shall consider only the finite time interval $I \equiv [0,T], T < \infty$, unless stated otherwise. We shall need the following result.

LEMMA 3.1. *Suppose that assumptions* (A1)–(A3) *hold. Then, for each* $\Gamma \in L^\infty(I, M(n\times m))$, *the error covariance* K *satisfies the following functional differential equation:*

$$\begin{aligned} \dot K(t) &= A_\Gamma(t)K + KA_\Gamma'(t) + \int_0^t \varphi_{H_1}(t-s)\big\{\Phi_\Gamma(t,s)\tilde Q(s,t) + \tilde Q'(s,t)\Phi_\Gamma'(t,s)\big\}ds \\ &\quad + \int_0^t \varphi_{H_2}(t-s)\big\{\Phi_\Gamma(t,s)\Gamma(s)\tilde Q_0(s,t)\Gamma'(t) \\ &\qquad + \Gamma(t)\tilde Q_0'(s,t)\Gamma'(s)\Phi_\Gamma'(t,s)\big\}ds, \qquad t \in I, \\ K(0) &= K_0, \end{aligned} \tag{3.7}$$

where $A_\Gamma(t) = A(t) - \Gamma(t)H(t)$ *and* $\tilde Q(s,t) \equiv \sigma(s)Q\sigma'(t), \tilde Q_0(s,t) \equiv \sigma_0(s)Q_0\sigma_0'(t)$.

Proof. For $\Gamma \in L^\infty(I, M(n \times m))$, it follows from assumption (A2) that A_Γ is locally integrable, and hence the transition operator is well defined and

$$\sup\{\| \Phi_\Gamma(t,s) \|, s,t \in I\} < \infty.$$

Hence for each fixed but arbitrary $t \in I$ the functions $s \longrightarrow \Phi_\Gamma(t,s)\sigma(s)$ and $s \longrightarrow \Phi_\Gamma(t,s)\Gamma(s)\sigma_0(s)$ belong to $L^2_{H_1}(I, M(n \times d))$ and $L^2_{H_2}(I, M(n \times m))$, respectively. Hence the fractional integrals appearing in the expression (3.5) are well defined and, by virtue of the property (p1) of Lemma 1.1, their expectations vanish. Clearly the expectation of the first term of (3.5) is zero. Hence the filter is unbiased. Further, by virtue of our assumption (A3), all three terms in (3.5) are mutually statistically independent, and hence for each pair of $\xi, \eta \in R^n$ we have

$$\begin{aligned} (K(t)\xi,\eta) &\equiv E\{(e(t),\xi)(e(t),\eta)\} = \big(\Phi_\Gamma(t,0)K_0\Phi_\Gamma'(t,0)\xi,\eta\big) \\ &\quad + \int_0^t\int_0^t \varphi_{H_1}(s-\tau)(\Phi_\Gamma(t,\tau)\tilde Q(\tau,s)\Phi_\Gamma'(t,s)\xi,\eta)dsd\tau \\ &\quad + \int_0^t\int_0^t \varphi_{H_2}(s-\tau)(\Phi_\Gamma(t,\tau)\Gamma(\tau)\tilde Q_0(\tau,s)\Gamma'(s)\Phi_\Gamma'(t,s)\xi,\eta)dsd\tau, \end{aligned} \tag{3.8}$$

where K_0 is the covariance of the random variable x_0. Differentiating this term by term, recalling that $\varphi_{H_i}, i = 1, 2$, are symmetric, and adding all the terms, we find that

$$
\begin{aligned}
(\dot{K}(t)\xi,\eta) &= (A_\Gamma(t)K(t)\xi,\eta) + (K(t)A_\Gamma^{'}(t)\xi,\eta) \\
&\quad + \int_0^t \varphi_{H_1}(t-s)([\Phi_\Gamma(t,s)\tilde{Q}(s,t) + \tilde{Q}^{'}(s,t)\Phi_\Gamma^{'}(t,s)]\xi,\eta)ds \\
&\quad + \int_0^t \varphi_{H_2}(t-s)([\Phi_\Gamma(t,s)\Gamma(s)\tilde{Q}_0(s,t)\Gamma^{'}(t) \\
&\qquad\qquad + \Gamma(t)\tilde{Q}_0(s,t)\Gamma^{'}(s)\Phi_\Gamma^{'}(t,s)]\xi,\eta)ds.
\end{aligned} \tag{3.9}
$$

Here we have used the basic properties of the transition operator Φ_Γ, such as

$$
\begin{aligned}
&(\partial/\partial t)\Phi_\Gamma(t,s) = A_\Gamma(t)\Phi_\Gamma(t,s) \equiv (A(t) - \Gamma(t)H(t))\Phi_\Gamma(t,s), \qquad 0 \le s \le t, \\
&\text{and} \quad \Phi_\Gamma(t,t) = I_d, \qquad t \ge 0.
\end{aligned}
$$

Since $\xi, \eta \in R^n$ are arbitrary, (3.7) follows from (3.9) and the definitions of $\tilde{Q}(s,t)$ and $\tilde{Q}_0(s,t)$. This completes the proof. □

Now we are prepared to formulate the filtering problem as a control problem. First we recall that (3.3), with Γ to be determined, gives an unbiased estimate of x. For a minimum variance estimate, we must now choose Γ so that $TrK(t)$ is minimum. We consider a more general problem which covers the filtering problem. Let T be any arbitrary but finite time with $I \equiv [0,T]$, and Σ any real positive definite symmetric matrix-valued function—for example, $\Sigma \in L^1(I, M_s^+(n \times n))$. Define

$$
J(\Gamma) = \int_0^T Tr(\Sigma(t)K(t))dt. \tag{3.10}
$$

For compactness of notation, set $\mathcal{G} \equiv L^\infty(I, M(n \times m))$. Then the optimum filtering problem is equivalent to the problem: find $\Gamma \in \mathcal{G}$ that imparts a minimum to the functional J subject to the dynamic constraint (3.7).

The first question that must be settled is whether the problem, as stated, has a solution. This is answered in the following corollary.

THEOREM 3.2. *Suppose that assumptions* (A1)–(A3) *hold and* $\Sigma \in L^1(I, M_s^+((n \times n))$. *Then, the optimal control problem as stated above has a solution.*

Proof. First assume that $\Sigma = I_d$, meaning that $J(\Gamma) = \int_I TrK(t)dt$. Using (3.8) and computing the $TrK(t)$ with respect to any basis, say $\{e_i, i = 1, 2, 3, \ldots, n\}$, because of the fourth right-hand-side term of (3.9), which involves $\Gamma(s), \Gamma(t)$, one finds that $J(\Gamma) \longrightarrow \infty$ as $\| \Gamma \|_{L^\infty(I, M(n \times m))} \longrightarrow \infty$. Further, for any fixed $\Gamma \in \mathcal{G}$, $J(\Gamma) < \infty$ and $J(\Gamma) \ge 0$. Define

$$
\mu \equiv \inf\{J(\Gamma), \Gamma \in \mathcal{G}\}.
$$

Clearly, it follows from the above comments that the infimum exists. Let $\{\Gamma_n\}$ be a minimizing sequence. Clearly this is a bounded sequence, and hence there exists a subsequence, relabeled as such, and an element $\Gamma_o \in \mathcal{G}$ such that $\Gamma_n \to \Gamma_o$ in the weak-star topology of $L^\infty(I, M(n \times m))$. This implies that $\Phi_{\Gamma_n}(t,s) \longrightarrow \Phi_{\Gamma_o}(t,s)$ pointwise on the triangle $0 \le s \le t \le T$ (see [6, Theorem 2.3.7]). This fact is proved using the generalized Gronwall inequality. Let K_n denote the solution of (3.7) corresponding to Γ_n, and K_o that corresponding to Γ_o. Using the weak-star convergence and the

pointwise convergence stated above, it follows from Lebesgue dominated convergence theorem that

$$Tr(K_o(t)) \leq \liminf_{n\to\infty} Tr(K_n(t)), \qquad \text{a.e. } t \in I.$$

Then by Fatou's lemma we obtain

$$\int_0^T Tr(K_o(t))\ dt \leq \liminf_{n\to\infty} \int_0^T Tr(K_n(t))dt.$$

This is equivalent to

$$J(\Gamma_o) \leq \liminf_{n\to\infty} J(\Gamma_n).$$

Hence $\mu = J(\Gamma_o)$. In other words, the infimum is attained. This proves that the optimization problem has a solution, given that Σ is an identity matrix. Note that $K \in C(I, M_s^+(n\times n)) \subset L^\infty(I, M_s^+(n\times n))$, and hence the functional (3.10) is well-defined. Thus, for a general symmetric positive definite matrix-valued function Σ, the result follows from orthogonal transformation. This completes the proof. □

Remark. The existence result given above also holds if the set $\mathcal{G}$ is a closed bounded convex subset of $L^\infty(I, M(n\times m))$. This follows from the facts that J is weak-star lower semicontinuous, as demonstrated above, and that $\mathcal{G}$ is weak-star compact.

4. Optimal filter. We have seen in the preceding section that an optimum linear filter exists and that it can be determined by solving the control problem

$$J(\Gamma) = \int_0^T Tr(\Sigma(t)K(t))dt \longrightarrow \min. \tag{4.1}$$

subject to the dynamic constraint

$$\begin{aligned}\dot{K}(t) = {} & A_\Gamma(t)K + KA_\Gamma'(t) + \int_0^t \varphi_{H_1}(t-s)\big\{\Phi_\Gamma(t,s)\tilde{Q}(s,t) + \tilde{Q}'(s,t)\Phi_\Gamma'(t,s)\big\}ds \\ & + \int_0^t \varphi_{H_2}(t-s)\big\{\Phi_\Gamma(t,s)\Gamma(s)\tilde{Q}_0(s,t)\Gamma'(t) \\ & \qquad + \Gamma(t)\tilde{Q}_0'(s,t)\Gamma'(s)\Phi_\Gamma'(t,s)\big\}ds, \qquad t \in I, \\ K(0) = {} & K_0,\end{aligned} \tag{4.2}$$

with $\Gamma \in \mathcal{G}$.

To obtain the necessary conditions of optimality and finally the optimum filter, we use the variational technique as in [4, 5]. For this we shall need the Gateaux differential of K with respect to Γ on $\mathcal{G}$. In general we may assume that $\mathcal{G}$ is any closed convex subset of $L^\infty(I, M(n\times m))$. Let $\Gamma_o \in \mathcal{G}$ be the optimal control, and $\Gamma \in \mathcal{G}$ any other arbitrary element. We show that the Gateaux differential of K at Γ_o in the direction $(\Gamma - \Gamma_o)$ is given by a functional differential equation. For this we must show that the transition operator Φ_Γ is Gateaux differentiable. This is stated in the following result.

LEMMA 4.1. *The Gateaux differential of the map* $\Gamma \longrightarrow \Phi_\Gamma$ *at* Γ_o *in the direction* $\Gamma - \Gamma_o$, *denoted by* $\tilde{\Phi}$, *is given by*

$$\tilde{\Phi}(t,\theta) = -\int_\theta^t ds\ \Phi_{\Gamma_o}(t,s)(\Gamma(s) - \Gamma_o(s))H(s)\Phi_{\Gamma_o}(s,\theta), \tag{4.3}$$

which satisfies the following differential equation:

$$(\partial/\partial t)\tilde{\Phi}(t,\theta) = (A(t) - \Gamma_o(t)H(t))\tilde{\Phi}(t,\theta) - (\Gamma(t) - \Gamma_o(t))H(t)\Phi_{\Gamma_o}(t,\theta), \quad 0 \le \theta \le t,$$
$$\tilde{\Phi}(\theta,\theta) = 0, \quad \theta \ge 0.$$

Further, as $\Gamma \to \Gamma_o$, $\tilde{\Phi}(t,\theta) \to 0$ *uniformly on* $I \times I$.

Proof. The proof follows directly from straightforward computation. □

LEMMA 4.2. *Let* $\mathcal{G}$ *be any closed convex subset of* $L^\infty(I, M(n\times m))$. *Then, for each pair of* $\Gamma_o, \Gamma \in \mathcal{G}$, *the Gateaux differential of* K *at* $\Gamma_o \in \mathcal{G}$ *in the direction* $\Gamma - \Gamma_o$, *denoted by* $\tilde{K}$, *is the solution of the functional differential equation*

$$\begin{aligned}
\dot{\tilde{K}}(t) &= A_{\Gamma o}(t)\tilde{K}(t) + \tilde{K}(t)A_{\Gamma o}^{'}(t) - (\Gamma(t) - \Gamma_o(t))H(t)K_o(t) \\
&\quad - K_o(t)H^{'}(t)(\Gamma(t) - \Gamma_o(t))^{'} \\
&\quad + \int_0^t \varphi_{H_1}(t-s)\big\{\tilde{\Phi}(t,s)\tilde{Q}(s,t) + \tilde{Q}^{'}(s,t)\tilde{\Phi}^{'}(t,s)\big\}ds \\
&\quad + \int_0^t \varphi_{H_2}(t-s)\big\{\tilde{\Phi}(t,s)\Gamma_o(s)\tilde{Q}_0(s,t)\Gamma_o^{'}(t) + \Gamma_o(t)\tilde{Q}_0^{'}(s,t)\Gamma_o^{'}(s)\tilde{\Phi}^{'}(t,s)\big\}ds \\
&\quad + \int_0^t \varphi_{H_2}(t-s)\big\{\Phi_{\Gamma_o}(t,s)(\Gamma - \Gamma_o)(s)\tilde{Q}_0(s,t)\Gamma_o^{'}(t) \\
&\qquad\qquad + \Gamma_o(t)\tilde{Q}_0^{'}(s,t)(\Gamma - \Gamma_o)^{'}(s)\Phi_{\Gamma_o}^{'}(t,s)\big\}ds \\
&\quad + \int_0^t \varphi_{H_2}(t-s)\big\{\Phi_{\Gamma_o}(t,s)\Gamma_o(s)\tilde{Q}_0(s,t)(\Gamma - \Gamma_o)^{'}(t) \\
&\qquad\qquad + (\Gamma - \Gamma_o)(t)\tilde{Q}_0^{'}(s,t)\Gamma_o^{'}(s)\Phi_{\Gamma_o}^{'}(t,s)\big\}ds,
\end{aligned} \tag{4.4}$$
$$\tilde{K}(0) = 0.$$

Proof. The proof is lengthy but straightforward. We give the basic outline. Let $\Gamma_o, \Gamma \in \mathcal{G}$ and $\varepsilon \in [0,1]$, and define $\Gamma_\varepsilon = \Gamma_o + \varepsilon(\Gamma - \Gamma_o)$. Since $\mathcal{G}$ is a closed convex set, $\Gamma_\varepsilon \in \mathcal{G}$. Let K_ε and K_o denote the solutions of (4.2) corresponding to Γ_ε and Γ_o, respectively. Using Lemma 4.1 and the continuous dependence of solutions $\Gamma \longrightarrow K$, one can verify that the limit

$$\lim_{\varepsilon \downarrow 0}(1/\varepsilon)\{K_\varepsilon - K_o\}$$

exists, and that this limit is the solution of (4.4). □

Now we are prepared to present the optimal filter equations. Define the following functionals:

$$\begin{aligned}
F_1(t,K,\Gamma) &\equiv A_\Gamma K + KA_\Gamma^{'} + \int_0^t \varphi_{H_1}(t-s)\big\{\Phi_\Gamma(t,s)\tilde{Q}(s,t) + \tilde{Q}^{'}(s,t)\Phi_\Gamma^{'}(t,s)\big\}ds \\
&\quad + \int_0^t \varphi_{H_2}(t-s)\big\{\Phi_\Gamma(t,s)\Gamma(s)\tilde{Q}_0(s,t)\Gamma^{'}(t) + \Gamma(t)\tilde{Q}_0^{'}(s,t)\Gamma^{'}(s)\Phi_\Gamma^{'}(t,s)\big\}ds
\end{aligned} \tag{4.5}$$

and

$$F_2(t,K,\Gamma) \equiv \int_0^t \varphi_{H_2}(t-s)\{\tilde{Q}_0(s,t)\Gamma'(t) + \tilde{Q}_0'(s,t)\Gamma'(s)\Phi_\Gamma'(t,s) - C_\Gamma(t,s)\Gamma(s)\tilde{Q}_0(s,t)\Gamma'(t)\}ds - \int_0^t \varphi_{H_1}(t-s)C_\Gamma(t,s)\tilde{Q}(s,t)ds - HK, \tag{4.6}$$

where C_Γ is given by

$$C_\Gamma(t,s) \equiv \int_s^t H(\theta)\Phi_\Gamma(\theta,s)d\theta. \tag{4.7}$$

THEOREM 4.3. *Suppose that the assumptions of Theorem* 3.2 *hold. Then, the optimum linear filter is given by the stochastic differential equation*

$$\begin{aligned} dz &= (A - \Gamma_o H)zdt + \Gamma_o dy, \\ z(0) &= \hat{x}_0, \end{aligned} \tag{4.8}$$

where the pair $\{\Gamma_o, K_o\}$ *satisfies the following functional differential equations:*

$$\begin{aligned} \dot{K}_o &= F_1(t,K_o,\Gamma_o), \quad K_o(0) = K_0, \qquad t \in I, \\ 0 &= F_2(t,K_o,\Gamma_o), \qquad t \in I. \end{aligned} \tag{4.9}$$

Proof. Suppose that $\Gamma_o \in \mathcal{G}$ minimizes the functional $J(\Gamma)$ as defined by (3.10). Let Γ be any element of $\mathcal{G}$, and define $\Gamma_\varepsilon \equiv \Gamma_o + \varepsilon(\Gamma - \Gamma_o)$ for $\varepsilon \in [0,1]$. Since $\mathcal{G}$ is a closed convex set, $\Gamma_\varepsilon \in \mathcal{G}$. Clearly, by the optimality of Γ_o,

$$J(\Gamma_\varepsilon) - J(\Gamma_o) \geq 0 \qquad \text{for all } \varepsilon \in [0,1].$$

Since by Lemma 4.2 the map $\Gamma \longrightarrow K$ is Gateaux differentiable, and its Gateaux derivative at Γ_o in the direction $\Gamma - \Gamma_o$ is given by the solution $\tilde{K}$ of (4.4), J is also Gateaux differentiable at Γ_o in the direction $\Gamma - \Gamma_o$, and hence

$$dJ(\Gamma_o, \Gamma - \Gamma_o) = \int_0^T Tr\{\Sigma\tilde{K}\}dt \geq 0 \qquad \text{for all } \Gamma \in \mathcal{G}. \tag{4.10}$$

Here we consider only the case $\mathcal{G} = L^\infty(I, M(n \times m))$, which is much easier. We shall have some comments later for the constrained case. In the unconstrained case the inequality reduces to the identity

$$dJ(\Gamma_o, \Gamma - \Gamma_o) = \int_0^T Tr\{\Sigma\tilde{K}\}dt = 0 \qquad \text{for all } \Gamma \in \mathcal{G}. \tag{4.11}$$

Since $\Sigma \in L^1(I, M_s^+(n \times n))$ is arbitrary and positive definite and T is any finite positive number, this identity can hold if and only if $\tilde{K}$ is identically zero. Since $\tilde{K}$ is the solution of the variational equation (4.4) with initial condition $\tilde{K}(0) = 0$, and the right-hand expression, from the third term through the last term, is a linear functional of the difference $(\Gamma - \Gamma_o)$, this is possible if and only if this nonhomogeneous term

vanishes identically for all $\Gamma \in \mathcal{G}$. Thus it is necessary that the following identity hold for all $\Gamma \in \mathcal{G}$:

$$
\begin{aligned}
0 = &- (\Gamma(t) - \Gamma_o(t))H(t)K_o(t) - K_o(t)H^{'}(t)(\Gamma(t) - \Gamma_o(t))^{'} \\
&+ \int_0^t \varphi_{H_1}(t-s)\big\{\tilde{\Phi}(t,s)\tilde{Q}(s,t) + \tilde{Q}^{'}(s,t)\tilde{\Phi}^{'}(t,s)\big\}ds \\
&+ \int_0^t \varphi_{H_2}(t-s)\big\{\tilde{\Phi}(t,s)\Gamma_o(s)\tilde{Q}_0(s,t)\Gamma_o^{'}(t) + \Gamma_o(t)\tilde{Q}_0^{'}(s,t)\Gamma_o^{'}(s)\tilde{\Phi}^{'}(t,s)\big\}ds \\
&+ \int_0^t \varphi_{H_2}(t-s)\big\{\Phi_{\Gamma_o}(t,s)(\Gamma - \Gamma_o)(s)\tilde{Q}_0(s,t)\Gamma_o^{'}(t) \\
&\qquad + \Gamma_o(t)\tilde{Q}_0^{'}(s,t)(\Gamma - \Gamma_o)^{'}(s)\Phi_{\Gamma_o}^{'}(t,s)\big\}ds \\
&+ \int_0^t \varphi_{H_2}(t-s)\big\{\Phi_{\Gamma_o}(t,s)\Gamma_o(s)\tilde{Q}_0(s,t)(\Gamma - \Gamma_o)^{'}(t) \\
&\qquad + (\Gamma - \Gamma_o)(t)\tilde{Q}_0^{'}(s,t)\Gamma_o^{'}(s)\Phi_{\Gamma_o}^{'}(t,s)\big\}ds
\end{aligned}
\tag{4.12}
$$

for all $t \in I$. For simplicity of presentation we denote the members on the right-hand side of (4.12) as follows:

$$
\alpha_1 \equiv -(\Gamma(t) - \Gamma_o(t))H(t)K_o(t) - K_o(t)H^{'}(t)(\Gamma(t) - \Gamma_o(t))^{'}, \tag{4.13}
$$

$$
\alpha_2 \equiv \int_0^t \varphi_{H_1}(t-s)\big\{\tilde{\Phi}(t,s)\tilde{Q}(s,t) + \tilde{Q}^{'}(s,t)\tilde{\Phi}^{'}(t,s)\big\}ds, \tag{4.14}
$$

$$
\alpha_3 \equiv \int_0^t \varphi_{H_2}(t-s)\big\{\tilde{\Phi}(t,s)\Gamma_o(s)\tilde{Q}_0(s,t)\Gamma_o^{'}(t) + \Gamma_o(t)\tilde{Q}_0^{'}(s,t)\Gamma_o^{'}(s)\tilde{\Phi}^{'}(t,s)\big\}ds, \tag{4.15}
$$

$$
\begin{aligned}
\alpha_4 \equiv \int_0^t \varphi_{H_2}(t-s)\big\{&\Phi_{\Gamma_o}(t,s)(\Gamma - \Gamma_o)(s)\tilde{Q}_0(s,t)\Gamma_o^{'}(t) \\
&+ \Gamma_o(t)\tilde{Q}_0^{'}(s,t)(\Gamma - \Gamma_o)^{'}(s)\Phi_{\Gamma_o}^{'}(t,s)\big\}ds,
\end{aligned}
\tag{4.16}
$$

$$
\begin{aligned}
\alpha_5 \equiv \int_0^t \varphi_{H_2}(t-s)\big\{&\Phi_{\Gamma_o}(t,s)\Gamma_o(s)\tilde{Q}_0(s,t)(\Gamma - \Gamma_o)^{'}(t) \\
&+ (\Gamma - \Gamma_o)(t)\tilde{Q}_0^{'}(s,t)\Gamma_o^{'}(s)\Phi_{\Gamma_o}^{'}(t,s)\big\}ds.
\end{aligned}
\tag{4.17}
$$

According to (4.12), the sum

$$
\alpha \equiv \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5 = 0 \qquad \text{for all } \ \Gamma \in \mathcal{G} \text{ and all } \ t \in I. \tag{4.18}
$$

Since this identity must hold for all $\Gamma \in \mathcal{G}$, it must also hold for any Γ of the form

$$
\Gamma(s) = \Gamma_o(s) + X_o(s)D \tag{4.19}
$$

for all constant matrices $D \in M(n \times m)$, where X_o is the fundamental solution of the equation

$$
\dot{X}_o(t) = (A(t) - \Gamma_o(t)H(t))X_0(t), \quad X(0) = I_d, \tag{4.20}
$$

with $I_d \in M(n \times n)$ being the identity matrix. This gives

$$\alpha_1 \equiv -X_o(t)DH(t)K_o(t) - K_o(t)H^{'}(t)(X_0(t)D)^{'}. \tag{4.21}$$

Denote

$$C_o(t,s) \equiv C_{\Gamma_o}(t,s) = \int_s^t H(\theta)\Phi_{\Gamma_o}(\theta,s)d\theta, \tag{4.22}$$

and note that for Γ given by (4.19), $\tilde{\Phi}$ of (4.3) takes the form

$$\begin{aligned}\tilde{\Phi}(t,\theta) &= -X_o(t)D\bigg(\int_\theta^t ds\ H(s)\Phi_{\Gamma_o}(s,\theta)\bigg)\\ &= -X_o(t)DC_o(t,\theta).\end{aligned} \tag{4.23}$$

Using these, we obtain

$$\begin{aligned}\alpha_2 =& -X_o(t)D\bigg\{\int_0^t \varphi_{H_1}(t-s)C_o(t,s)\tilde{Q}(s,t)ds\bigg\}\\ &-\bigg\{\int_0^t \varphi_{H_1}(t-s)\tilde{Q}^{'}(s,t)C_o^{'}(t,s)ds\bigg\}D^{'}X_o^{'}(t),\end{aligned} \tag{4.24}$$

$$\begin{aligned}\alpha_3 =& -X_o(t)D\bigg\{\int_0^t \varphi_{H_2}(t-s)C_o(t,s)\Gamma_o(s)\tilde{Q}_0(s,t)\Gamma_o^{'}(t)ds\bigg\}\\ &-\bigg\{\int_0^t \varphi_{H_2}(t-s)\Gamma_o(t)\tilde{Q}_0^{'}(s,t)\Gamma_o^{'}(s)C_o^{'}(t,s)ds\bigg\}D^{'}X_o^{'}(t),\end{aligned} \tag{4.25}$$

$$\begin{aligned}\alpha_4 =& X_o(t)D\bigg\{\int_0^t \varphi_{H_2}(t-s)\tilde{Q}_0(s,t)\Gamma_o^{'}(t)ds\bigg\}\\ &+\bigg\{\int_0^t \varphi_{H_2}(t-s)\Gamma_o(t)\tilde{Q}_0^{'}(s,t)ds\bigg\}D^{'}X_o^{'}(t),\end{aligned} \tag{4.26}$$

and

$$\begin{aligned}\alpha_5 =& X_o(t)D\bigg\{\int_0^t \varphi_{H_2}(t-s)\tilde{Q}_0^{'}(s,t)\Gamma_o^{'}(s)\Phi_{\Gamma_o}^{'}(t,s)ds\bigg\}\\ &+\bigg\{\int_0^t \varphi_{H_2}(t-s)\Phi_{\Gamma_o}(t,s)\Gamma_o(s)\tilde{Q}_0(s,t)ds\bigg\}D^{'}X_o^{'}(t).\end{aligned} \tag{4.27}$$

Adding all the α's, we arrive at an expression of the form

$$\alpha = X_o(t)DM(t) + M^{'}(t)D^{'}X_o^{'}(t), \tag{4.28}$$

where

$$\begin{aligned}M(t) \equiv& -H(t)K_o(t) + \int_0^t \varphi_{H_2}(t-s)\bigg\{-C_o(t,s)\Gamma_o(s)\tilde{Q}_0(s,t)\Gamma_o^{'}(t)\\ &+\tilde{Q}_0(s,t)\Gamma_o^{'}(t) + \tilde{Q}_0^{'}(s,t)\Gamma_o^{'}(s)\Phi_{\Gamma_o}^{'}(t,s)\bigg\}ds\\ &-\int_0^t \varphi_{H_1}(t-s)C_o(t,s)\tilde{Q}(s,t)ds.\end{aligned}$$

By virtue of (4.18) and (4.28), we have

$$\alpha = X_o(t)DM(t) + M^{'}(t)D^{'}X_o^{'}(t) = 0 \qquad \text{for all } D \in M(n \times m),\ t \in I. \tag{4.29}$$

Thus

$$Tr\{X_o(t)DM(t)\} = 0 \qquad \text{for all } D \in M(n \times m),\ t \in I. \tag{4.30}$$

Since $X_o(t)$ is always nonsingular, this implies that $M(t) \equiv 0$, and hence

$$\begin{aligned} F_2(t, K_o, \Gamma_o) \equiv -H(t)K_o(t) + \int_0^t \varphi_{H_2}(t-s)\Big\{ & \tilde{Q}_0(s,t)\Gamma_o^{'}(t) + \tilde{Q}_0^{'}(s,t)\Gamma_o^{'}(s)\Phi_{\Gamma_o}^{'}(t,s) \\ & - C_o(t,s)\Gamma_o(s)\tilde{Q}_0(s,t)\Gamma_o^{'}(t)\Big\}ds \\ - \int_0^t \varphi_{H_1}(t-s)C_o(t,s)\tilde{Q}(s,t)ds = 0. & \end{aligned} \tag{4.31}$$

Thus it follows from (3.7), (4.5), and (4.31) that the optimal pair must satisfy the functional differential equation (4.9), and consequently the optimum filter must be given by (4.8), with the pair $\{\Gamma_o, K_o\}$ being the solution of (4.9). This completes the proof. □

Remark. When $\mathcal{G}$ is a closed convex, possibly bounded, (proper) subset of $L^\infty(I, M(n \times m))$, the necessary conditions of optimality involve an additional equation, the adjoint equation associated with the (state) functional equation (4.2). Readers interested in constrained filtering may see [4, 5].

Remark. If $\sigma_0(t)$ is nonsingular, the filter equation (4.8) can be rewritten as

$$dz = A(t)zdt + \Gamma_o\ \sigma_0\ d\nu_H(t),$$

where ν_H, given by

$$\nu_H(t) = \int_0^t \sigma_0^{-1}(s)\{dy(s) - H(s)z(s)ds\}, \qquad t \geq 0,$$

is an $\mathcal{F}_t^y$-measurable Gaussian process. However, according to [16], ν_H is not the innovation process, and thus, it is not an FBM with the same Hurst parameter.

Some special cases. Here we adapt the results derived for the special cases, when either of the two FBMs are replaced by standard Brownian motions, and we further show that the convergence methodology yields the Kalman filter equations when both noises are standard Brownian motions. Since the difference between the FBM and the standard Brownian motion is captured through the function $\varphi_{H_j}, j = 1, 2$, we shall show that as the Hurst parameters tend to $1/2$, φ_{H_j} converges to a delta function. Once this result is derived, it can be used in the previous general filter equations to study special cases.

(C1) We show here that if the Hurst parameters $H_i \to (1/2), i = 1, 2$, our filter equations reduce to the classical Kalman filter equations. We need the following lemma.

LEMMA 4.4. *As* $H_i \to (1/2)$, $\varphi_{H_i}(t) \to (1/2)\delta(t), i = 1, 2$.

Proof. Let $H = H_i = (1/2) + \varepsilon$, $i = 1, 2$, $\varepsilon > 0$. Define $\psi_\varepsilon(t) = \varphi_{(1/2)+\varepsilon}(t)$. Let $C_0^\infty[0, \infty)$ denote the class of C^∞ functions with compact supports, and take $\xi \in C_0^\infty[0, \infty)$. Consider the functional

$$f_\varepsilon(\xi) \equiv \int_0^\infty \xi(t)\psi_\varepsilon(t)dt.$$

Then it follows from our definition of φ_H (1.3) that

$$\begin{aligned} f_\varepsilon(\xi) &\equiv \varepsilon(1+2\varepsilon)\int_0^\infty \xi(t)|t|^{2\varepsilon-1}dt \\ &= \varepsilon(1+2\varepsilon)\int_0^\infty \xi(t)\ t^{2\varepsilon-1}dt \\ &= \frac{(1+2\varepsilon)}{2}\int_0^\infty \xi(t^{1/2\varepsilon})dt. \end{aligned}$$

We split the integral as

$$\begin{aligned} f_\varepsilon(\xi) &= \frac{(1+2\varepsilon)}{2}\int_0^\infty \xi(t^{1/2\varepsilon})\ dt \\ &= \frac{(1+2\varepsilon)}{2}\left\{\int_0^{1-0} \xi(t^{1/2\varepsilon})\ dt + \int_{1+0}^\infty \xi(t^{1/2\varepsilon})\ dt\right\}. \end{aligned}$$

Since ξ has compact support on $[0,\infty)$ and is C^∞, one can easily justify that

$$\lim_{\varepsilon\downarrow 0}\frac{(1+2\varepsilon)}{2}\int_0^{1-0} \xi(t^{1/2\varepsilon})\ dt = (1/2)\xi(0)$$

and

$$\lim_{\varepsilon\downarrow 0}\frac{(1+2\varepsilon)}{2}\int_{1+0}^\infty \xi(t^{1/2\varepsilon})\ dt = 0.$$

Using these facts, we arrive at the following identity:

$$\lim_{\varepsilon\downarrow 0} f_\varepsilon(\xi) = \left(\frac{1}{2}\right)\xi(0) \qquad \text{for every } \xi \in C_0^\infty[0,\infty). \tag{4.32}$$

In fact, it follows from the same arguments that, for any $\xi \in C_0^\infty(R)$,

$$\lim_{\varepsilon\downarrow 0} f_\varepsilon(\xi) = \xi(0) \qquad \text{for every } \xi \in C_0^\infty(R). \tag{4.33}$$

In other words, as $H_i \downarrow (1/2)$, $\varphi_{H_i}, i = 1,2$, converge to the Dirac measures concentrated at the origin. This proves the assertion. □

Now we are prepared to derive the classical Kalman filter from our main result given in Theorem 4.3. For given σ and σ_0, define, for each $t \geq 0$, the (auto) covariance matrices:

$$\begin{aligned} Q(t) &\equiv \tilde{Q}(t,t) = \sigma(t)Q\sigma'(t), \\ Q_0(t) &\equiv \tilde{Q}_0(t,t) = \sigma_0(t)Q_0\sigma_0'(t). \end{aligned} \tag{4.34}$$

COROLLARY 4.5. *Suppose that $Q_0(t)$, as defined by* (4.34)*, is nonsingular. Then the Kalman filter is given by the error covariance equation,*

$$\begin{aligned} &\dot{K}_o(t) = A(t)K_o(t) + K_o(t)A'(t) + Q(t) - K_o(t)H'(t)Q_0^{-1}(t)H(t)K_o(t), \\ &K(0) = K_0, \end{aligned} \tag{4.35}$$

and the filter equation,

$$\begin{aligned} &dz(t) = A(t)z(t)dt + K_o(t)H'(t)Q_0^{-1}(t)(dy(t) - H(t)z(t)dt), \\ &z(0) = \hat{x}_0. \end{aligned} \tag{4.36}$$

Proof. By virtue of Lemma 4.4, letting $H_i \downarrow (1/2), i = 1, 2$, it follows from the expressions (4.5) and (4.6) that

$$\begin{aligned} F_1(t, K, \Gamma) &= A_\Gamma(t)K(t) + K(t)A_\Gamma^{'}(t) + Q(t) + \Gamma Q_0(t)\Gamma^{'}(t), \\ F_2(t, K, \Gamma) &= Q_0(t)\Gamma^{'}(t) - H(t)K(t). \end{aligned} \tag{4.37}$$

For optimality, it follows from Theorem 4.3 that both the equations of (4.9) must be satisfied. The second equation of (4.9) requires that

$$F_2(t, K_0, \Gamma_o) = Q_0(t)\Gamma_o^{'}(t) - H(t)K_o(t) = 0.$$

Taking its transpose and using the assumption that $Q_0(t)$ is invertible, we obtain the optimal Kalman gain

$$\Gamma_o(t) = K_o(t)H^{'}(t)Q_0(t)^{-1}. \tag{4.38}$$

Substituting this into F_1 given by (4.37), we obtain

$$F_1(t, K_o, \Gamma_o) = A(t)K_o(t) + K_o(t)A^{'}(t) + Q(t) - K_o(t)H^{'}(t)Q_0^{-1}(t)H(t)K_o(t).$$

Using this expression in the first equation of (4.9), we obtain the classical matrix Riccati differential equation (4.35). Substituting the expression for Γ_o from (4.38) into (4.8), we obtain the filter equation (4.36). Hence the Kalman filter follows from our Theorem 4.3. This completes the proof. □

Remark. Note that if $Q_0(t)$ is not invertible, we do not have the Kalman filter, or more precisely, there are many Kalman filter solutions. To choose a unique solution, one can consider minimizing the norm of K_o subject to the constraint (4.9).

Two other special cases, in which only one of the two Brownian motions is fractional, are given in the following corollary.

COROLLARY 4.6. *Suppose that the following conditions hold:*

(C2) *The signal process $\{x\}$ is perturbed by Q-Brownian motion, and the measurement process $\{y\}$ is driven by Q_0-fractional Brownian motion.*

(C3) *The signal process $\{x\}$ is perturbed by Q-fractional Brownian motion, and the measurement process $\{y\}$ is driven by Q_0-Brownian motion.*

Then the filter equations for these cases are given by (4.8)*, while the covariance equations are given by*

$$\begin{aligned} \dot{K} = A_\Gamma K + KA_\Gamma^{'} + Q(t) + \int_0^t \varphi_{H_2}(t-s)\Big\{ &\Phi_\Gamma(t,s)\Gamma(s)\tilde{Q}_0(s,t)\Gamma^{'}(t) \\ &+ \Gamma(t)\tilde{Q}_0^{'}(s,t)\Gamma^{'}(s)\Phi_\Gamma^{'}(t,s)\Big\}ds, \\ 0 = \int_0^t \varphi_{H_2}(t-s)\Big\{ &\tilde{Q}_0(s,t)\Gamma^{'}(t) + \tilde{Q}_0^{'}(s,t)\Gamma^{'}(s)\Phi_\Gamma^{'}(t,s) \\ &- C_\Gamma(t,s)\Gamma(s)\tilde{Q}_0(s,t)\Gamma^{'}(t)\Big\}ds - H(t)K(t) \end{aligned} \tag{4.39}$$

for (C2)*, and by*

$$\begin{aligned} \dot{K} = A_\Gamma K + KA_\Gamma^{'} + \int_0^t \varphi_{H_1}(t-s)\Big\{ &\Phi_\Gamma(t,s)\tilde{Q}(s,t) + \tilde{Q}^{'}(s,t)\Phi_\Gamma^{'}(t,s)\Big\}ds \\ &+ \Gamma(t)Q_0(t)\Gamma^{'}(t), \\ 0 = Q_0(t)\Gamma^{'}(t) - \int_0^t \varphi_{H_1}(t-s)C_\Gamma(t,s)\tilde{Q}(s,t)ds &- H(t)K(t) \end{aligned} \tag{4.40}$$

for (C3).

Proof. The proof is a straightforward application of Theorem 4.3, with $\varphi_{H_i}, i = 1, 2$, replaced by the Dirac measures in the appropriate sections of (4.5) and (4.6). □

5. Optimal filter for dynamically coupled systems. Here the signal and the observed processes are governed by the following set of stochastic differential equations:

$$\begin{aligned} dx(t) &= A(t)x(t)dt + B(t)y(t)dt + \sigma(t)dB_{H_1}(t), \quad x(0) = x_0, \\ dy(t) &= H(t)x(t)dt + C(t)y(t)dt + \sigma_0(t)dV_{H_2}(t), \quad y(0) = 0, \end{aligned} \tag{5.1}$$

where B and C are the coupling matrices taking values from $M(n\times m)$ and $M(m\times m)$, respectively. In this case we have the following result.

THEOREM 5.1. *Suppose that the assumptions of Theorem* 4.3 *hold and, further, that B and C are locally integrable. Then the optimum linear filter for the system* (5.1) *is given by*

$$\begin{aligned} dz &= (A - \Gamma_o H)zdt + (B - \Gamma_o C)ydt + \Gamma_o dy, \\ z(0) &= \hat{x}_0, \end{aligned} \tag{5.2}$$

where the optimum gain Γ_o is given by the solutions of (4.9).

Proof. Since the proof is very similar to the previous case, we give only an outline of the proof. Consider the following linear structure for the optimum filter:

$$\begin{aligned} dz &= Fzdt + Gydt + \Gamma dy, \\ z(0) &= \hat{x}_0, \end{aligned} \tag{5.3}$$

where the matrix-valued functions $\{F, G, \Gamma\}$ must be chosen to give an unbiased minimum variance filter. Define $e = x - z$, and note that e satisfies the stochastic differential equation

$$\begin{aligned} de = (A - \Gamma H)edt + (A - \Gamma H - F)zdt + (B - G - \Gamma C)ydt \\ + \sigma dB_{H_1} - \Gamma\sigma_0 dV_{H_2}, \\ e(0) = e_0 \equiv x_0 - \hat{x}_0. \end{aligned} \tag{5.4}$$

One can show that for the unbiased filter we must have $F = A - \Gamma H$ and $G = B - \Gamma C$. This reduces the error equation to (3.4), and consequently the error covariance equation remains unchanged. This means that the optimum Γ must be determined by the solution of (4.9) as in Theorem 4.3. Once this is done, we obtain the optimum filter (5.2). This completes the outline of our proof. □

6. An algorithm for computation. Here we briefly present a conceptual algorithm for computing the optimum filter gain. We do this by using the result of Theorem 4.3. Suppose that the nth stage of iteration has been reached and Γ^n has been determined.

Step 1: Use $\Gamma = \Gamma^n$ to solve for K^n using the first equation of (4.9).

Step 2: Use K^n in the second equation of (4.9). If $F_2(t, K^n, \Gamma^n) \neq 0$, solve the corresponding functional equation given by $F_2(t, K^n, \Gamma) = 0$, and call the solution Γ^{n+1}.

Step 3: Use a suitable metric, for example, the metric induced by the standard L^∞ norm, to compute

$$d(\Gamma^{n+1}, \Gamma^n) \equiv \| \Gamma^{n+1} - \Gamma^n \|_{L^\infty(I, M(n\times m))},$$

and stop if a predetermined level of accuracy has been met, and print; if not, go to step 1.

As discussed before, the nonuniqueness of solutions can be addressed by further minimizing the norm of Γ, subject to the constraint $0 = F_2(t, K, \Gamma)$, which will result in a unique solution of the filtering problem (see [4, 5]).

It may be interesting to investigate whether the functional equation (4.9) can be solved by the use of singular perturbation techniques. An interesting point raised by the referee is the investigation of the robustness properties of the filter with respect to the Hurst parameters. In principle, this problem is tractable through the computation of the norm of the difference between two Γ's corresponding to different Hurst parameters.

Acknowledgment. The authors wish to express their thanks to the referee for his valuable comments.

REFERENCES

[1] B.B. Mandelbrot and J.W. Van Ness, *Fractional Brownian, motions, fractional noises and applications*, SIAM Rev., 10 (1968), pp. 422–437.

[2] T.E. Duncan, Y. Hu, and B. Pasik-Duncan, *Stochastic calculus for fractional Brownian motion I. Theory*, SIAM J. Control Optim., 38 (1999), pp. 582–612.

[3] T.E. Duncan, Y. Hu, and B. Pasik-Duncan, *Some methods of stochastic calculus for fractional Brownian motions*, in Proceedings of the 38th Conference on Decision and Control, Phoenix, AZ, 1999, pp. 2390–2393.

[4] N.U. Ahmed and P. Li, *Quadratic regulator theory and linear filtering under system constraints*, IMA J. Math. Control Inform., 8 (1991), pp. 93–107.

[5] N.U. Ahmed, *Linear and Nonlinear Filtering for Scientists and Engineers*, World Scientific Publishers, London, River Edge, NJ, Hong Kong, 1999.

[6] N.U. Ahmed, *Elements of Finite Dimensional Systems and Control Theory*, Pitman Monographs and Surveys in Pure and Applied Mathematics 37, Longman Scientific and Technical, Harlow, UK; John Wiley, New York, 1988.

[7] J. Beran, R. Sherman, M.S. Taqqu, and W. Willinger, *Long-range dependence in variable-bit-rate video traffic*, IEEE Trans. Commun., 43 (1995), pp. 1566–1579.

[8] B. Tsybakov and N. Georganas, *Self-similar processes in communication networks*, IEEE Trans. Inform. Theory, 44 (1998), pp. 1713–1725.

[9] L. Coutin and L. Decreusefond, *Abstract nonlinear filtering theory in the presence of fractional Brownian motion*, Ann. Appl. Probab., 9 (1999), pp. 1058–1090.

[10] L. Decreusefond and A.S. Ustunel, *Stochastic analysis of the fractional Brownian motion*, Potential Anal., 10 (1999), pp. 177–214.

[11] M.L. Kleptsyna, M.L. Kloeden, P.E. Ahn, and V.V. Ahn, *Linear filtering with fractional Brownian motion*, Stochastic Anal. Appl., 16 (1998), pp. 907–914.

[12] I. Norros, I. Valkeila, and J. Vitramo, *An elementary approach to a Girsanov formula and other analytical results on fractional Brownian motions*, Bernoulli, 5 (1999), pp. 571–587.

[13] A. Le Breton, *Filtering and parameter estimation in a simple linear model driven by a fractional Brownian motion*, Statist. Probab. Lett., 38 (1998), pp. 263–274.

[14] G. Gripenberg and I. Norros, *On the prediction of fractional Brownian motion*, J. Appl. Probab., 33 (1996), pp. 400–410.

[15] J. A. Beran, *Statistics for Long-Memory Processes,* Chapman and Hall, London, 1999.

[16] M.L. Kleptsyna, M.L. Breton, and M.C. Roubaud, *An Elementary Approach to Filtering in Systems with Fractional Brownian Observation Noise*, Rapport de Recherche N 3439, INRIA Le Chesnay, France, 1998.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 2, pp. 331–344

ON A CONSTRAINED DIRICHLET PROBLEM*

ARRIGO CELLINA†

Abstract. We consider a Dirichlet minimum problem with a pointwise constraint on the gradient, i.e., $\|\nabla u(x)\| \leq 1$ a.e., or, equivalently, an unconstrained minimum problem with an extended-valued integrand. Since the subdifferential of this integrand is defined on the whole effective domain, the problem of the validity of the Euler–Lagrange equation (or, equivalently, of the Pontryagin maximum principle) for a solution w can be posed. To show that this equation is verified along a solution, the equivalence of the problem considered here and of a problem with obstacles is proved, and a generalization of Stampacchia's bounded slope condition result is presented.

Key words. Euler–Lagrange equation, Pontryagin maximum principle, constraint on the gradient

AMS subject classifications. 35B50, 49N60

PII. S0363012900380425

Introduction. The author has been interested in the following specific problem: consider a *constrained* Dirichlet problem consisting of minimizing $\int_{S_R} \frac{1}{2}\|\nabla u\|^2$ over those functions u satisfying the boundary condition $u|_{\partial S_R} = u^0|_{\partial S_R}$ and, in addition, the constraint $\|\nabla u(x)\| \leq 1$ a.e. in S_R. Here $S_R \subset \Re^2$ is the planar sector obtained as the intersection of the disk of radius R and of the first orthant. The boundary datum u^0 is defined to be $u^0(x,y) = 1 - \frac{1}{R}\sqrt{x^2+y^2}$. When $R < 1$, there are no Lipschitzian functions satisfying at once the boundary conditions and the constraint on the gradient, and so the problem is of interest for $R \geq 1$. A computation shows that the bounded slope condition of constant 1 is satisfied when $R \geq \sqrt{2}$. In this case, since the solution to the Dirichlet problem is regular, Stampacchia's original theorem [6] proves that the solution w to the unconstrained Dirichlet problem satisfies $\|\nabla w(x)\| \leq 1$, i.e., that the constrained Dirichlet problem is actually equivalent to the unconstrained problem; hence, in this case, the solution to the constrained problem inherits all the regularity of the standard unconstrained problem. The motivation for the present paper was the desire of understanding what happens for $1 \leq R < \sqrt{2}$. In particular, it is our purpose to show that, for all $R \geq 1$, the Euler–Lagrange equation holds for the solution to this problem. To be more precise, consider the following *unconstrained* equivalent formulation to the minimization problem presented above:

$$(DP) \qquad \text{minimize} \int_{S_R} \left(\frac{1}{2}\|\nabla u(x)\|^2\right)^\infty dx \quad \text{on } u - u^0 \in H_0^1(S_R),$$

where $(\frac{1}{2}\|\xi\|^2)^\infty$ denotes the map

$$\xi \to \begin{cases} \frac{1}{2}\|\xi\|^2 & \text{if } \|\xi\| \leq 1, \\ +\infty & \text{otherwise.} \end{cases}$$

The (extended-valued) function $(\frac{1}{2}\|\xi\|^2)^\infty$ is convex, lower semicontinuous, and coercive: solutions to this minimization problem exist. Although the integrand is not

*Received by the editors November 3, 2000; accepted for publication (in revised form) January 10, 2002; published electronically June 18, 2002.
http://www.siam.org/journals/sicon/41-2/38042.html
†Dipartimento di Matematica e Applicazioni, Università di Milano Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy (cellina@matapp.unimib.it).

everywhere differentiable, its subdifferential is defined on the whole effective domain of the integrand itself: given one solution w, $\partial((\frac{1}{2}\|\nabla w(x)\|^2)^\infty)$ is defined a.e. We wish to prove the existence of an integrable function $p(.)$, a selection from the map $x \to \partial((\frac{1}{2}\|\nabla w(x)\|^2)^\infty)$, such that, for every $\phi \in H_0^1(S_R)$, one has

$$\int_{S_R} \langle p(x), \nabla\phi(x)\rangle \, dx = 0.$$

Equivalently, we wish to prove the existence of p, a weak solution to the system of equations

$$p(x) \in \partial\left(\left(\frac{1}{2}\|\nabla w(x)\|^2\right)^\infty\right), \quad \text{div } p(x) = 0.$$

Alternatively, from a control theory point of view, the above problem can be seen as the problem of minimizing

$$\int_{S_R} \left(\frac{1}{2}\|v\|^2\right) dx$$

for $u \in u^0 + W_0^\infty(S_R)$, subject to the Hamilton–Jacobi control equation

$$\nabla u(x) = v, \qquad v \in B,$$

where B is the Euclidean unit ball of $\Re^2$. Denote by H the map $H(p,v) = -\frac{1}{2}\|v\|^2 + \langle p, v\rangle$, and notice that $H(p,v) = \max_{z\in B} H(p,z)$ iff $p \in \partial(\frac{1}{2}\|\nabla v\|^2)^\infty$. Hence proving the validity of the Euler–Lagrange equation in the sense specified above amounts to proving the existence of p satisfying the equations

$$\nabla u(x) = \nabla H_p(p(x), v(x)), \qquad \text{div} p(x) = -\frac{\partial H}{\partial u}$$

such that a.e.

$$H(p(x), v(x)) = \max_{z\in B} H(p(x), z),$$

i.e., to proving that the solution w to the optimum control problem satisfies Pontryagin's maximum principle. The same control formulation, for a different minimum problem, was already considered in [4].

Problems like the one we consider here, i.e., the problem of minimizing $\int_\Omega f(\nabla u(x))dx$ on a set $\mathbf{K}$ of functions instead of on a linear space, are not new in the literature: these and similar problems have been treated from the point of view of variational inequalities. In that context, one wishes to show that a solution w is such that

$$\int_\Omega \langle \nabla f(\nabla w(x)), \nabla\phi(x)\rangle \, dx \geq 0$$

whenever $\phi \in C_c^\infty(\Omega)$ is admissible, i.e., such that $w + \phi$ is in $\mathbf{K}$. This kind of condition is *not* equivalent to the answer we seek here. Consider, for instance, the very degenerate problem arising when the boundary conditions are such that there is only one function in the space $W^{1,1}$ satisfying at once the boundary condition and the constraint on the gradient, as in the case considered here when $R = 1$. Then the boundary datum itself is the solution to this problem. However, the relevant

information on this solution that we gain from the validity of a variational inequality is nil since the only admissible variation ϕ is $\phi = 0$. On the other hand, the Euler–Lagrange equation holds on a linear space of test functions ϕ, independent of whether or not the test functions are admissible, and this fact should provide useful information on the solution w even in a degenerate case like this one.

The interest for the special region and boundary condition we examine here arises from the following considerations: it is known that a key to the understanding of problems with a constraint on the gradient comes from proving the equivalence of the given problem with an obstacle problem, as in Brézis and Sibony [1] and in Treu and Vornicescu [5]. When the constraint is $\|\nabla u(x)\| \leq K$, one is lead to consider the Natural Obstacles: the lower obstacle $N_K^-(x) = \sup_{\xi\in\partial S_R} \quad (u^0(\xi) - K\|x-\xi\|)$ and the upper obstacle $N_K^+(x) = \inf_{\xi\in\partial S_R} \quad (u^0(\xi) + K\|x-\xi\|)$: simply, any function u satisfying at once the boundary conditions and the constraint $\|\nabla u(x)\| \leq K$ must lie between the lower and the upper Natural Obstacle function. However computations show that for the region S_R and the boundary condition u^0 described above, these Natural Obstacles are not only not smooth but fail to have those reasonable properties that one can hopefully use to prove regularity of the solutions. Hence this region and boundary conditions provide possibly the simplest non-trivial example that illustrates the needs of more precise techniques to handle the problem.

In order to define the Effective Obstacle functions, that we will use instead of the Natural Obstacles, let us recall the definition of the Bounded Slope Condition [6].

DEFINITION ($(BSC)_K$). *Let K be a positive real, Ω a bounded convex set. The boundary datum u^0 satisfies $(BSC)_K$ if for every $x^0 \in \partial\Omega$ there exist vectors $k^+ = k^+(x^0)$ and $k^- = k^-(x^0)$, $\|k^+\| \leq K, \|k^-\| \leq K$, such that for every $x \in \partial\Omega$, we have*

$$C_+ \qquad u^0(x) - u^0(x^0) \leq \langle k^+, x - x^0\rangle$$

and

$$C_- \qquad u^0(x) - u^0(x^0) \geq \langle k^-, x - x^0\rangle.$$

The next definition presents a lower and an upper obstacle for the solutions, different from the Natural Obstacles.

DEFINITION (effective obstacles). *Let $V_K^+ = \{x^0 \in \partial\Omega$: there exist no $k^+, \|k^+\| \leq K$ satisfying C_+ for every $x \in \partial\Omega\}$; let $V_K^- = \{x^0 \in \partial\Omega$: there exist no $k^-, \|k^-\| \leq K$ satisfying C_- for every $x \in \partial\Omega\}$. Let $\Phi_K^+(x)$ be identically $+\infty$ when $V_K^+ = \emptyset$ and $\inf_{\xi\in V_K^+}(u^0(\xi) + K\|x-\xi\|)$ otherwise. Let $\Phi_K^-(x)$ be identically $-\infty$ when $V_K^- = \emptyset$ and $\sup_{\xi\in V_K^-}(u^0(\xi) - K\|x-\xi\|)$ otherwise.*

Comparing the definition of the natural obstacles with the definition above, one can see that the consideration of a smaller set of points (V_K^+ and V_K^- instead of $\partial\Omega$) gives rise to a larger and smoother upper bound and to a smaller and smoother lower bound for the solution and hence to a lesser constrained problem. Consider the (trivial) one-dimensional problem, where Ω is the interval $(0,1)$ and $u^0(0) = a, u^0(1) = b$. In this case, either $|b-a| > K$ and no solution exists, or $|b-a| \leq K$: in this second case, V_K^+ and V_K^- are empty, and, from our theorems below, the constrained problem is equivalent to the unconstrained problem, as it has to be, since its solutions are affine functions with the absolute value of the slope $= |b-a| \leq K$. The purpose of section 1 is to prove, under general assumptions on the region of integration Ω and the integrand f, that the two minimum problems, the problem with the constraint on the gradient $\|\nabla u(x)\| \leq 1$ and the problem constrained by the effective obstacles

(a less constrained problem than the analogous problem with the natural obstacles), are actually equivalent; this fact, that we can indeed work with a less constrained problem, possessing smoother solutions, will be the key to our proof of the validity of the Euler–Lagrange equation for the constrained Dirichlet problem to be presented in section 2.

From another point of view, our Theorem 1 below can be seen as a generalization of Stampacchia's theorem on the bounded slope condition [6]: in fact, the peculiarity of Theorem 1 is that the obstacles in its statement disappear (i.e., become, respectively, $+\infty$ and $-\infty$) when the bounded slope condition is satisfied.

1. The equivalence of variational problems. Theorem 2 of section 2 will provide suitable regularity for solutions to the problem with obstacles, assuming regularity of the obstacles. To check that the required assumptions are satisfied by the specific problem on S_R we have in mind, let us first explicitly discuss the effective obstacles for this problem: when $R \geq \sqrt{2}, \Phi_1^+(x,y) = +\infty$ and $\Phi_1^-(x,y) = -\infty$; for $1 \leq R < \sqrt{2}$, after some computations we obtain

$$\Phi_1^-(x,y) = \begin{cases} 1 - \frac{1}{R}y - \frac{\sqrt{R^2-1}}{R}x & \text{on } S_R^3 = S_R \cap \{y \geq x\frac{1}{\sqrt{R^2-1}}\}, \\ 1 - \sqrt{x^2+y^2} & \text{on } S_R^2 = S_R \cap \{\sqrt{R^2-1}\,x \leq y \leq x\frac{1}{\sqrt{R^2-1}}\}, \\ 1 - \frac{1}{R}x - \frac{\sqrt{R^2-1}}{R}y & \text{on } S_R^1 = S_R \cap \{0 \leq y \leq \sqrt{R^2-1}\,x\}. \end{cases}$$

Hence

$$\nabla\Phi_1^-(x,y) = \begin{cases} \left(-\frac{\sqrt{R^2-1}}{R}, -\frac{1}{R}\right) & \text{on } S_R^3, \\ \left(-\frac{x}{\sqrt{x^2+y^2}}, -\frac{y}{\sqrt{x^2+y^2}}\right) & \text{on } S_R^2, \\ \left(-\frac{1}{R}, -\frac{\sqrt{R^2-1}}{R}\right) & \text{on } S_R^1, \end{cases}$$

and

$$\Delta\Phi_1^-(x,y) = \begin{cases} 0 & \text{on } S_R^3, \\ -\frac{1}{\sqrt{x^2+y^2}} & \text{on } S_R^2, \\ 0 & \text{on } S_R^1. \end{cases}$$

Moreover,

$$\Phi_1^+(x,y) = +\infty.$$

Hence, in this example, when finite, the obstacle is smooth: the map $x, y \to \nabla\Phi_1^-(x,y)$ is locally Lipschitzian.

Let us also consider, for this special problem, a change of variables τ that will be of use in the proof of the validity of the Euler–Lagrange equation. Let θ^* be such that $\cos\theta^* = \frac{1}{R}$, $\sin\theta^* = \frac{\sqrt{R^2-1}}{R}$; let $\theta^{**} = \frac{\pi}{2} - \theta^*$ and $\theta^{***} = \theta^{**} - \theta^*$. Consider the regions $\Sigma^1 = \{(\xi,\eta) : \xi^2+\eta^2 < R^2; -\tan\theta^*\xi \leq \eta \leq 0\}$; $\Sigma^2 = \{(\xi,\eta) : 0 < \xi < R; 0 \leq \eta \leq \theta^{***}\}$; $\Sigma^3 = \{(\xi,\eta) : \xi^2+\eta^2 < R^2; \theta^{***} \leq \eta \leq \theta^{***} + \tan\theta^{**}\xi\}$. Set $\Sigma_R = \Sigma^1 \cup \Sigma^2 \cup \Sigma^3$,

and define the piecewise smooth transformation $\tau : \Sigma_R \subset \Re^2 \to S_R$ by

$$\begin{pmatrix} x \\ y \end{pmatrix} = \tau(\xi,\eta) = \begin{cases} \begin{pmatrix} \cos\theta^{**} & -\sin\theta^{**} \\ \sin\theta^{**} & \cos\theta^{**} \end{pmatrix} \begin{pmatrix} \xi \\ \eta - \theta^{***} \end{pmatrix} \text{ on } \Sigma^3, \\ \begin{pmatrix} \xi\cos(\eta+\theta^*) \\ \xi\sin(\eta+\theta^*) \end{pmatrix} \text{ on } \Sigma^2, \\ \begin{pmatrix} \cos\theta^* & -\sin\theta^* \\ \sin\theta^* & \cos\theta^* \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix} \text{ on } \Sigma^1. \end{cases}$$

One can verify that $\Phi_1^-(\tau(\xi,\eta))$ is (the restriction to Σ_R of) the affine function $1-\xi$ so that

$$\frac{\partial}{\partial\xi}\Phi_1^-(\tau(\xi,\eta)) \equiv -1; \quad \frac{\partial}{\partial\eta}\Phi_1^-(\tau(\xi,\eta)) \equiv 0.$$

The properties of τ that will be of later use are: (i) τ is orthogonal, i.e. $x_\xi x_\eta + y_\xi y_\eta = 0$, (ii) $\frac{\partial\tau}{\partial\xi}(\xi,\eta) = -(\nabla\Phi_1^-)(\tau(\xi,\eta))$, and hence, in particular, $\|\frac{\partial\tau}{\partial\xi}\| = \|(x_\xi, y_\xi)\| \equiv 1$, and (iii) for every fixed η, the map $\xi \to x_\eta^2 + y_\eta^2$ is nondecreasing. The unit vector $\frac{1}{\sqrt{x_\eta^2+y_\eta^2}}(x_\eta, y_\eta)$ inherits the regularity of $\frac{\partial\tau}{\partial\xi}(\xi,\eta)$.

The following properties of the natural obstacle functions will be of use.

PROPOSITION 1. *When u^0 is continuous and $\Phi_K^- \not\equiv -\infty$, Φ_K^- is Lipschitzian of constant K; whenever $\nabla\Phi_K^-$ exists, we have $\|\nabla\Phi_K^-\| = K$ and a similar case for Φ_K^+.*

The following theorem, on the equivalence of variational problems, is also a generalization of Stampacchia's theorem on the bounded slope condition [6] and of its generalization in [3].

THEOREM 1. *Let Ω be open and bounded; let f be a (possibly extended-valued) strictly convex function. Let K be positive real, let u^0 be in $C(\overline{\Omega}) \cap W^{1,1}(\Omega)$, and let the effective obstacles $\Phi_K^-(x)$, $\Phi_K^+(x)$ be defined as above. Moreover, assume that w, a solution to*

$$(P_{\Phi_K}) \qquad \textit{minimize} \int_\Omega f(\nabla u(x))\,dx,\ u - u^0 \in W_0^{1,1}(\Omega), \Phi_K^-(x) \le u(x) \le \Phi_K^+(x),$$

is in $C(\Omega)$. Then w is Lipschitzian of Lipschitz constant K, and it is a solution to

$$(P)_K \qquad \textit{minimize} \int_\Omega f(\nabla u(x))\,dx,\ u - u^0 \in W_0^{1,1}(\Omega), \|\nabla u(x)\| \le K.$$

Notice that the result is false if we assume f to be convex but not strictly convex: consider the problem of minimizing

$$\int_{-1}^{1} f(|x'(t)|)dt, \quad x(-1) = x(1) = 0,$$

subject to the constraint $|x'(t)| \le \frac{1}{2}$, where $f : \Re^+ \to \Re^+$ is the indicator function of the interval $[0,1]$. Then the bounded slope condition (of constant $\frac{1}{2}$) is verified both at $t=-1$ and at $t=+1$; i.e., the effective obstacles are $+\infty$ and $-\infty$, but it is not true that all solutions to the (unconstrained) resulting problem of minimizing $\int_{-1}^{1} f(|x'(t)|)dt, x(-1) = x(1) = 0$, are Lipschitzian with Lipschitz constant $\frac{1}{2}$: it is enough to consider $x(t) = -1 - t$ for $t \le 0$ and $= -1 + t$ for $t > 0$.

The following corollary to Theorem 1 is Theorem 2 of [3].

COROLLARY 1. *Let Ω be open, convex and bounded; let f be a (possibly extended-valued) strictly convex function. Let K be positive real, let u^0 be in $C(\overline{\Omega}) \cap W^{1,1}(\Omega)$, and let it satisfy $(BSC)_K$. Let w, a solution to*

$$\text{minimize} \int_\Omega f(\nabla u(x))\, dx,\ u - u^0 \in W_0^{1,1}(\Omega),$$

be in $C(\Omega)$. Then w is Lipschitzian of Lipschitz constant K, and it is a solution to

$$\text{minimize} \int_\Omega f(\nabla u(x))\, dx,\ u - u^0 \in W_0^{1,1}(\Omega), \|\nabla u(x)\| \le K.$$

Proof of Theorem 1. (a) The proof follows the same steps as the proof of Theorem 2 of [3]. Any function u satisfying the constraints of problem (P_K) must satisfy those of problem (P_{Φ_K}); hence $\inf(P_K) \ge \inf(P_{\Phi_K})$. Let $I(w)$ be $\int_\Omega f(\nabla w(x))\, dx$; if we show that a continuous solution w to (P_{Φ_K}) satisfies $\|\nabla w(x)\| \le K$, we have $\inf(P_K) \le I(w) = \inf(P_{\Phi_K})$. It is enough to show that there cannot exist a unit vector $\mathbf{v}$, a scalar $M > K$, and a set $E \subset \Omega$ with $\mu(E) > 0$ such that, for x in E, the derivative of the map $t \to w(x + tv)$ at $t = 0$ exists, equals $\langle \nabla w(x), \mathbf{v} \rangle$, and is $> M$. Let us assume that $M, \mathbf{v}, E$ exist. As in [3], the key consists in proving the following claim.

Claim. Let x^{**} be a point in Ω such that $t \to w(x^{**} + t\mathbf{v})$ is differentiable in $t = 0$ with derivative $D^{**} > M$, and let $h^{**} > 0$ be such that, for every $0 < h < h^{**}$, $x^{**} - h\mathbf{v} \in \Omega$, and

$$w(x^{**} + h\mathbf{v}) - w(x^{**}) - Mh > 0.$$

Then w is affine on $[x^{**}, x^{**} + h^{**}\mathbf{v}]$ with derivative D^{**}.

Proof of the claim. Fix $h > 0$. Consider $\Omega_h = \{\Omega \cap (\Omega - h\mathbf{v})\}$ and the set $E_h^+ = \{x \in \Omega_h; w(x + h\mathbf{v}) > w(x) + hM\}$. E_h^+ is open and nonempty since it contains x^{**}. For $x \in E_h^+$, we have that $y = x + h\mathbf{v}$ is such that $y - h\mathbf{v}$ is in Ω and $w(y - h\mathbf{v}) < w(y) - hM$. The set $E_h^- = \{y \in \Omega_{-h}; w(y - h\mathbf{v}) < w(y) - hM\}$ is a translation of E_h^+: $E_h^- - h\mathbf{v} = E_h^+$, and it contains $x^{**} + h\mathbf{v}$. The proof is based on taking variations $\phi_h^+(x) = \max\{w(x + h\mathbf{v}) - w(x) - hM, 0\}$ on Ω_h and $\phi_h^-(x) = \min\{w(x - h\mathbf{v}) - w(x) + hM, 0\}$ on Ω_{-h}. However in the present situation, these variations need not yield admissible functions: for $\lambda > 0$, $w(x) + \lambda\phi_h^+(x) > w(x)$ on E_h^+, but every admissible function has to be bounded above by Φ_K^+ and analogously for $w(x) + \lambda\phi_h^-(x)$. Let us show (for ϕ_h^-) that this is not the case. Let x be in E_h^-. Since the map $\Phi_K^-(x)$ is Lipschitzian with constant K, one has

$$\Phi_K^-(x) - hK \le \Phi_K^-(x - h\mathbf{v}) \le w(x - h\mathbf{v}) < w(x) - hM = \Phi_K^-(x) - hM + (w - \Phi_K^-)(x);$$

i.e., $(w - \Phi_K^-)(x) > h(M - K)$. Again, since Φ_K^- and Φ_K^- are Lipschitzian, $|w|$ is uniformly bounded, and so is $|\phi_h^-|$. Let $\lambda^0 \ge 0$ be so small that $|\lambda^0 \phi_h^-| \le h(M - K)$. Then, for all $0 \le \lambda \le \lambda^0$, we have

$$\Phi_K^-(x) = w(x) - (w(x) - \Phi_K^-(x)) < w(x) - h(M - K) \le w(x) + \lambda\phi_h^-(x) \le w(x) \le \Phi_K^+(x);$$

i.e., the variation $\lambda\phi_h^-(x)$ is admissible. A similar case holds for $\lambda\phi_h^+(x)$. Since w is a minimum, we must have that, for all λ sufficiently small,

$$\int_\Omega f(\nabla w(x) + \lambda \nabla \phi_h^+(x))\, dx \ge \int_\Omega f(\nabla w(x))\, dx,$$

$$\int_\Omega f(\nabla w(x) + \lambda \nabla \phi_h^-(x))\, dx \geq \int_\Omega f(\nabla w(x))\, dx.$$

The above inequalities yield

$$\int_{E_h^+} \{f(\nabla w(x) + \lambda[\nabla w(x + h\mathbf{v}) - \nabla w(x)]) - f(\nabla w(x))\}\, dx \geq 0,$$

$$\int_{E_h^-} \{f(\nabla w(x) + \lambda[\nabla w(x - h\mathbf{v}) - \nabla w(x)]) - f(\nabla w(x))\}\, dx \geq 0.$$

Making the change of variables $y = x + h\mathbf{v}$ and adding the two inequalities, we obtain

$$\int_{E_h^+} \{f(\nabla w(x) + \lambda[\nabla w(x + h\mathbf{v}) - \nabla w(x)]) - f(\nabla w(x)) + f(\nabla w(x + h\mathbf{v})$$

$$-\lambda[\nabla w(x + h\mathbf{v}) - \nabla w(x)]) - f(\nabla w(x + h\mathbf{v}))\}\, dx \geq 0.$$

Reasoning as in [3], we obtain that, a.e. in E_h^+, $\nabla w(x) = \nabla w(x + h\mathbf{v})$. As a consequence, we have that there exists $\Lambda > 0$ such that the map $t \to w(x^* + t\mathbf{v})$ is affine on $[0, \Lambda]$ with derivative $M + \zeta$, $\zeta > 0$, and, in addition, that $x^* + \Lambda\mathbf{v}$ is in $\partial\Omega$. Set x^{**} to be $x^* + \Lambda\mathbf{v}$ (hence, $\Lambda = \|x^{**} - x^*\|$). We have obtained that $w(x^*) = w(x^{**} - \Lambda\mathbf{v}) = w(x^{**}) - \Lambda(M + \zeta)$.

When $x^{**} \in \partial\Omega \setminus V_K^-$, i.e., the bounded slope condition is verified at x^{**}, there exists $k^-, \|k^-\| \leq K$ such that, for every $x \in \partial\Omega$, we have $u^0(x) - u^0(x^{**}) \geq \langle k^-, x - x^{**}\rangle$, or, equivalently, for every $x \in \partial\Omega$, $u^0(x) \geq u^0(x^{**}) + \langle k^-, x - x^{**}\rangle$. Hence Theorem 3 of [3] applies with $\ell(x) = u^0(x^{**}) + \langle k^-, x - x^{**}\rangle$, and we infer that the solution w satisfies the inequality

$$w(x^*) \geq u^0(x^{**}) + \langle k^-, x^* - x^{**}\rangle \geq w(x^{**}) + \langle k^-, x^* - x^{**}\rangle \geq w(x^{**}) - K\Lambda,$$

contradicting $w(x^*) = w(x^{**}) - \Lambda(M + \zeta)$.

On the other hand, when x^{**} belongs to V_K^-, from the very definition of $\Phi_K^-(x^*)$, we have

$$w(x^*) \geq \Phi_K^-(x^*) \geq w(x^{**}) - K\Lambda,$$

again contradicting $w(x^*) = w(x^{**}) - \Lambda(M + \zeta)$.

The assumption that M, $\mathbf{v}$, and E exist leads to a contradiction. □

COROLLARY 2. *Under the same assumptions on f, Ω, and u^0, problems (P_K) and (P_{Φ_K}) coincide in the sense that they have the same solutions.*

2. Regularity of the solutions and the validity of the Euler–Lagrange equation. Our purpose is to apply the results of the previous section to prove the validity of the Euler–Lagrange equation to the minimization problem described in the introduction. To do so, we shall need further regularity of the solution. Notice that, for the specific example we have in mind, the lower obstacle function, as computed in section 1, is not in $W^{2,2}(S_R)$, but it is in $W^{2,2}(\Omega^*)$ for every $\Omega^* \subset\subset S_R$. Theorem 2, in particular, can be applied to the case $\Omega = S_R$ and $f = \Phi_1^-$; in this case, Ω^* is any smooth subset of S_R bounded away from $(0, 0)$.

We recall that $u^0 \geq f$ on $\partial\Omega$ in the sense of $W^{2,2}(\Omega)$ iff $(u^0 - f)^- \in W_0^{2,2}(\Omega)$.

THEOREM 2. *Let Ω be an open and bounded set with Lipschitzian boundary, let $\Omega^* \subset \Omega$ be open with smooth boundary, and let $f \in W^{1,2}(\Omega) \cap W^{2,2}(\Omega^*)$. Let $u^0 \in W^{1,2}(\Omega)$ be such that $u^0 \geq f$ on $\partial\Omega$ in the sense of $W^{1,2}(\Omega)$. Let w be a solution to problem (P):*

$$(P) \qquad \text{minimize} \int_\Omega \|\nabla u(x)\|^2 dx, \quad u(x) \geq f(x) \text{ a.e.}, \quad u - u^0 \in W^{1,1}(\Omega);$$

then w is in $W^{2,2}(\Omega^*)$.

Proof. (a) Since $(w^0 - f)^- = 0$ a.e. in Ω^*, $w^0 \geq f$ on Ω^* in the sense of $W^{1,2}(\Omega^*)$. Hence [7, p. 82] there exists a sequence of smooth maps, $\left(w_n^0\right)$, converging to w in $W^{1,2}(\Omega^*)$ such that $w_n^0 \geq f$ on $\partial\Omega^*$.

Consider the minimization problem:

$$(P^*) \qquad \text{minimize} \int_{\Omega^*} \|\nabla u(x)\|^2 dx, \quad u(x) \geq f(x) \text{ a.e.}, \quad u - w_n^0 \in W^{1,2}(\Omega^*),$$

and let w_n be a solution.

(b) Following the technique of Brézis–Stampacchia [2], we will approximate w_n by smoother maps. For positive ε, let w_n^ε be a solution to the following minimum problem:

$$(P_n) \qquad \text{minimize} \quad \int_{\Omega^*} \varepsilon \|\nabla v - \nabla f\|^2 + (v - w_n)^2 \quad \text{on } w_n^0 + W^{1,2}(\Omega^*).$$

Claim. $w_n^\varepsilon(x) \geq f(x)$ a.e. on Ω^*.

Proof of the claim. The functional to be minimized is convex and coercive; hence a solution w_n^ε in $W^{1,2}(\Omega^*)$ exists; from known results, it satisfies the following Euler–Lagrange equation: for every η in $W_0^{1,2}(\Omega^*)$,

$$\int_{\Omega^*} \varepsilon \langle \nabla w_n^\varepsilon - \nabla f, \nabla \eta \rangle + [w_n^\varepsilon - w_n]\eta = 0.$$

Let $\eta^- = (w_n^\varepsilon - f)^-$, and assume it is not zero a.e. Since

$$0 \geq (w_n^\varepsilon - f)^- \geq (w_n^\varepsilon - w_n^0)^- + (w_n^0 - f)^-,$$

we have $\eta^- \in W_0^{1,2}(\Omega^*)$, and, from the Euler–Lagrange equation, we obtain

$$\int_{\{w_n^\varepsilon < f\}} \varepsilon \langle \nabla w_n^\varepsilon - \nabla f, \nabla w_n^\varepsilon - \nabla f \rangle + (w_n^\varepsilon - w)(w_n^\varepsilon - f) = 0.$$

Since, on the set $\{w_n^\varepsilon < f\}$, we have $w_n \geq f > w_n^\varepsilon$, we obtain $(w_n - w_n^\varepsilon)(f - w_n^\varepsilon) > 0$ and hence a contradiction. Thus $w_n^\varepsilon(x) \geq f(x)$ a.e. on Ω^*.

Moreover, by the regularity results for linear equations, $w_n^\varepsilon \in W^{2,2}(\Omega^*)$, and w_n^ε is a solution of the equation $-\varepsilon\Delta w_n^\varepsilon + (w_n^\varepsilon - w_n) + \varepsilon\Delta f = 0$.

(c) Notice that, for every η in $W_0^{1,2}(\Omega^*)$ admissible, i.e., such that $w_n + \eta \geq f$, one has

$$\int_{\Omega^*} \langle \nabla w_n(x), \nabla \eta(x) \rangle dx \geq 0;$$

we have obtained in (b) that $\eta = w_n^\varepsilon - w_n$ is an admissible variation on Ω^*; hence

$$0 \leq \int_{\Omega^*} \langle \nabla w_n, \nabla \eta \rangle = \int_{\Omega^*} \langle \nabla w_n, \nabla w_n^\varepsilon - \nabla w_n \rangle \leq \int_{\Omega^*} \langle \nabla w_n^\varepsilon, \nabla w_n^\varepsilon - \nabla w_n \rangle.$$

From

$$\int_{\Omega^*} \langle \nabla w_n^\varepsilon, \nabla w_n^\varepsilon - \nabla w_n \rangle = \int_{\Omega^*} (-\Delta w_n^\varepsilon)(w_n^\varepsilon - w) = \varepsilon \int_{\Omega^*} (-\Delta w_n^\varepsilon)(\Delta w_n^\varepsilon - \Delta f),$$

we obtain

$$\varepsilon \int_{\Omega^*} |\Delta w_n^\varepsilon|^2 \le \varepsilon \left(\int_{\Omega^*} |\Delta w_n^\varepsilon|^2 \right)^{\frac{1}{2}} \left(\int_{\Omega^*} |\Delta f|^2 \right)^{\frac{1}{2}};$$

i.e.,

$$\left(\int_{\Omega^*} |\Delta w_n^\varepsilon|^2 \right)^{\frac{1}{2}} \le \left(\int_{\Omega^*} |\Delta f|^2 \right)^{\frac{1}{2}}.$$

As a consequence, from (b) we infer

$$\|w_n^\varepsilon - w_n\|_2 \le \varepsilon 2 \|\Delta f\|_2$$

and also

$$\int_{\Omega^*} -\Delta w_n^\varepsilon (w_n^\varepsilon - w_n) \to 0$$

as $\varepsilon \to 0$. Hence $\int_{\Omega^*} \nabla w_n^\varepsilon (\nabla w_n^\varepsilon - \nabla w_n)$ and $\int_{\Omega^*} \nabla w_n (\nabla w_n^\varepsilon - \nabla w_n) \to 0$ so that $\|\nabla w - \nabla w_n^\varepsilon\|_2 \to 0$ as $\varepsilon \to 0$.

(d) Fix any $\phi \in C_c^\infty(\Omega^*)$. Then

$$\left| \int_{\Omega^*} \langle \nabla w_n, \nabla \phi \rangle \right| = \lim_{\varepsilon \to 0} \left| \int_{\Omega^*} \langle \nabla w_n^\varepsilon, \nabla \phi \rangle \right| \le K \|\phi\|_2$$

(K independent of n) so that $\|\Delta w_n\|_2 \le K$. The sequence (w_n) is weakly precompact in $W^{1,2}(\Omega^*)$.

(e) A subsequence (still called (w_n)) converges weakly in $W^{1,2}(\Omega^*)$ to w^*. We claim that $w^* - u^0 \in W_0^{1,2}(\Omega^*)$. To begin with, we have

$$d((w_n - u^0), W_0^{1,2}(\Omega^*)) \le \|(w_n - u^0) - [(w_n - w_n^0) + (w - u^0)]\|_{1,2} = \|w_n^0 - w\|_{1,2} \to 0.$$

Let (y_n) be a sequence of convex combinations converging to w in $W^{1,2}(\Omega^*)$: more precisely, for every n, there are $\nu(n)$ indices $j_1(n), \dots, j_\nu(n)$ such that $j_1(n) < \cdots < j_\nu(n)$ and $\lim_{n\to\infty} j_1(n) = \infty$ and coefficients $\alpha_{j_1(n)}, \dots, \alpha_{j_\nu(n)}$ such that $\alpha_{j_1(n)} + \cdots + \alpha_{j_\nu(n)} = 1$ and $\alpha_{j_i(n)} \ge 0$ such that

$$y_n = \sum_{j_1(n),\dots,j_\nu(n)} \alpha_j w_j \to w \text{ as } n \to \infty.$$

Also, let $v_n \in W_0^{1,2}(\Omega^*)$ be such that $\|(w_n - u^0) - v_n\| \le 2d((w_n - u^0), W_0^{1,2}(\Omega^*))$. Then

$$d((y_n - u^0), W_0^{1,2}(\Omega^*)) \le \|(y_n - u^0) - v_n\| \le \sum_{j_1(n),\dots,j_{\nu(n)}(n)} \alpha_j \|(w_j - u^0) - v_n\|$$

$$\le 2 \sup_{j \in \{j_1(n),\dots,j_{\nu(n)}(n)\}} d(w_j - u^0, W_0^{1,2}(\Omega^*)) \to 0 \text{ as } n \to \infty$$

so that $w^* - u^0 \in W_0^{1,2}(\Omega^*)$.

(f) We claim that w^* is the restriction to Ω^* of w. Let $I(u)$ be $\int_{\Omega^*} \|\nabla u(x)\|^2 dx$; then, since the restriction of w to Ω^* is a minimum for P^*, $I(w) \le I(w^*)$. Assume there exists $\sigma > 0$ such that $I(w) \le I(w^*) - \sigma$. By the weak lower semicontinuity of the functional to be minimized, for all n sufficiently large, $I(w^*) \le I(w_n) + \frac{\sigma}{2}$. In problems (P_n), compute the integral on w_n^0 to obtain

$$I(w_n) \le I(w_n^0) \le I(w_n^0 - w) + I(w) \le I(w_n^0 - w) + I(w_n) - \frac{\sigma}{2},$$

which is a contradiction since $I(w_n^0 - w) \to 0$. So w^* is a solution to problem (P^*). By the strict convexity of the functional, the solutions are unique. So w^* is the restriction to Ω^* of w, and the whole sequence (w_n) converges weakly to w.

(g) Consider again y^n. Fix any $\phi \in C_c^\infty(\Omega^*)$. Then

$$\left| \int \langle \nabla w, \nabla \phi \rangle \right| = \lim_{n\to\infty} \left| \int \langle \nabla y^n, \nabla \phi \rangle \right| \le K \|\phi\|_2.$$

Hence $w \in W^{2,2}(\Omega^*)$. □

COROLLARY 3. *Let $u^0(x,y) = 1 - \frac{1}{R}\sqrt{x^2+y^2}$; let $\Phi_1^-(z)$ be the natural obstacle for problem (DP) as computed in section 1. Then problems*

$$\text{minimize} \int_{S_R} \frac{1}{2} \|\nabla u(z)\|^2 dz : \quad u - u^0 \in W^{1,1}(S_R), \quad \Phi_1^-(z) \le u(z)$$

and

$$\text{minimize} \int_{S_R} \frac{1}{2} \|\nabla u(z)\|^2 dz : \quad u - u^0 \in W^{1,1}(S_R), \quad \|\nabla u(z)\| \le 1$$

are equivalent in the sense that they have the same solutions.

Proof. Solutions to the obstacle problem are continuous on S_R by the previous theorem. □

The following theorem presents the construction of p.

THEOREM 3. *Let $\Omega \subset \Re$ be a bounded convex region. Let w be a solution to the constrained Dirichlet problem (DP) on S. Let $K = 1$ and $\Phi_1^+(x)$ and $\Phi_1^-(x)$ be the effective obstacles. Assume that $\Phi_1^+(x) = \infty$ and that $\Phi_1^- \in W^{2,2}(\Omega^*)$ on any $\Omega^* \subset\subset \Omega$. Assume that there exists a Lipschitzian one-to-one transformation $\tau : \Sigma \to \Omega$ such that*

(i) *τ is orthogonal;*

(ii) *$\frac{\partial \tau}{\partial \xi}(\xi,\eta) = -(\nabla \Phi_1^-)(\tau(\xi,\eta))$;*

(iii) *for every η, the map $\xi \to x_\eta^2 + y_\eta^2$ is nondecreasing;*

(iv) *τ is piecewise C^2 on Σ.*

Then there exists $p(.)$, a selection from $z \to \partial((\frac{1}{2}\|\nabla w(z)\|^2)^\infty)$, such that, for every $\phi \in W_0^{1,2}(\Omega)$, one has

$$\int_\Omega \langle p(z), \nabla \phi(z) \rangle \, dz = 0.$$

Proof. (a) Apply the previous theorem to infer that w is continuous on Ω. Let $E = \{z \in \Omega : w(z) = \Phi_1^-(z)\}$: E is closed (in Ω). We notice that, on the open set $\Omega \setminus E$, w is a solution to the (unconstrained) minimization problem

$$\text{minimize} \int_{\Omega\setminus E} \|\nabla u(z)\|^2 \, dz;$$

hence w is harmonic in $\Omega \setminus E$.

(b) Consider the transformation τ, and set $\mathbf{\Phi}(\xi,\eta) = \Phi_1^-(\tau(\xi,\eta))$. We have

$$\mathbf{\Phi}_\xi = \Phi_1^-{}_x x_\xi + \Phi_1^-{}_y y_\xi = -\|\nabla\Phi_1^-\|^2 = -1, \qquad \mathbf{\Phi}_\eta = 0.$$

Set $\mathbf{E} = \tau^{-1}(E), \omega(\xi,\eta) = w(\tau(\xi,\eta))$. On $\mathbf{E}$, $\nabla\omega = \nabla\mathbf{\Phi} = (-1,0)$. From the regularity of w and hence of ω, on a.e. line $\eta = \text{const}$, the map $\xi \to \nabla\omega(\xi,\eta)$ is absolutely continuous. Let $\eta = \eta^*$ be one such line. Let (ξ^*,η^*) be in $\mathbf{E}$, so that $\omega(\xi^*,\eta^*) = \mathbf{\Phi}(\xi^*,\eta^*)$. Then, when $0 < \xi \le \xi^*$, $(\xi,\eta^*) \in \mathbf{E}$. In fact, assume there exists ξ^{**} such that $\omega(\xi^{**},\eta^*) > \mathbf{\Phi}(\xi^{**},\eta^*)$. Since $\Phi_\xi \equiv 1$ and $\omega_\xi = w_x x_\xi + w_y y_\xi$ with $\|(x_\xi, y_\xi)\| = 1$, on $\{\eta = \eta^*\} \cap \mathbf{E}$ we would have points where the gradient of w is in norm larger than 1. Hence there exist $\xi^E = \xi^E(\eta^*)$ such that $\{\eta = \eta^*\} \cap \mathbf{E}$ equals either $\{(\xi,\eta^*) : 0 < \xi \le \xi^E\}$ or $\{(\xi,\eta^*) : 0 < \xi < \xi^E\}$ according to whether ξ^E belongs to $\tau^{-1}(\Omega)$ or to $\partial(\tau^{-1}(\Omega))$.

(c) We wish to show that there exists $\alpha(z)$, $\alpha = 1$, when $\|\nabla w(z)\| < 1, \alpha \ge 1$, when $\|\nabla w(z)\| = 1$ such that, for every $\phi \in C_c^\infty(\Omega)$, one has

$$\int_\Omega \alpha(z)\langle\nabla w(z), \nabla\phi(z)\rangle\, dz = 0.$$

If this is the case, the map $\alpha(.)\nabla w(.)$ is $p(.)$, the required selection from the map $z \to \partial((\frac{1}{2}\|\nabla w(z)\|^2)^\infty)$. The computation will be performed in the variables (ξ,η). Let $\Xi(\eta) = \sqrt{(x_\eta^2 + y_\eta^2)}(\xi^E(\eta),\eta)$, and define $A(\xi,\eta) = \frac{1}{\sqrt{(x_\eta^2+y_\eta^2)(\xi,\eta)}}\Xi(\eta)$ for $\xi \le \xi^E(\eta)$, i.e., for (ξ,η) in E, and $= 1$ otherwise. Then, by assumption (iii), $(\xi,\eta) \in E$ implies $A(\xi,\eta) \ge 1$. We will show that

$$\int_{\tau^{-1}(\Omega)} A(\xi,\eta)[\langle\nabla w, \nabla\phi\rangle(\tau(\xi,\eta))]J(\xi,\eta)\, d(\xi,\eta) = 0.$$

Then $\alpha(z) = A(\tau^{-1}(z))$ will be the sought-for function α.

Fix ϕ. Let $P(\xi,\eta) = \phi(\tau(\xi,\eta))$; let s_P be the interior of $\text{supp}(P)$. Recalling that τ is an orthogonal transformation and that $(x_\xi^2 + y_\xi^2) = 1$, we obtain

$$\langle\nabla w, \nabla\phi\rangle(\tau(\xi,\eta)) = \left[\omega_\xi P_\xi + \frac{\omega_\eta P_\eta}{(J)^2}\right].$$

Setting $B(\xi,\eta) = \Xi(\eta)$ for $\xi \ge \xi^E(\eta)$ and $= J(\xi,\eta)$ otherwise, the integral above becomes

$$\int_{s_P} B(\xi,\eta)\left[\omega_\xi P_\xi + \frac{\omega_\eta P_\eta}{(J)^2}\right] d(\xi,\eta).$$

We will compute separately

$$I_1 = \int_{s_P} B(\xi,\eta)[\omega_\xi P_\xi]\, d(\xi,\eta)$$

and

$$I_2 = \int_{s_P} B(\xi,\eta)\left[\omega_\eta P_\eta \frac{1}{(x_\eta^2 + y_\eta^2)}\right] d(\xi,\eta).$$

(d) From the regularity provided by Theorem 2, on a.e. line $\eta = \eta^*$, the maps $\xi \to \nabla\omega(\xi,\eta)$ and $\xi \to \sqrt{(x_\eta^2 + y_\eta^2)}\omega_\xi P(\xi,\eta)$ are absolutely continuous. Consider

$$I_1 = \int \left(\int_{\{\eta=\eta^*\}\cap \mathbf{E}} \Xi(\eta^*)[\omega_\xi P_\xi]\, d\xi \right) d\eta^*$$

$$+ \int \left(\int_{\{\eta=\eta^*\}\cap(\Sigma\setminus\mathbf{E})} \sqrt{(x_\eta^2 + y_\eta^2)}[\omega_\xi P_\xi]\, d\xi \right) d\eta^*.$$

On $\mathbf{E}$, $\omega_\xi = -1$; at ∂s_P, $P = 0$, and we obtain

$$I_1 = \int \left(\int_0^{\xi^E} \Xi(\eta^*)[-P_\xi]\, d\xi \right) d\eta^* + \int \left(\int_{\xi^E}^{\xi^\partial} \sqrt{(x_\eta^2 + y_\eta^2)}[\omega_\xi P_\xi]\, d\xi \right) d\eta^*$$

$$= \int \Xi(\eta^*)[-P(\xi^E(\eta^*),\eta^*)] d\eta^* + \int \left(\int_{\xi^E}^{\xi^\partial} \sqrt{(x_\eta^2 + y_\eta^2)}[\omega_\xi P_\xi]\, d\xi \right) d\eta^*.$$

To compute the second integral above, since

$$\left(\sqrt{(x_\eta^2 + y_\eta^2)}\omega_\xi P\right)(\xi^\partial,\eta^*) - \left(\sqrt{(x_\eta^2 + y_\eta^2)}\omega_\xi P\right)(\xi^E,\eta^*)$$

$$= \int_{\xi^E}^{\xi^\partial} \left(\sqrt{(x_\eta^2 + y_\eta^2)}\omega_\xi P_\xi + P\frac{d}{d\xi}\left(\sqrt{(x_\eta^2 + y_\eta^2)}\omega_\xi\right)\right) d\xi,$$

$\omega_\xi(\xi^E,\eta^*) = -1$ (the continuity of ω_ξ is used here), and $P(\xi^\partial,\eta^*) = 0$, we obtain

$$\int \left(\int_{\xi^E}^{\xi^\partial} \sqrt{(x_\eta^2 + y_\eta^2)}[\omega_\xi P_\xi]\, d\xi \right) d\eta^*$$

$$= \int \left(-\left(\sqrt{(x_\eta^2 + y_\eta^2)}P\right)(\xi^E,\eta^*) - \int_{\xi^E}^{\xi^\partial} P\frac{d}{d\xi}\left(\sqrt{(x_\eta^2 + y_\eta^2)}\omega_\xi\right) d\xi \right) d\eta^*.$$

Hence

$$I_1 = \int \left(-\int_{\xi^E}^{\xi^\partial} P\frac{d}{d\xi}\left(\sqrt{(x_\eta^2 + y_\eta^2)}\omega_\xi\right) d\xi \right) d\eta^*$$

$$= \int_{s_P\setminus E} -\frac{d}{d\xi}\left(\sqrt{(x_\eta^2 + y_\eta^2)}\omega_\xi\right) P\, d(\xi,\eta).$$

(e) Consider

$$I_2 = \int_{s_P} B(\xi,\eta)\left[\omega_\eta P_\eta \frac{1}{(x_\eta^2 + y_\eta^2)}\right] d(\xi,\eta).$$

For a.e. ξ, the map $\eta \to \omega_\eta \frac{1}{\sqrt{(x_\eta^2+y_\eta^2)}} P$ is absolutely continuous. Fix one such ξ^*; the intersection of the line $\xi = \xi^*$ with the open set $s_P \setminus \mathbf{E}$ is the union of (at most countably many) open intervals $((\xi^*, \alpha_i(\xi^*)), (\xi^*, \beta_i(\xi^*)))$, and we obtain

$$\int_{s_P} B(\xi,\eta) \left[\omega_\eta P_\eta \frac{1}{(x_\eta^2+y_\eta^2)}\right] d(\xi,\eta)$$

$$= \int \left(\int_{\Sigma \cap \{\xi = \xi^*\}} B(\xi,\eta) \left[\omega_\eta P_\eta \frac{1}{(x_\eta^2+y_\eta^2)}\right] d\eta \right) d\xi^*$$

$$= \int \left(\int_{\cup((\xi^*,\alpha_i(\xi^*)),(\xi^*,\beta_i(\xi^*)))} \left[\omega_\eta P_\eta \frac{1}{\sqrt{(x_\eta^2+y_\eta^2)}}\right] d\eta \right) d\xi^*$$

$$+ \int \left(\int_{\mathbf{E} \cap \{\xi = \xi^*\}} \Xi(\eta) \left[\omega_\eta P_\eta \frac{1}{(x_\eta^2+y_\eta^2)}\right] d\eta \right) d\xi^*.$$

The last integral is 0 since, on $\mathbf{E}$, $\omega_\eta = 0$. Consider one interval $((\xi^*, \alpha_i(\xi^*)), (\xi^*, \beta_i(\xi^*)))$: by the absolute continuity of $\eta \to \omega_\eta \frac{1}{\sqrt{(x_\eta^2+y_\eta^2)}} P$, we have

$$\omega_\eta \frac{1}{\sqrt{(x_\eta^2+y_\eta^2)}} P|_{\alpha_i}^{\beta_i} = \int_{\alpha_i}^{\beta_i} \left[P \frac{d}{d\eta} \left(\omega_\theta \frac{1}{\sqrt{(x_\eta^2+y_\eta^2)}} \right) + \left(\omega_\eta \frac{1}{\sqrt{(x_\eta^2+y_\eta^2)}} \right) P_\eta \right] d\eta.$$

The points $(\xi^*, \alpha_i(\xi^*))$ and $(\xi^*, \beta_i(\xi^*))$ are either at ∂s_P, so that $P = 0$, or on $\mathbf{E}$, and, in this case, by the regularity of $\omega_\eta \frac{1}{\sqrt{(x_\eta^2+y_\eta^2)}}$, we have $\omega_\eta \frac{1}{\sqrt{(x_\eta^2+y_\eta^2)}} = 0$. Hence

$$I_2 = -\sum_i \int \left(\int_{\alpha_i}^{\beta_i} \left[P \frac{d}{d\eta} \left(\omega_\eta \frac{1}{\sqrt{(x_\eta^2+y_\eta^2)}} \right) \right] d\eta \right) d\xi^*$$

$$= \int_{s_P \setminus \mathbf{E}} \left[\frac{d}{d\eta} \left(\omega_\eta \frac{1}{\sqrt{(x_\eta^2+y_\eta^2)}} \right) \right] P \, d(\xi,\eta).$$

Adding the results of (d) and of (e), we obtain

$$\int_\Omega \alpha(z) \langle \nabla w(z), \nabla \phi(z) \rangle \, dz = I_1 + I_2$$

$$= \int_{s_P \setminus \mathbf{E}} \left[\frac{d}{d\eta} \left(\omega_\eta \frac{1}{\sqrt{(x_\eta^2+y_\eta^2)}} \right) + \frac{d}{d\xi} \sqrt{(x_\eta^2+y_\eta^2)} \omega_\xi \right] P \, d(\xi,\eta)$$

$$= \int_{s_P \setminus \mathbf{E}} \left[\omega_{\eta\eta} \frac{1}{J^2} - \omega_\eta \frac{J_\eta}{J^3} + \omega_\xi \frac{J_\xi}{J} + \omega_{\xi\xi} \right] PJ \, d(\xi,\eta).$$

(f) Expressing the Laplacian of w at the point $\tau(\xi,\eta)$ in terms of ω, for a generic τ twice differentiable at the point (ξ,η), one obtains

$$J^2\Delta(w(\tau(\xi,\eta))) = \omega_{\xi\xi}(y_\eta^2+x_\eta^2)+\omega_{\eta\eta}(x_\xi^2+y_\xi^2)-2\omega_{\xi\eta}(x_\xi x_\eta+y_\xi y_\eta)$$

$$+\omega_\xi\left(x_\eta x_{\eta\xi}-x_\xi x_{\eta\eta}+y_\eta y_{\eta\xi}-y_\xi y_{\eta\eta}-x_\eta^2\frac{J_\xi}{J}-y_\eta^2\frac{J_\xi}{J}+x_\eta x_\xi\frac{J_\eta}{J}+y_\eta y_\xi\frac{J_\eta}{J}\right)$$

$$+\omega_\eta\left(x_\xi x_{\xi\eta}+y_\xi y_{\xi\eta}-x_\eta x_{\xi\xi}-y_\eta y_{\xi\xi}-x_\xi^2\frac{J_\eta}{J}-y_\xi^2\frac{J_\eta}{J}+x_\xi x_\eta\frac{J_\xi}{J}+y_\xi y_\eta\frac{J_\xi}{J}\right).$$

This equality, for a transformation τ with properties (i) and (ii), becomes

$$J^2\Delta(w(\tau(\xi,\eta))) = \omega_{\xi\xi}(y_\eta^2+x_\eta^2)+\omega_{\eta\eta}+\omega_\xi(JJ_\xi)+\omega_\eta\left(-\frac{J_\eta}{J}\right).$$

Since w is harmonic on $\Omega\setminus E$, its Laplacian computed at $\tau(\xi,\eta)$ for every $(\xi,\eta)\in\Omega\setminus E$ is 0; by the change of variables formula, one obtains then that the integral in (e) is zero.

This ends the proof. □

Remark. In the case when $\Omega = S_R$ with $R=1$ and $u^0 = 1-\sqrt{x^2+y^2}$, the previous construction yields an explicit definition of α: in fact, in this case, we have that $(\xi^E(\eta),\eta)=(1,\eta)$, $\Xi(\eta)=1$, and $A(\xi,\eta)=\frac{1}{\sqrt{(x_\eta^2+y_\eta^2)(\xi,\eta)}}\Xi(\eta)=\frac{1}{\xi}$. Hence we obtain that, setting $z=(x,y)$, $\alpha(z)\nabla w(z)=\frac{1}{\sqrt{x^2+y^2}}(\frac{x}{\sqrt{x^2+y^2}},\frac{y}{\sqrt{x^2+y^2}})$ is $p(z)$, the required selection from the map $z\to\partial((\frac{1}{2}\|\nabla w(z)\|^2)^\infty)$ having the property that, for any $\phi\in C_c^\infty(S_1)$,

$$\int_{S_1}\langle p(x),\nabla\phi(x)\rangle\,dx = 0.$$

In this case, p turns out to be the gradient of the harmonic function $\log\sqrt{x^2+y^2}$.

Summarizing the results of Corollary 2 and Theorem 3, we obtain the following corollary.

COROLLARY 4. *For every $R\geq 1$ and $u^0(x,y)=1-\frac{1}{R}\sqrt{x^2+y^2}$, a solution w to problem (DP) on S_R satisfies the Euler–Lagrange equation.*

REFERENCES

[1] H. BRÉZIS AND M. SIBONY, *Équivalence de deux inéquations variationnelles et applications*, Arch. Ration. Mech. Anal., 41 (1971), pp. 254–265.

[2] H. BRÉZIS AND G. STAMPACCHIA, *Sur la régularité de la solution d'inéquations elliptiques*, Bull. Soc. Math. France, 96 (1968), pp. 153–180.

[3] A. CELLINA, *On the bounded slope condition and the validity of the Euler–Lagrange equation*, SIAM J. Control Optim., 40 (2001), pp. 1270–1279.

[4] A. CELLINA AND S. PERROTTA, *On the validity of the maximum principle and of the Euler–Lagrange equation for a minimum problem depending on the gradient*, SIAM J. Control Optim., 36 (1998), pp. 1987–1998.

[5] G. TREU AND M. VORNICESCU, *On the equivalence of two variational problems*, Calc. Var. Partial Differential Equations, 11 (2000), pp. 307–319.

[6] G. STAMPACCHIA, *On some regular multiple integral problems in the calculus of variations*, Comm. Pure Appl. Math., 16 (1963), pp. 383–421.

[7] G. M. TROIANIELLO, *Elliptic Differential Equations and Obstacle Problems*, Plenum Press, New York, 1987.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 2, pp. 345–361

NOTIONS OF OBSERVABILITY FOR UNCERTAIN LINEAR SYSTEMS WITH STRUCTURED UNCERTAINTY*

IAN R. PETERSEN†

Abstract. This paper introduces a notion of observability for a class of uncertain linear systems with structured uncertainty. In the uncertain systems under consideration, the uncertainty is described by averaged integral quadratic constraints. In order to define a notion of observability for uncertain linear systems, the paper introduces a robust observability function which extends the usual definition of the observability Gramian to the case of uncertain systems. Using this observability function, a corresponding unobservable cone is defined, and an uncertain system is said to be robustly observable if this cone contains only the origin. The paper presents an algorithm for finding the robust observability function and corresponding unobservable cone. This algorithm involves solving a parameterized Riccati differential equation.

Key words. uncertain systems, observability, Riccati equation, integral quadratic constraints, structured uncertainty

AMS subject classifications. 93B07, 49N10, 93B35, 93B36

PII. S0363012900368077

1. Introduction.

An important aspect of control theory is the insight it gives the control engineer into the control system design problem at hand. Recent research in robust control theory has resulted in a number of powerful techniques for the synthesis of robust control systems; see, e.g., [19, 3, 9]. However, there remains a need for new results on robust control theory which give additional insight into a given robust control problem. For example, it is extremely helpful for the control system designer to know what factors are currently limiting the performance of the control system design and what might be done to the system in order to achieve an improved level of performance.

One approach to providing the control system designer with additional insight into the control problem under consideration is the current research into "fundamental limitations" in control systems; see, e.g., [15]. In this approach, bounds on achievable performance are given in terms of plant transfer function properties such as poles and zeros. Although this approach is very useful in many applications, it does not fit directly in the uncertain systems approach to robust control system design. Also, in multivariable control systems, it would be useful to have some insight into the performance limitations imposed by the structure of the uncertain dynamics. This fact has motivated us to look at extending the modern control theory notion of observability to the case of uncertain systems. Naturally, it would also be of interest to look at the dual notion of controllability for uncertain systems. However, this question is beyond the scope of the current paper.

The notion of observability is one of the fundamental properties of a linear system; see, e.g., [4]. Together with controllability, the notion of observability can be used to determine if a given linear system can be stabilized via feedback control. Also, it can be used to determine if the unobserved states of a system can be estimated via

*Received by the editors February 18, 2000; accepted for publication (in revised form) December 14, 2001; published electronically June 18, 2002. This work was supported by the Australian Research Council.

http://www.siam.org/journals/sicon/41-2/36807.html

†School of Electrical Engineering, Australian Defence Force Academy, Canberra, ACT, 2600, Australia (irp@ee.adfa.edu.au).

a state estimator. However, in considering practical control system and filter design problems, it was found that the notion of observability offered little real insight to the designer. To some extent, this fact can be traced to the lack of concern with robustness issues in the standard definition of observability. For example, the fact that a given plant model is controllable and observable gives the designer very little insight into whether a controller can be designed to achieve adequate performance in the face of model uncertainties. Also, a plant model which fails to be observable can usually be recognized as deficient from physical principles without the need for the concept of observability.

The fact that many of the limitations on achievable control system performance arise from the presence of model uncertainty leads us to think that the notion of observability may provide considerably more insight into a particular controller or state estimator design problem if extended to classes of uncertain systems. Thus the aim of this paper is to introduce a notion of observability for uncertain systems which will provide insight into the structure of uncertain systems and the limitations on achievable performance which arise.

The notion of observability for uncertain systems has previously been considered by a number of authors. In particular, the notion of observability considered in the papers [5, 6] is most closely related to the notion of observability considered here. In these papers, the notion of observability introduced is motivated by a certain set valued state estimation problem. In particular, it is assumed that the state of the uncertain system is initially completely unknown. Then the uncertain system is said to be observable if the set of possible initial states consistent with output measurements over a finite time interval is bounded. This definition of observability for uncertain systems fits in quite well with problems of set valued state estimation such as those considered in [6, 8]. However, it cannot be easily extended to the case of uncertain systems with structured uncertainty. Furthermore, it does not naturally lead to a notion extending the idea of the unobservable subspace, which arises in linear systems theory; see, e.g., [17].

The notion of observability introduced in this paper involves extending the definition of the observability Gramian to the case of uncertain systems; see also [2], where this notion is extended to the case of nonlinear systems. The observability Gramian can be defined in terms of the energy in the output signal resulting from a given initial condition. In our definition, we define the robust observability function as the minimum possible energy in the output signal for the given initial condition and for any admissible uncertainties. Thus, in our definition of observability, we think of the uncertainty as attempting to prevent the energy in the initial condition from being reflected in the output signal. This approach to defining a robust observability function for an uncertain system naturally leads to an extension of the notion of unobservable subspace. That is, we consider the set of all states for which the robust observability function is zero. This set is no longer a subspace but rather a cone in the state space. It is hoped that consideration of this cone will give new insight into the structure of uncertain systems.

As in the papers [5, 6], the uncertain systems considered in this paper will use an integral quadratic constraint (IQC) uncertainty description. However, in [5, 6], only one IQC is considered. Thus, in these papers, the uncertainty is unstructured. The notion of observability introduced in this paper allows for the case of structured uncertainty in a straightforward way, in particular, using the averaged IQC uncertainty description used in [12, 13]. This uncertainty description allows us to exploit a

certain S-procedure theorem in order to calculate the observability function in terms of a certain parameter dependent optimal control problem. A similar result could have been achieved using stochastic IQC uncertainty descriptions such as in [7, 16]. However, in this paper, we used the averaged IQC approach for the sake of simplicity.

2. Problem formulation. We consider the following time-varying uncertain system defined on the finite time interval $[0, T]$:

$$\begin{aligned}
\dot{x}(t) &= A(t)x(t) + \sum_{s=1}^{k} B_s(t)\xi_s(t),\\
y(t) &= C(t)x(t) + \sum_{s=1}^{k} D_s(t)\xi_s(t),\\
z_1(t) &= K_1(t)x(t),\\
z_2(t) &= K_2(t)x(t),\\
&\vdots\\
z_k(t) &= K_k(t)x(t),
\end{aligned} \tag{2.1}$$

where $x(t) \in \mathbf{R}^n$ is the *state*, $y(t) \in \mathbf{R}^l$ is the *measured output*, $z_1(t) \in \mathbf{R}^{h_1}$, $z_2(t) \in \mathbf{R}^{h_2}, \ldots, z_k(t) \in \mathbf{R}^{h_k}$ are the *uncertainty outputs*, $\xi_1(t) \in \mathbf{R}^{r_1}$, $\xi_2(t) \in \mathbf{R}^{r_2}, \ldots, \xi_k(t) \in \mathbf{R}^{r_k}$ are the *uncertainty inputs*, and $A(\cdot)$, $B_1(\cdot), \ldots, B_k(\cdot)$, $C(\cdot)$, $K_1(\cdot)$, $K_2(\cdot)$, $\ldots$, $K_k(\cdot)$ are bounded piecewise continuous matrix functions defined on $[0, T]$. The uncertainty inputs can be thought of as the outputs of uncertainty blocks as shown in Figure 2.1. Also, the uncertainty outputs can be thought of as the inputs to these uncertainty blocks. The bounds on the uncertainties are described below.

System uncertainty. The uncertainty inputs and outputs may be collected together into two vectors. That is, we define

$$\xi(t) \triangleq [\xi_1(t)'\ \xi_2(t)'\ \ldots\ \xi_k(t)'\ ]'$$

and

$$z(t) \triangleq [z_1(t)'\ z_2(t)'\ \ldots\ z_k(t)'\ ]'.$$

The uncertainty is required to satisfy a certain averaged IQC. That is, we consider finite collections of uncertainty inputs such that the following constraint is satisfied.

Averaged IQC. Let $d_1 > 0, d_2 > 0, \ldots, d_k > 0$ be given positive constants associated with the system (2.1). We will consider collections of uncertainty inputs $\mathcal{S} = \{\xi^1(\cdot), \xi^2(\cdot), \ldots \xi^q(\cdot)\}$. The number of elements q in any such collection is arbitrary. A collection of uncertainty functions of the form $\mathcal{S} = \{\xi^1(\cdot), \xi^2(\cdot), \ldots \xi^q(\cdot)\} \subset \mathbf{L}_2[0, T]$ is an *admissible uncertainty collection* for the system (2.1) if the following conditions hold: Given any $\xi^i(\cdot) \in \mathcal{S}$ and any corresponding solution $\{x^i(\cdot), \xi^i(\cdot), z^i(\cdot)\}$ to (2.1) defined on $[0, T]$, we have

$$\begin{aligned}
\frac{1}{q}\sum_{i=1}^{q}\int_0^T \left(\|\xi_1^i(t)\|^2 - \|z_1^i(t)\|^2\right) dt \le d_1,\\
\frac{1}{q}\sum_{i=1}^{q}\int_0^T \left(\|\xi_2^i(t)\|^2 - \|z_2^i(t)\|^2\right) dt \le d_2,
\end{aligned}$$

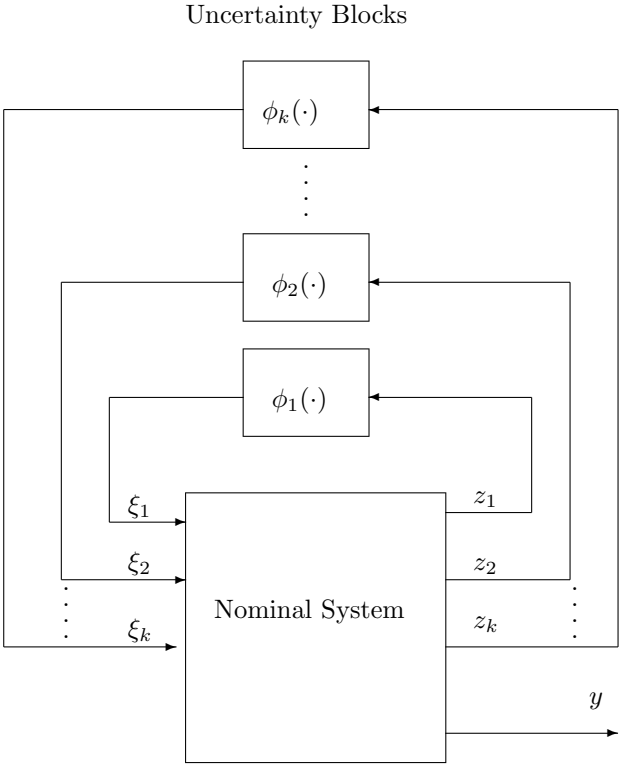


FIG. 2.1. *An uncertain system with structured feedback uncertainty.*

$$\vdots \qquad \vdots$$

$$\frac{1}{q}\sum_{i=1}^{q}\int_0^T \left(\|\xi_k^i(t)\|^2 - \|z_k^i(t)\|^2\right) dt \le d_k. \tag{2.2}$$

Here $\mathbf{L}_2[0,T]$ denotes the set of square integrable vector functions defined on the set $[0,T]$, and $\|\cdot\|$ denotes the standard Euclidean norm. The class of all such admissible uncertainty collections is denoted $\mathbf{\Xi}$. One way in which such uncertainty could be generated is via structured feedback uncertainty, as shown in the block diagram in Figure 2.1.

Remarks. The above definition extends the definition of the IQC given in [18, 10, 11]. In these papers, only individual uncertainty inputs are considered rather than collections of uncertainty inputs. One interpretation of the uncertainty class described above is a probabilistic one. That is, given any admissible uncertainty collection $\mathcal{S} \in \mathbf{\Xi}$, each uncertainty input $\xi^i(\cdot) \in \mathcal{S}$ is assigned an equal probability. Then condition (2.2) amounts to a bound on the expected value of the "measure of mismatch" between given uncertainty inputs $\xi_s(\cdot)$ and the following $\mathbf{L}_2[0,T]$ induced norm bound condition:

$$\int_0^T \left(\|\xi_s(t)\|^2 - \|z_s(t)\|^2\right) dt \le 0;$$

see also, e.g., [10] and [11]. It should be noted that the IQC on the uncertainty allows for nonlinear time-varying dynamic uncertainty. Indeed, the average measure

of mismatch bound d_s can be regarded as a bound on the average size of the initial condition for the uncertainty dynamics.

DEFINITION 2.1. *The* robust observability function *for the uncertain system* (2.1), (2.2) *is defined as*

$$L_o(x_0; d_1, \ldots, d_k; T) \triangleq \inf_{\mathcal{S}\in\Xi} \frac{1}{q} \sum_{i=1}^{q} \int_0^T \|y(t)\|^2 dt, \tag{2.3}$$

where $x(0) = x_0$ *in* (2.1).

This definition extends the standard definition of the observability Gramian for linear systems. In our robust observability function, we consider the worst case observability in which the uncertainty is trying to force the output of the system to zero for the given initial condition.

DEFINITION 2.2. *A state* $x_0 \in \mathbf{R}^n$ *is said to be* unobservable *for the uncertain system* (2.1), (2.2) *if*

$$L_o(x_0; d_1, \ldots, d_k; T) = 0$$

for all constants $d_1 > 0$, $d_2 > 0$,$\ldots d_k > 0$ *in* (2.2). *The set of all unobservable states for the uncertain system* (2.1), (2.2) *is referred to as the* unobservable cone $\mathcal{U}$*; i.e.,*

$$\mathcal{U} \triangleq \{x \in \mathbf{R}^n : L_o(x; d_1, \ldots, d_k; T) = 0 \ \ \forall d_1 > 0, d_2 > 0, \ldots, d_k > 0\}.$$

Also, the uncertain system (2.1), (2.2) *is said to be* robustly observable *if* $\mathcal{U} = \{0\}$*, i.e., if the origin is the only unobservable state.*

The above definition of the unobservable cone extends the standard definition of the unobservable subspace for a linear system (see [17]) to the case of uncertain systems. The aim of this paper is to find a means of constructing the observability function $L_o(x; d_1, \ldots, d_k; T)$ and the unobservable cone $\mathcal{U}$.

Note that the above definition will depend on the time horizon T. Roughly speaking, the observability of a state x_0 may depend on the time horizon T since, as the time horizon is lengthened, more information is obtained from the measurement y, but also more uncertainty may be allowed by the averaged IQC (2.2).

3. The main result. The robust observability function $L_o(x; d_1, \ldots, d_k; T)$ defined in (2.3) is defined as the value function for a constrained optimization problem. In order to solve this constrained optimization problem, we will use a version of the S-procedure theorem to convert the constrained optimization problem into an unconstrained optimization problem dependent on a set of Lagrange multiplier parameters.

3.1. The unconstrained optimization problem. For the uncertain system (2.1), (2.2), we define a function $V_\tau(x_0)$ as follows:

$$V_\tau(x_0) \triangleq \inf_{\xi(\cdot)\in\mathbf{L}_2[0,T]} \int_0^T \left(\|y(t)\|^2 + \sum_{s=1}^{k} \tau_s \|\xi_s(t)\|^2 - \sum_{s=1}^{k} \tau_s \|z_s(t)\|^2 \right) dt. \tag{3.1}$$

Here $\tau_1 \geq 0$, $\tau_2 \geq 0$,$\ldots$,$\tau_k \geq 0$ are given constants. For a given vector of Lagrange multiplier parameters $\tau = [\tau_1 \ \tau_2 \ \ldots \ \tau_k]$, the quantity $V_\tau(x_0)$ can be calculated as the solution to a standard linear quadratic optimal control problem. The solution to this optimal control problem, which will be given below, is given in terms of a Riccati differential equation dependent on the vector τ. Subsequently, we will show, using an

S-procedure theorem, that the robust observability function $L_o(x_0; d_1, \ldots, d_k; T)$ can be calculated in terms of $V_\tau(x_0)$.

In order to calculate $V_\tau(x_0)$, we first introduce some notation. Given $\tau = [\tau_1 \ \tau_2 \ \ldots \ \tau_k]$, let

$$\begin{aligned} B(t) &= [B_1(t) \ B_2(t) \ \ldots \ B_k(t)], \\ D(t) &= [D_1(t) \ D_2(t) \ \ldots \ D_k(t)], \\ K_\tau(t) &= \sum_{s=1}^{k} \tau_s K_s(t)' K_s(t), \\ \Lambda_\tau &= \begin{bmatrix} \tau_1 I & & 0 \\ & \ddots & \\ 0 & & \tau_k I \end{bmatrix}. \end{aligned}$$

Using this notation, it follows that the system (2.1) can be rewritten as

$$\dot{x}(t) = A(t)x(t) + B(t)\xi(t), \quad x(0) = x_0. \tag{3.2}$$

Also, the function $V_\tau(x_0)$ can be rewritten as

$$V_\tau(x_0) = \inf_{\xi(\cdot) \in \mathbf{L}_2[0,T]} J_\tau(\xi(\cdot)), \tag{3.3}$$

where

$$\begin{aligned} J_\tau(\xi(\cdot)) &= \int_0^T \begin{bmatrix} (C(t)x(t) + D(t)\xi(t))'(C(t)x(t) + D(t)\xi(t)) \\ +\xi(t)\Lambda_\tau\xi(t) - x(t)'K_\tau(t)x(t) \end{bmatrix} dt \\ &= \int_0^T \begin{pmatrix} x(t)[C(t)'C(t) - K_\tau(t)]x(t) \\ +2x(t)'C(t)'D(t)\xi(t) + \xi(t)'\Lambda_\tau\xi(t) + \xi(t)'D'D\xi(t) \end{pmatrix} dt. \end{aligned} \tag{3.4}$$

If $\tau = [\tau_1 \ \tau_2 \ \ldots \ \tau_k]$ is such that $\tau_1 > 0$, $\tau_2 > 0, \ldots, \tau_k > 0$, then the optimization problem (3.3) can be solved in terms of the following Riccati differential equation:

$$\begin{aligned} -\dot{P} &= A'P + PA - (PB + C'D)[\Lambda_\tau + D'D]^{-1}(D'C + B'P) + C'C - K_\tau, \\ P(T) &= 0. \end{aligned} \tag{3.5}$$

LEMMA 3.1. *Let $\tau = [\tau_1 \ \tau_2 \ \ldots \ \tau_k]$ be given such that $\tau_1 > 0$, $\tau_2 > 0, \ldots, \tau_k > 0$, and consider the corresponding system* (3.2) *and cost functional* (3.4). *Then the optimal control problem* (3.3) *is such that*

$$V_\tau(x_0) > -\infty$$

if and only if the Riccati differential equation (3.5) *has a solution $P_\tau(t)$ defined on $[0, T]$. In this case,*

$$V_\tau(x_0) = x_0' P_\tau(0) x_0. \tag{3.6}$$

Proof. This lemma follows directly from a standard result on the linear quadratic regulator problem; see, e.g., page 55 of [1]. ∎

3.2. An S-procedure result. In order to use the formula (3.6) to calculate the robust observability function (2.3), we will use the following S-procedure result. Indeed, consider a set of functionals

$$F_0(\xi(\cdot)),\ F_1(\xi(\cdot)),\dots,F_k(\xi(\cdot))$$

defined for the system (3.2).

LEMMA 3.2. *Suppose that, for any collection of input functions* $\{\xi^1(\cdot),\ \xi^2(\cdot),\dots,\xi^q(\cdot)\}$ *such that* $\xi^i(\cdot)\in\mathbf{L}_2[0,T]$ *for all* i *and*

$$\begin{aligned}
&\sum_{i=1}^{q} F_1(\xi^i(\cdot))\geq 0,\\
&\sum_{i=1}^{q} F_2(\xi^i(\cdot))\geq 0,\\
&\qquad\vdots\\
&\sum_{i=1}^{q} F_k(\xi^i(\cdot))\geq 0,
\end{aligned}\tag{3.7}$$

we have

$$\sum_{i=1}^{q} F_0(\xi^i(\cdot))\geq 0.\tag{3.8}$$

Then there exist constants $\tau_0\geq 0$, $\tau_1\geq 0$, $\dots$,$\tau_k\geq 0$ *such that* $\sum_{i=0}^{k}\tau_i>0$ *and*

$$\tau_0 F_0(\xi(\cdot))\geq\sum_{i=1}^{k}\tau_i F_k(\xi(\cdot))\tag{3.9}$$

for all inputs $\xi(\cdot)\in\mathbf{L}_2[0,T]$.

Proof. We first define the set

$$\mathcal{P}\triangleq\{[F_0(\xi(\cdot)),F_1(\xi(\cdot)),\dots,F_k(\xi(\cdot))]'\,:\,\xi(\cdot)\in\mathbf{L}_2[0,T]\}\subset\mathbf{R}^{k+1}.$$

Then condition (3.7), (3.8) implies that this set satisfies the assumptions of Theorem 3.1 of [14]. From this theorem, (3.9) follows. □

OBSERVATION 1. *If there exists an input* $\xi(\cdot)\in\mathbf{L}_2[0,T]$ *such that* $F_1(\xi(\cdot))>0,\ F_2(\xi(\cdot))>0,\dots,F_k(\xi(\cdot))>0$, *and the assumptions of the above lemma hold, then* τ_0 *may be chosen as* $\tau_0=1$ *in* (3.9)*; see Observation* 3.1 *in* [14].

3.3. A formula for the robust observability function. In order to present our main result, which is a formula for the robust observability function $L_o(x;d_1,\dots,d_k;T)$, we first introduce the following notation:

$$\Gamma\triangleq\{\tau=[\tau_1\ \tau_2\ \dots\ \tau_k]:\tau_1>0\ \tau_2>0\ \dots\ \tau_k>0\text{ and }V_\tau(x_0)>-\infty\}.$$

Also,

$$\bar{\Gamma}\triangleq\{\tau=[\tau_1\ \tau_2\ \dots\ \tau_k]:\tau_1\geq 0\ \tau_2\geq 0\ \dots\ \tau_k\geq 0\text{ and }V_\tau(x_0)>-\infty\}.$$

THEOREM 3.3. *Consider the uncertain system* (2.1), (2.2) *and the corresponding robust observability function* (2.3). *Then, for any initial condition* $x(0) = x_0$,

$$L_o(x_0; d_1, \ldots, d_k; T) = \max_{\tau \in \bar{\Gamma}} \left\{ V_\tau(x_0) - \sum_{s=1}^{k} \tau_s d_s \right\}. \tag{3.10}$$

Proof. Given any admissible uncertainty input collection $\mathcal{S} \in \Xi$ for the uncertain system (2.1), (2.2) with initial condition $x(0) = x_0$ and vector $\tau \in \bar{\Gamma}$, we claim

$$\frac{1}{q} \sum_{i=1}^{q} \int_0^T \|y^i(t)\|^2 dt \geq V_\tau(x_0) - \sum_{s=1}^{k} \tau_s d_s. \tag{3.11}$$

To establish this claim, we first note that it follows from the definition of $V_\tau(x_0)$ (3.1) that

$$\int_0^T \left(\|y(t)\|^2 + \sum_{s=1}^{k} \tau_s \|\xi_s(t)\|^2 - \sum_{s=1}^{k} \tau_s \|z_s(t)\|^2 \right) dt \geq V_\tau(x_0)$$

for all $\xi(\cdot) \in \mathbf{L}_2[0, T]$. In particular, this inequality holds for every element in the given collection $\mathcal{S}$. Hence

$$\begin{aligned} &\frac{1}{q} \sum_{i=1}^{q} \int_0^T \left(\|y^i(t)\|^2 + \sum_{s=1}^{k} \tau_s \|\xi_s^i(t)\|^2 - \sum_{s=1}^{k} \tau_s \|z_s^i(t)\|^2 \right) dt \\ &\geq \frac{1}{q} \sum_{i=1}^{q} V_\tau(x_0) \\ &= V_\tau(x_0). \end{aligned} \tag{3.12}$$

However, $\mathcal{S} \in \Xi$ implies that (2.2) is satisfied, and hence from (3.12) we obtain

$$\frac{1}{q} \sum_{i=1}^{q} \int_0^T \|y^i(t)\|^2 dt + \sum_{s=1}^{k} \tau_s d_s \geq V_\tau(x_0).$$

Thus (3.11) holds.

Now, since (3.11) holds for any $\mathcal{S} \in \Xi$, we have

$$\inf_{\mathcal{S} \in \Xi} \frac{1}{q} \sum_{i=1}^{q} \int_0^T \|y^i(t)\|^2 dt \geq V_\tau(x_0) - \sum_{s=1}^{k} \tau_s d_s \tag{3.13}$$

for all $\tau \in \bar{\Gamma}$. We now claim that there exists a $\tau \in \bar{\Gamma}$ such that

$$\inf_{\mathcal{S} \in \Xi} \frac{1}{q} \sum_{i=1}^{q} \int_0^T \|y^i(t)\|^2 dt \leq V_\tau(x_0) - \sum_{s=1}^{k} \tau_s d_s. \tag{3.14}$$

To establish this claim, we let

$$c \triangleq \inf_{\mathcal{S} \in \Xi} \frac{1}{q} \sum_{i=1}^{q} \int_0^T \|y^i(t)\|^2 dt. \tag{3.15}$$

Also, we define the functionals in Lemma 3.1 as follows:

$$\begin{aligned}
F_0(\xi(\cdot)) &\triangleq \int_0^T \|y(t)\|^2 dt - c,\\
F_1(\xi(\cdot)) &\triangleq \int_0^T \left(\|z_1(t)\|^2 - \|\xi_1(t)\|^2\right) dt + d_1,\\
F_2(\xi(\cdot)) &\triangleq \int_0^T \left(\|z_2(t)\|^2 - \|\xi_2(t)\|^2\right) dt + d_2,\\
&\vdots\\
F_k(\xi(\cdot)) &\triangleq \int_0^T \left(\|z_k(t)\|^2 - \|\xi_k(t)\|^2\right) dt + d_k.
\end{aligned}$$

Now, for any uncertainty input collection $\mathcal{S}$ such that

$$\frac{1}{q}\sum_{i=1}^{q} F_1(\xi^i(\cdot)) \geq 0,\ \frac{1}{q}\sum_{i=1}^{q} F_2(\xi^i(\cdot)) \geq 0, \ldots, \frac{1}{q}\sum_{i=1}^{q} F_k(\xi^i(\cdot)) \geq 0,$$

the averaged IQCs (2.2) are satisfied, and hence $\mathcal{S} \in \Xi$. Then it follows from (3.15) that

$$\frac{1}{q}\sum_{i=1}^{q} F_0(\xi^i(\cdot)) \geq 0.$$

Thus the conditions of the S-procedure result, Lemma 3.2, are satisfied. Also, note that, since $d_1 > 0,\ d_2 > 0, \ldots, d_k > 0$, then $F_1(0) > 0,\ F_2(0) > 0, \ldots, F_k(0) > 0$. Thus it follows from Lemma 3.2 and Observation 1 that there exist constants $\tau_1 \geq 0$, $\tau_2 \geq 0, \ldots, \tau_k \geq 0$ such that

$$F_0(\xi(\cdot)) \geq \sum_{s=1}^{k} \tau_s F_s(\xi(\cdot))$$

for all $\xi(\cdot) \in \mathbf{L}_2[0,T]$. That is,

$$\int_0^T \|y(t)\|^2 dt - c \geq \sum_{s=1}^{k} \tau_s \left[\int_0^T \left(\|z_s(t)\|^2 - \|\xi_s(t)\|^2\right) dt + d_s\right]$$

for all $\xi(\cdot) \in \mathbf{L}_2[0,T]$. Hence

$$\inf_{\xi(\cdot)\in\mathbf{L}_2[0,T]} \int_0^T \left(\|y(t)\|^2 + \sum_{s=1}^{k} \tau_s \|\xi_s(t)\|^2 - \sum_{s=1}^{k} \tau_s \|z_s(t)\|^2\right) dt \geq c + \sum_{s=1}^{k} \tau_s d_s.$$

Then, using (3.1) and (3.15), we have

$$V_\tau(x_0) \geq \inf_{\mathcal{S}\in\Xi} \frac{1}{q}\sum_{i=1}^{q} \int_0^T \|y(t)\|^2 dt + \sum_{s=1}^{k} \tau_s d_s \geq 0.$$

That is, (3.14) is satisfied. Furthermore, since $V_\tau(x_0) \geq 0$, then $\tau = [\tau_1\ \tau_2\ \ldots\ \tau_k]' \in \bar{\Gamma}$. Combining (3.13) and (3.14) now leads to (3.10). This completes the proof of the theorem. □

COROLLARY 3.4. *Consider the uncertain system* (2.1), (2.2) *and the corresponding robust observability function* (2.3). *Then, for any initial condition* $x(0) = x_0$,

$$L_o(x_0; d_1, \ldots, d_k; T) = \sup_{\tau \in \Gamma} \left\{ x_0' P_\tau(0) x_0 - \sum_{s=1}^{k} \tau_s d_s \right\}. \tag{3.16}$$

Proof. It is straightforward to verify that

$$\max_{\tau \in \bar{\Gamma}} \left\{ V_\tau(x_0) - \sum_{s=1}^{k} \tau_s d_s \right\} = \sup_{\tau \in \Gamma} \left\{ V_\tau(x_0) - \sum_{s=1}^{k} \tau_s d_s \right\}.$$

Hence, using Lemma 3.1, (3.16) follows. □

COROLLARY 3.5. *Consider the uncertain system* (2.1), (2.2). *Then a state* $x_0 \in \mathbf{R}^n$ *is unobservable if and only if*

$$x_0' P_\tau(0) x_0 \leq 0$$

for all $\tau \in \Gamma$.

Proof. If x_0 is unobservable, then $L_o(x_0; d_1, \ldots, d_k; T) = 0$ for all $d_1 > 0,\ d_2 > 0, \ldots, d_k$. Hence, using Corollary 3.4,

$$x_0' P_\tau(0) x_0 - \sum_{s=1}^{k} \tau_s d_s \leq 0$$

for all $d_1 > 0,\ d_2 > 0, \ldots, d_k > 0$, and $\tau \in \Gamma$. Thus

$$x_0' P_\tau(0) x_0 \leq 0$$

for all $\tau \in \Gamma$.

Conversely, if

$$x_0' P_\tau(0) x_0 \leq 0$$

for all $\tau \in \Gamma$, then

$$x_0' P_\tau(0) x_0 - \sum_{s=1}^{k} \tau_s d_s \leq 0$$

for all $d_1 > 0,\ d_2 > 0, \ldots, d_k > 0$, and $\tau \in \Gamma$. Thus, using Corollary 3.4,

$$L_o(x_0; d_1, \ldots, d_k; T) = \sup_{\tau \in \Gamma} \left\{ x_0' P_\tau(0) x_0 - \sum_{s=1}^{k} \tau_s d_s \right\} \leq 0$$

for all $d_1 > 0,\ d_2 > 0, \ldots, d_k > 0$. However, it follows from the definition of the observability function that $L_o(x_0; d_1, \ldots, d_k; T) \geq 0$. Thus

$$L_o(x_0; d_1, \ldots, d_k; T) = 0$$

for all $d_1 > 0,\ d_2 > 0, \ldots, d_k$. That is, x_0 is unobservable. □

OBSERVATION 2. *From the above corollary, it follows immediately that the unobservable cone* $\mathcal{U}$ *can be written in the form*

$$\mathcal{U} = \{x \in \mathbf{R}^n : x_0' P_\tau(0) x_0 \leq 0 \ \ \forall \tau \in \Gamma\}.$$

Also, it follows that the uncertain system (2.1), (2.2) *is robustly observable if and only if, for all* $x_0 \in \mathbf{R}^n : x_0 \neq 0$, *there exists a* $\tau \in \Gamma$ *such that*

$$x_0' P_\tau(0) x_0 > 0.$$

4. An alternative definition of robust observability. In this section, we compare the notion of robust observability defined above with the notion of robust observability considered in the paper [6]. This paper considers a class of uncertain systems in which the uncertainty is described by a single IQC. Hence the uncertainty in the system is unstructured. Also, in [6] there was no need to use averaged IQCs.

The notion of observability considered in [6] relates to a certain set valued state estimation problem. The result of [6] shows that the uncertain system under consideration has the property of robust observability if and only if the solution to a certain Riccati differential equation is positive-definite at time zero. The main result of this section shows that the definition of robust observability given in [6] is, in fact, equivalent to our definition of robust observability given in section 2.

The uncertain system considered in [6] can be considered as a special case of the uncertain system (2.1), (2.2), where $k = 1$,

$$\begin{aligned} \xi_1(t) &= \begin{bmatrix} w(t) \\ v(t) \end{bmatrix}, \\ A(t) &= A(t); \quad B_1(t) = [B_1(t)\ 0], \\ K_1(t) &= K(t); \quad C(t) = C(t); \quad D_1(t) = [0\ I]. \end{aligned} \tag{4.1}$$

Also, the uncertainty is assumed to satisfy the IQC

$$\int_0^T \left(\|w(t)\|^2 + \|v(t)\|^2 - \|z(t)\|^2 \right) dt \leq d_1. \tag{4.2}$$

This IQC can be considered as a special case of the averaged IQCs (2.2) when we restrict our attention to uncertainty input collections with one element. Also, note that we have assumed that, in [6], $R(t) \equiv I$, $Q \equiv I$.

According to the definition given in [6], an uncertain system is robustly observable if the set of all possible states at time $t = 0$, consistent with the uncertain system model and given output measurements on $[0,\ T]$, is bounded (for all $d_1 > 0$). The main result of [6] on robust observability gives a characterization of robust observability in terms of the Riccati differential equation:

$$\begin{aligned} -\dot{Y}(t) = {} & Y(t)A(t) + A(t)'Y(t) - Y(t)B_1(t)B_1(t)'Y(t) \\ & -K(t)'K(t) + C(t)'C(t), \quad Y(T) = 0. \end{aligned} \tag{4.3}$$

Indeed, it is shown in [6] that the uncertain system under consideration is robustly observable (in the sense of [6]) if and only if (4.3) has a solution on $[0,\ T]$ and $Y(0) > 0$.

In order to compare our definition of robust observability with the definition given in [6], we first observe that the Riccati differential equation (4.3) can be given an optimal control interpretation. Indeed, if we consider the system (2.1), (4.1) with initial condition $x(0) = x_0$ and the Riccati differential equation (4.3), then

$$x_0'Y(0)x_0 = \inf_{w(\cdot)\in \mathbf{L}_2[0,T]} \int_0^T \left(\|C(t)x(t)\|^2 - \|z_1(t)\|^2 + \|w(t)\|^2 \right) dt. \tag{4.4}$$

Now consider the quantity (3.1) for the uncertain system (2.1), (4.1), (4.2). This quantity can be calculated as follows:

$$\begin{aligned} V_\tau(x_0) &= \inf_{\xi(\cdot)\in \mathbf{L}_2[0,T]} \int_0^T \left(\|y(t)\|^2 - \tau\|z(t)\|^2 + \tau\|w(t)\|^2 + \tau\|v(t)\|^2 \right) dt \\ &= \inf_{w(\cdot)\in \mathbf{L}_2[0,T]} \inf_{v(\cdot)\in \mathbf{L}_2[0,T]} \int_0^T \begin{pmatrix} \|C(t)x(t) + v(t)\|^2 - \tau\|z(t)\|^2 \\ +\tau\|w(t)\|^2 + \tau\|v(t)\|^2 \end{pmatrix} dt. \end{aligned}$$

However,

$$\begin{aligned}
&\inf_{v(\cdot)\in\mathbf{L}_2[0,T]} \int_0^T \left(x(t)'C(t)'C(t)x(t) + 2x(t)'C(t)v(t) + v(t)'v(t) + \tau\|v(t)\|^2\right) dt \\
&= \inf_{v(\cdot)\in\mathbf{L}_2[0,T]} \int_0^T \left(\begin{array}{l} [\frac{1}{\sqrt{1+\tau}}C(t)x(t) + \sqrt{1+\tau}v(t)]'[\frac{1}{\sqrt{1+\tau}}C(t)x(t) + \sqrt{1+\tau}v(t)] \\ -\frac{1}{1+\tau}x(t)'C(t)'C(t)x(t) + x(t)'C(t)'C(t)x(t) \end{array} \right) dt \\
&= \int_0^T \frac{\tau}{1+\tau}x(t)'C(t)'C(t)x(t)dt.
\end{aligned}$$

Thus

$$V_\tau(x_0) = \tau \inf_{w(\cdot)\in\mathbf{L}_2[0,T]} \int_0^T \left(\frac{1}{1+\tau}x(t)'C(t)'C(t)x(t) - \|z(t)\|^2 + \|w(t)\|^2\right) dt.$$

Now observe that, for any $\tau > 0$, $V_\tau(x_0) \leq 0$ if and only if $\bar{V}_\tau(x_0) \leq 0$, where

$$\bar{V}_\tau(x_0) = \inf_{w(\cdot)\in\mathbf{L}_2[0,T]} \int_0^T \left(\frac{1}{1+\tau}x(t)'C(t)'C(t)x(t) - \|z(t)\|^2 + \|w(t)\|^2\right) dt. \tag{4.5}$$

Also note that it follows from (4.5) that, for any $x_0 \in \mathbf{R}^n$, $\bar{V}_\tau(x_0)$ is monotone increasing as $\tau \to 0$. Hence it follows from Theorem 3.3 that a state $x_0 \in \mathbf{R}^n$ is unobservable for the uncertain system (2.1), (4.1), (4.2) if and only if

$$\inf_{w(\cdot)\in\mathbf{L}_2[0,T]} \int_0^T \left(x(t)'C(t)'C(t)x(t) - \|z(t)\|^2 + \|w(t)\|^2\right) dt \leq 0.$$

Therefore, it follows from (4.4) that the system has no nonzero unobservable state if and only if

$$Y(0) > 0.$$

That is, our robust observability condition is equivalent to the robust observability condition of [6] for uncertain systems of the form (2.1), (4.1), (4.2). However, in contrast to the robust observability condition of [6], our robust observability condition can be applied to systems with structured uncertainty. Also, our robust observability condition leads naturally to a notion of an unobservable cone.

5. Illustrative examples. In this section, we consider two examples which illustrate our notions of robust observability and the unobservable cone for an uncertain system.

Example 1. We consider an uncertain system of the form (2.1), (2.2), where $k = 1$, $T = 10$,

$$\begin{gathered}
A(t) \equiv \begin{bmatrix} -2 & 0 \\ 0 & -1 \end{bmatrix}, \quad B_1(t) \equiv \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \\
K_1(t) \equiv [1 \;\; 0], \quad C(t) \equiv [1 \;\; 1], \quad D_1(t) \equiv 0.
\end{gathered}$$

In order to characterize the unobservable cone for this uncertain system, we will apply Observation 2. This involves solving the Riccati differential equation (3.5) for different values of the parameter $\tau > 0$. For each value of $\tau > 0$, we form the set

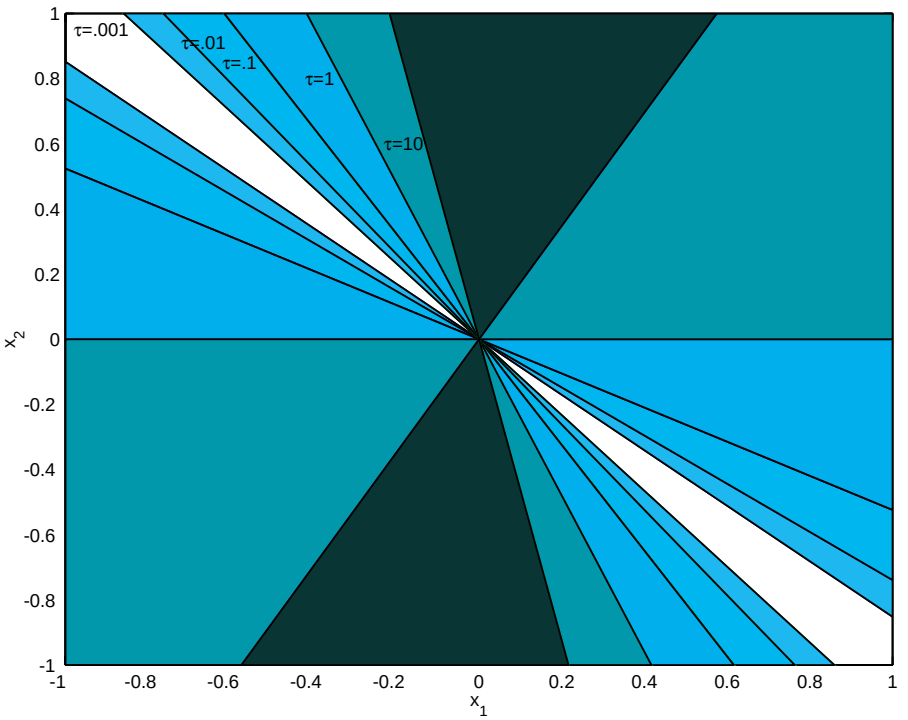


FIG. 5.1. *Unobservable cone for Example* 1.

$\{x \in \mathbf{R}^n : x_0' P_\tau(0) x_0 \leq 0\}$. Then the unobservable cone is the intersection of all of these sets. This process is illustrated in Figure 5.1.

In this figure, the unshaded region corresponds to the states such that $x_0' P_\tau(0) x_0 \leq 0$. As more and more values of τ are considered, the remaining unshaded region becomes smaller and smaller. Indeed, the unobservable cone in this example consists of a single line passing through the origin. This can be seen as follows: For each $\tau > 0$, the set $\{x \in \mathbf{R}^n : x_0' P_\tau(0) x_0 \leq 0\}$ is bounded by two lines which pass through the origin. Furthermore, it is straightforward to verify that, as $\tau \to 0$, both slopes converge to -1; see also Figure 5.2, in which we plot the slope of these lines as a function of τ. From this figure, we can conclude that, in this example, the unobservable cone is

$$\mathcal{U} = \{x = [x_1 \ x_2]' : x_2 = -x_1\}.$$

To further understand this result, note that, for this example, the averaged IQC (2.2) allows for norm bounded uncertainties of the form $w(t) = \delta(t) z(t)$, where $\delta(t)$ is a norm bounded uncertain parameter satisfying the bound $|\delta(t)| \leq 1$. In this case, the uncertain system can be rewritten as

$$\begin{aligned} \dot{x} &= \begin{bmatrix} -2 + 2\delta(t) & 0 \\ 0 & -1 \end{bmatrix} x, \\ y &= [1 \quad 1] x. \end{aligned}$$

In particular, if $\delta(t) \equiv 0.5$, this system becomes

$$\dot{x} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} x,$$

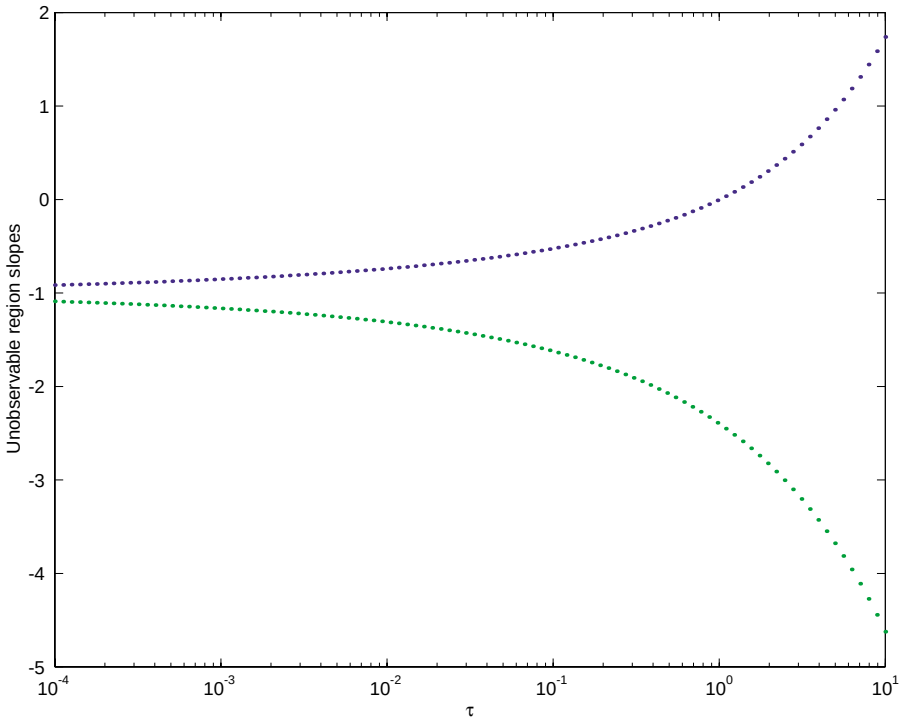


FIG. 5.2. *Bounding slopes vs* τ.

$$y = [1 \quad 1]x,$$

which is an unobservable linear system with unobservable subspace

$$\{x = [x_1 \ x_2]' : x_2 = -x_1\}.$$

This is the same as the unobservable cone for the uncertain system calculated above.

Note that, for this example, we have $D_1(t) \equiv 0$. That is, we have no uncertainty in the "C matrix." In this situation, it is straightforward to verify that the unobservable cone must lie in this null space of the matrix C. Hence, in a two dimensional system such as this one, any nontrivial unobservable cone must be a linear space. In the next example, we allow for uncertainty in the "C matrix" and show that this can lead to an unobservable cone which is not a linear space.

Example 2. We now consider another uncertain system of the form (2.1), (2.2), where $k = 1$, $T = 10$,

$$A(t) \equiv \begin{bmatrix} 0 & 1 \\ -1 & -5 \end{bmatrix}, \quad B_1(t) \equiv \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$
$$K_1(t) \equiv [0 \quad 0.7], \quad C(t) \equiv [1 \quad 0], \quad D_1(t) \equiv 1.$$

As above, we characterize the unobservable cone for this system using Observation 2. This leads to the results illustrated in Figure 5.3.

As in Example 1, the unshaded region in this figure corresponds to the states such that $x_0' P_\tau(0) x_0 \leq 0$ for various values of τ. As more and more values of τ are

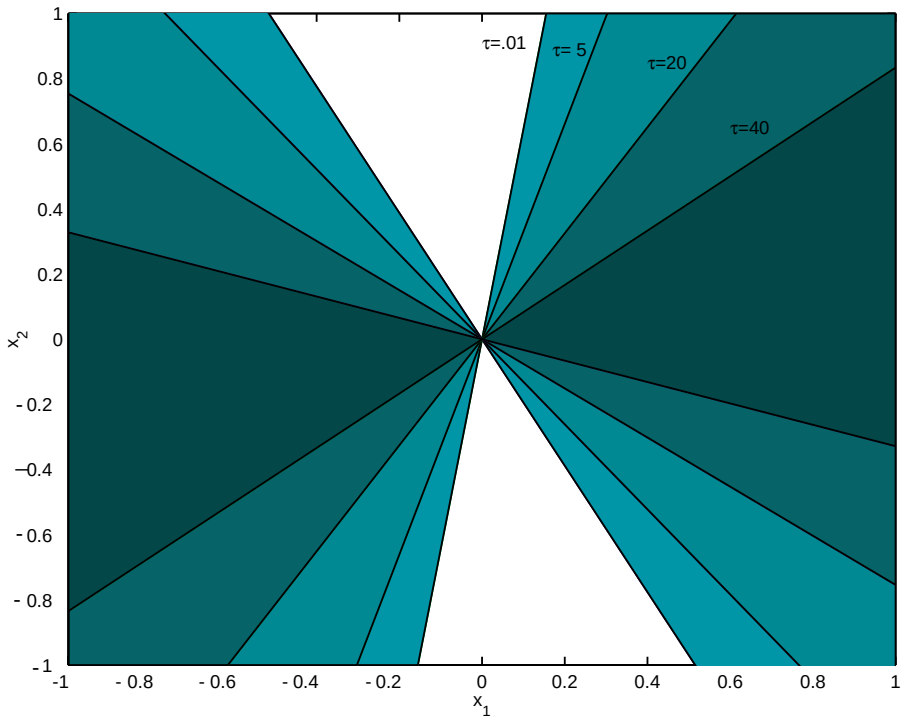


FIG. 5.3. *Unobservable cone for Example* 2.

considered, the remaining unshaded region becomes smaller and smaller. Also, as in Example 1, the regions $x_0' P_\tau(0) x_0 \leq 0$ are bounded by two straight lines passing through the origin. By considering the range of values of these slopes, we conclude that the unobservable cone for this uncertain system is a region in the state space bounded between the lines $\{x_2 = -1.94x_1\}$ and $\{x_2 = 6.47x_1\}$; see, e.g., Figure 5.4, in which we plot the slopes of these lines as a function of τ. Note that we consider only $\tau < 54.4$ since, for τ greater than this value, $P_\tau(0)$ is negative-definite.

For this uncertain system, the corresponding uncertain system with constant norm bounded uncertainty is described by the state equations

$$\dot{x} = \begin{bmatrix} 0 & 1 + 0.7\delta_1 \\ -1 & -5 \end{bmatrix} x,$$
$$y = [1 \quad 0.7\delta_2] x,$$

where $\delta_1^2 + \delta_2^2 \leq 1$. For this system, the corresponding observability matrix is given by

$$\begin{bmatrix} 1 & 0.7\delta_2 \\ -0.7\delta_2 & 1 + 0.7\delta_1 - 3.5\delta_2 \end{bmatrix}.$$

This matrix is nonsingular for some values of δ_1, δ_2 such that $\delta_1^2 + \delta_2^2 \leq 1$. This is consistent with the fact that the original uncertain system considered in this example has a nontrivial unobservable cone.

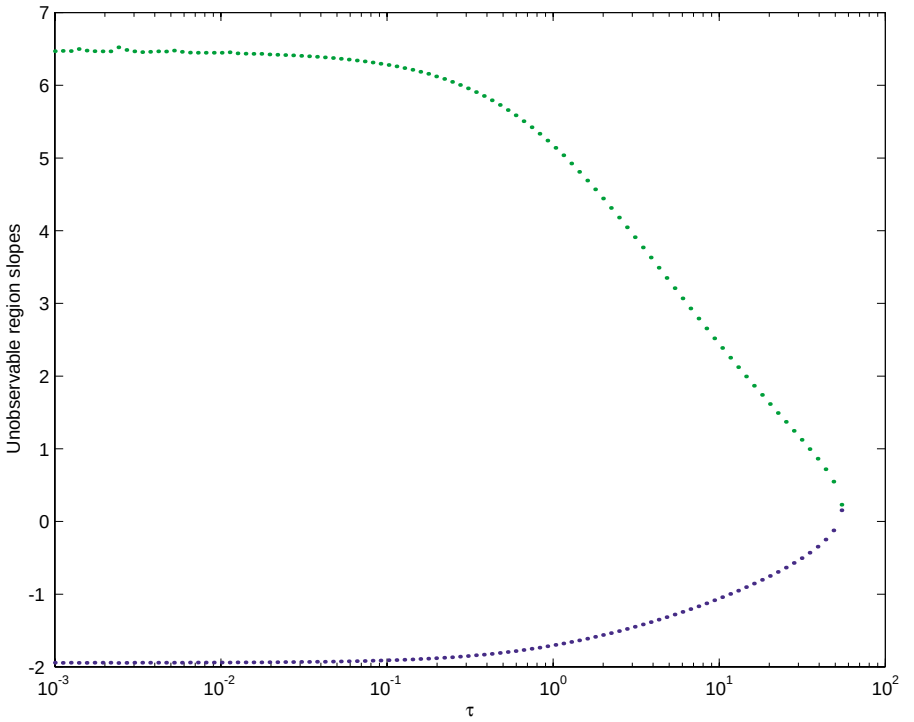


FIG. 5.4. *Bounding slopes vs* τ.

6. Conclusions. In this paper, we have introduced a new notion of robust observability for a class of uncertain systems. We also presented some results which show how this condition can be tested. A feature of this notion of robust observability is that it applies to uncertain systems with structured uncertainty. Furthermore, it also leads naturally to a notion of the unobservable cone for an uncertain system. The paper shows that the notion of robust observability introduced here is equivalent to an earlier notion of robust observability which was only applicable to uncertain systems with unstructured uncertainty. Also, the earlier notion of robust observability did not lead to a corresponding unobservable cone.

The paper presents two simple examples which illustrate the notion of robust observability and the calculation of the unobservable cone. These examples also illustrate the relation between the notion of robust observability and corresponding observability ideas for uncertain systems with constant norm bounded uncertainty.

One of the main areas of future research arising from this paper concerns the computation of the unobservable cone. The results presented in this paper allow the unobservable cone to be found by performing a search over the vector of Lagrange multiplier parameters τ. Future research is required to determine if efficient means can be found to perform this search, e.g., linear matrix inequality methods, etc.

REFERENCES

[1] D. J. CLEMENTS AND B. D. O. ANDERSON, *Singular Optimal Control: The Linear-Quadratic Problem*, Springer-Verlag, Berlin, Germany, 1978.

[2] W. S. Gray and J. P. Mesko, *Observability functions for linear and nonlinear systems*, Systems Control Lett., 38 (1999), pp. 99–113.
[3] M. Green and D. J. N. Limebeer, *Linear Robust Control*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
[4] T. Kailath, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
[5] S. O. R. Moheimani, A. V. Savkin, and I. R. Petersen, *Robust observability for a class of time-varying discrete-time uncertain systems*, Systems Control Lett., 27 (1996), pp. 261–266.
[6] S. O. R. Moheimani, A. V. Savkin, and I. R. Petersen, *Robust filtering, prediction, smoothing and observability of uncertain systems*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 45 (1998), pp. 446–457.
[7] I. R. Petersen and M. R. James, *Performance analysis and controller synthesis for nonlinear systems with stochastic uncertainty constraints*, Automatica J. IFAC, 32 (1996), pp. 959–972.
[8] I. R. Petersen and A. V. Savkin, *Robust Kalman Filtering for Signals and Systems with Large Uncertainties*, Birkhäuser Boston, Boston, 1999.
[9] I. R. Petersen, V. Ugrinovski, and A. V. Savkin, *Robust Control Design using H^∞ Methods*, Springer-Verlag, London, 2000.
[10] A. V. Savkin and I. R. Petersen, *A connection between H^∞ control and the absolute stabilizability of uncertain systems*, Systems Control Lett., 23 (1994), pp. 197–203.
[11] A. V. Savkin and I. R. Petersen, *Minimax optimal control of uncertain systems with structured uncertainty*, Internat. J. Robust Nonlinear Control, 5 (1995), pp. 119–137.
[12] A. V. Savkin and I. R. Petersen, *An uncertainty averaging approach to optimal guaranteed cost control of uncertain systems with structured uncertainty*, Automatica J. IFAC, 31 (1995), pp. 1649–1654.
[13] A. V. Savkin and I. R. Petersen, *Robust state estimation for uncertain systems with averaged integral quadratic constraints*, Internat. J. Control, 64 (1996), pp. 923–939.
[14] A. V. Savkin and I. R. Petersen, *Uncertainty averaging approach to output feedback optimal guaranteed cost control of uncertain systems*, J. Optim. Theory Appl., 88 (1996), pp. 321–337.
[15] M. M. Seron, J. H. Braslavsky, and G. C. Goodwin, *Fundamental Limitations in Filtering and Control*, Springer-Verlag, London, 1997.
[16] V. A. Ugrinovskii and I. R. Petersen, *Finite horizon minimax optimal control of nonlinear continuous time systems with stochastic uncertainty*, Dynam. Control, 10 (2000), pp. 63–86.
[17] W. M. Wonham, *Linear Multivariable Control: A Geometric Approach*, 2nd ed., Springer-Verlag, New York, 1979.
[18] V. A. Yakubovich, *Dichotomy and absolute stability of nonlinear systems with periodically nonstationary linear part*, Systems Control Lett., 11 (1988), pp. 221–228.
[19] K. Zhou, J. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice-Hall, Upper Saddle River, NJ, 1996.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 2, pp. 362–379

A STUDY ON THE OPTIMAL COST FOR THE H_∞ PROBLEM OF DISCRETE LINEAR PERIODICALLY TIME-VARYING SYSTEMS*

SHUNJI TANAKA†, TOMOMICHI HAGIWARA†, AND MITUHIKO ARAKI†

Abstract. In this paper, we propose a new method to obtain the optimal cost (the infimum of the achievable H_∞ norm of a closed-loop system) for discrete linear periodically time-varying (LPTV) systems. This method reduces via the lifting technique the original problem to the H_∞ problems of linear time-invariant (LTI) systems without the causality constraint, and therefore the optimal cost can be easily calculated. It is one of the primary advantages over other existing methods. We also show, by applying our method to LTI systems, that the optimal cost for the H_∞ problem of discrete LTI systems cannot be improved even if we use a class of noncausal controllers. An intriguing implication of this fact is further shown in the context of causal controllers.

Key words. H_∞ problem, discrete linear periodically time-varying system, causality constraint, lifting technique, optimal cost

AMS subject classifications. 93B36, 93C55

PII. S0363012900376517

1. Introduction.

This paper is devoted to the study of the H_∞ problem of discrete linear periodically time-varying (LPTV) systems. Up to now, several methods of solution have been proposed for this problem, which can be categorized, roughly speaking, into the following two approaches:

(1) time-varying approach,
(2) time-invariant (lifting) approach.

In the time-varying approach, the problem of LPTV systems is directly solved in the framework of general discrete linear time-varying (LTV) systems. As for the H_∞ problem of general discrete LTV systems, Feintuch and Francis [4] first derived a complete solution in 1985. It was based on function space analysis, and the solution was not given in such a form that engineers could easily apply it to their practical problems. In fact, the optimal cost, i.e., the infimum of the achievable H_∞ norm of the closed-loop system, was given in terms of an infinite number of operators in the function space. After the research of Feintuch and Francis [4], not much was reported on this topic for a while. Recently, however, new advances have emerged, inspired by the developments in the H_∞ theory of time-invariant systems. Actually, Dragan, Halanay, and Ionescu [3], Katayama and Ichikawa [7], and Scherpen and Verhaegen [18] gave similar expressions of the solution in terms of algebraic Riccati equations (AREs), while their approaches were mutually different. Although numerical algorithms to solve time-varying AREs are still under development, we can, in principle, obtain the solution numerically by applying these results.

In the time-invariant approach, the solution for linear time-invariant (LTI) systems is directly utilized to solve the problem of LPTV systems. Here, the lifting technique [11, 8] plays a key role. This technique associates a class of LPTV systems with an equivalent class of LTI systems. More specifically, the class of m-input,

*Received by the editors August 9, 2000; accepted for publication August 9, 2001; published electronically June 18, 2002. A preliminary form of this research was presented at the 2nd International Conference on Circuits, Systems and Computers (CSC'98) under the title *On the H_∞ problems for discrete periodically time-varying systems.*
http://www.siam.org/journals/sicon/41-2/37651.html
†Graduate School of Engineering, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan (tanaka@kuee.kyoto-u.ac.jp, hagiwara@kuee.kyoto-u.ac.jp, araki@kuee.kyoto-u.ac.jp).

p-output, discrete, linear, N-periodic systems can be shown to be equivalent to the class of mN-input, pN-output discrete LTI systems with the transfer matrix $P(\lambda)$ satisfying the condition that $P(0)$ be block lower triangular. Note that $P(0)$ gives the throughput term in the state space expression, and the above condition corresponds to the causality requirement. For this reason, the above condition on $P(0)$ is referred to as the "causality constraint." This lifting technique enables us to translate the solution of time-invariant systems to that of periodic systems and reduces the difficulties in dealing with time-varying systems to one point: "how to secure the causality constraint, i.e., how to make the controller $K(\lambda)$ satisfy the condition that $K(0)$ be block lower triangular." The methods belonging to this "time-invariant (lifting) approach" category can be classified into several subcategories according to the ways of coping with this causality constraint.

Feintuch, Khargonekar, and Tannenbaum [5] solved a sensitivity minimization problem of periodic systems based on the result in [4]. Georgiou and Khargonekar [6] proposed a constructive algorithm. Voulgaris, Dahleh, and Valavani [19] showed another algorithm for both H_∞ and H_2 problems of general multirate systems including LPTV systems. However, these three methods focus on the so-called one-block H_∞ problem, and it is not easy to extend them to the four-block H_∞ problem.

A method to solve the general four-block H_∞ problem was proposed by Chen and Qiu [2, 13]. They treated the problem in the framework of multirate sampled-data systems and introduced the notion of "nest operators." Sågfors, Toivonen, and Lennartson [15, 16] also proposed another method using the game theoretic approach and formulated the solution in terms of AREs.

In this paper, we propose an alternative way to solve the four-block H_∞ problem of discrete LPTV systems, which can be categorized into the time-invariant approach.

Our method is based on the result in [4] and thus bears a certain similarity to the results in [5, 6] in that the optimal cost for the N-periodic H_∞ problem is given in terms of the maximum of "N values." Here, even though the "N values" in the method in [5, 6] are the norms of N infinite dimensional matrices, our method gives such values as the optimal costs for the N time-invariant H_∞ problems without the causality constraint. Hence our method is more advantageous than other methods in that the optimal cost is given by solving the ordinary (in the sense that no causality constraint is imposed on them) LTI H_∞ problems.

Another important advantage of our method is that it enables us to show an interesting property of the discrete LTI H_∞ problem. More specifically, we can show that the optimal cost for the H_∞ problem of discrete LTI systems cannot be improved even if we extend the class of controllers to a certain class of noncausal systems. Although the practical meaning of this property might be unclear since noncausal controllers cannot be implemented, it certainly lays a fundamental theoretical basis for practical problems. In fact, in section 5, we study a discrete H_∞ problem in the context of sampled-data control (with a causal controller) and show, with this property, the performance limitation that arises if the sampler has a periodic delay.

The outline of this paper is as follows. In section 2, we explain some notation and definitions used in this paper, and, in section 3, we formulate the discrete LPTV H_∞ problem and convert it into a four-block model matching problem. Then, in section 4, we state our main result, which gives the optimal cost for this H_∞ problem based on the result in [4]. In section 5, we show the properties that can be obtained by applying our result to LTI systems. Finally, section 6 summarizes the results obtained in this paper.

2. Preliminaries. In this section, we introduce the notation and definitions used in this paper.

2.1. Spaces and norms. The space of complex $n \times 1$ vector valued sequences $x = \{x_k : k \geq 0\}$ is denoted by s^n or simply s. The subspace of s^n of square-summable sequences is denoted by l_2^n or simply l_2.

The norm on l_2, denoted by $\|\cdot\|_{l_2}$, is defined as

$$\|x\|_{l_2} = \left(\sum_{k=0}^{\infty} x_k^* x_k \right)^{1/2}, \tag{1}$$

where * denotes complex conjugate transpose.

The space of bounded linear operators from l_2^n to l_2^m is denoted by $\mathcal{B}^{m\times n}$ or just $\mathcal{B}$. The norm on $\mathcal{B}$, denoted by $\|\cdot\|$, is defined as

$$\|F\| = \sup_{x\in l_2} \frac{\|Fx\|_{l_2}}{\|x\|_{l_2}}. \tag{2}$$

Any operator F in $\mathcal{B}$ can be expressed by the matrix representation

$$\begin{bmatrix} F_{00} & F_{01} & F_{02} & \cdots \\ F_{10} & F_{11} & F_{12} & \cdots \\ F_{20} & F_{21} & F_{22} & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}. \tag{3}$$

The subspace of $\mathcal{B}^{m\times n}$ of causal operators is denoted by $\mathcal{C}^{m\times n}$, or simply $\mathcal{C}$, and the matrix representation of such an operator has a block lower triangular form. The subspace of $\mathcal{B}^{m\times n}$ of time-invariant operators is denoted by $\mathcal{T}^{m\times n}$, or just $\mathcal{T}$, and the matrix representation of such an operator has a block Toeplitz form.

The space of essentially bounded, matrix valued functions defined on the unit circle is denoted by L_∞. The norm on L_∞, denoted by $\|\cdot\|_\infty$, is defined as

$$\|f\|_\infty = \operatorname*{ess\,sup}_{\theta\in[0,2\pi]} \overline{\sigma}(f(e^{j\theta})), \tag{4}$$

where $\overline{\sigma}(\cdot)$ denotes the maximum singular value. The subspace of L_∞, whose element has analytic continuation into the open unit disc, is denoted by H_∞.

Let F be a time-invariant operator in $\mathcal{T}$. From its matrix representation

$$\begin{bmatrix} F_0 & F_{-1} & F_{-2} & \cdots \\ F_1 & F_0 & F_{-1} & \cdots \\ F_2 & F_1 & F_0 & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}, \tag{5}$$

define the transfer function $\widehat{F}(\lambda)$ of F by

$$\widehat{F}(\lambda) = \sum_{k=-\infty}^{\infty} F_k \lambda^k. \tag{6}$$

Then $\widehat{F}(e^{j\theta}) \in L_\infty$, and $\|F\| = \|\widehat{F}\|_\infty$. If $F \in \mathcal{C} \cap \mathcal{T}$,

$$\widehat{F}(\lambda) = \sum_{k=0}^{\infty} F_k \lambda^k, \tag{7}$$

and $\widehat{F}(e^{j\theta}) \in H_\infty$.

2.2. Shift operator and truncation operator. The kth shift operator Λ_k is defined by

$$\begin{aligned}\Lambda_k :\ & \{x_0, x_1, \ldots\} \\ & \to \begin{cases} \{\underbrace{0, \ldots, 0}_{k \text{ times}}, x_0, x_1, \ldots\} & \text{if} \quad k \geq 0, \\ \{x_{-k}, x_{-k+1}, \ldots\} & \text{if} \quad k < 0. \end{cases}\end{aligned} \tag{8}$$

For any F from s to s, the kth input/output-shift operator $S_k(F)$ $(k = 0, 1, \ldots)$ is defined by

$$S_k(F) = \Lambda_{-k} F \Lambda_k. \tag{9}$$

The kth truncation operator Π_k $(k = -1, 0, \ldots)$ is defined by

$$\begin{aligned}\Pi_k :\ & \{x_0, x_1, \ldots\} \\ & \to \begin{cases} \{0, 0, \ldots\} & \text{if} \quad k = -1, \\ \{x_0, x_1, \ldots, x_k, 0, \ldots\} & \text{if} \quad k \geq 0. \end{cases}\end{aligned} \tag{10}$$

2.3. Periodic operator and lifting operator. Periodic operators are defined as follows via the input/output-shift operator.

DEFINITION 2.1. *The operator F is called N-periodic if*

$$F = S_N(F) = \Lambda_{-N} F \Lambda_N. \tag{11}$$

The subspace consisting of N-periodic operators in $\mathcal{B}^{m\times n}$ is denoted by $\mathcal{P}_N^{m\times n}$ or just $\mathcal{P}_N$.

Let Ξ_N be the isomorphism defined by

$$\begin{aligned}\Xi_N :\ & \{x_0, x_1, \ldots\} \in s^n \\ & \to \left\{ \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{N-1} \end{bmatrix}, \begin{bmatrix} x_N \\ x_{N+1} \\ \vdots \\ x_{2N-1} \end{bmatrix}, \ldots \right\} \in s^{nN},\end{aligned} \tag{12}$$

and let $L_N(\cdot)$ be the lifting operator defined by

$$L_N(F) = \Xi_N F \Xi_N^{-1}, \quad F : s^n \to s^m. \tag{13}$$

Then, for any F in $\mathcal{C}^{m\times n} \cap \mathcal{P}_N^{m\times n}$, $L_N(F)$ belongs to $\mathcal{C}^{mN\times nN} \cap \mathcal{T}^{mN\times nN}$, and

$$\|L_N(F)\| = \|F\|. \tag{14}$$

However, the converse is not true in that for an operator F^L in $\mathcal{C}^{mN\times nN} \cap \mathcal{T}^{mN\times nN}$, it might not be expressed as $F^L = L_N(F)$ for any $F \in \mathcal{C}^{m\times n} \cap \mathcal{P}_N^{m\times n}$ because of the *causality constraint*. Namely, for $L_N^{-1}(F^L)$ to belong to $\mathcal{C}^{m\times n} \cap \mathcal{P}_N^{m\times n}$, the transfer function $\widehat{F}^L$ of F^L should have such a structure that the throughput term $\widehat{F}^L(0)$ is block lower triangular. Hence, for convenience hereafter, we define the subspace $\mathcal{W}_N$ of $\mathcal{C}^{mN\times nN} \cap \mathcal{T}^{mN\times nN}$ by

$$\mathcal{W}_N = \{F^L : L_N^{-1}(F^L) \in \mathcal{C}^{m\times n} \cap \mathcal{P}_N^{m\times n}\}. \tag{15}$$

Then any function in $\mathcal{W}_N$ can be associated with a function in $\mathcal{C}^{m\times n} \cap \mathcal{P}_N^{m\times n}$ via $L_N(\cdot)$ and $L_N^{-1}(\cdot)$. We also define the subspace $\hat{\mathcal{W}}_N$ of H_∞, the space of transfer functions $\widehat{F}^L$ whose throughput term $\widehat{F}^L(0)$ is block lower triangular. $\hat{\mathcal{W}}_N$ in H_∞ corresponds to $\mathcal{W}_N$ in $\mathcal{T}$.

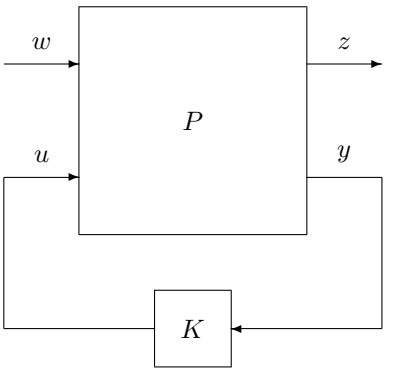


FIG. 1. *The block diagram of the discrete-time system.*

3. The discrete LPTV H_∞ problem. In this section, we formulate the discrete LPTV H_∞ problem and convert it into a model matching problem.

Consider the discrete system shown in Figure 1. In Figure 1, w is the exogenous input, u is the control input, z is the controlled output, and y is the measurement output. P denotes a discrete, linear, N-periodic, causal, finite-dimensional generalized plant that can be partitioned according to w, u, z, y into

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}. \tag{16}$$

K denotes a discrete linear causal controller.

Let us denote by $\mathcal{F}_l(P, K)$ the linear fractional transformation (LFT) of K on P; namely,

$$\mathcal{F}_l(P, K) = P_{11} + P_{12}K(I - P_{22}K)^{-1}P_{21}. \tag{17}$$

The discrete LPTV H_∞ problem of P is to find K such that

- the closed-loop system is internally stable,
- the norm of $\mathcal{F}_l(P, K)$ (the operator from w to z) is minimized.

In other words, this H_∞ problem is the following optimization problem:

$$\nu = \inf_{K:\text{causal}} \|\mathcal{F}_l(P, K)\|. \tag{18}$$

Let us assume that P_{22} admits a doubly coprime factorization and that the following assumption holds.

ASSUMPTION 3.1. *$\widehat{S}^L(\lambda)$ and $\widehat{T}^{L\sim}(\lambda) = \widehat{T}^{L^T}(\lambda^{-1})$ are injective for every λ on the unit circle.*

Then (18) can be transformed into a model matching problem of the form (see Appendix A)

$$\nu = \inf_{Q \in \mathcal{C} \cap \mathcal{P}_N} \left\| \begin{bmatrix} X - Q & Y \\ Z & U \end{bmatrix} \right\|, \tag{19}$$

where X, Y, Z, and U belong to $\mathcal{P}_N$.

In the next section, we will present our result, which gives the solution to this problem.

4. Main result: A new method to obtain the optimal cost for the discrete LPTV H_∞ problem. In this section, we give a new type of method to obtain the optimal cost for the discrete LPTV H_∞ problem: a method to calculate ν in (19). In the following, we will show three versions of our result, which gives the optimal cost ν in terms of the optimal costs for ordinary LTI H_∞ problems.

Our main result is summarized in the following theorem.

THEOREM 4.1. *For $k = 0, 1, \ldots, N-1$ let us define*

$$X_k^L = L_N(S_k(X)), \quad Y_k^L = L_N(S_k(Y)), \tag{20}$$

$$Z_k^L = L_N(S_k(Z)), \quad U_k^L = L_N(S_k(U)), \tag{21}$$

and

$$\nu_k = \inf_{Q_k^L \in \mathcal{C}\cap\mathcal{T}} \left\| \begin{bmatrix} X_k^L - Q_k^L & Y_k^L \\ Z_k^L & U_k^L \end{bmatrix} \right\|. \tag{22}$$

Then

$$\nu = \max(\nu_0,\ \nu_1, \ldots,\ \nu_{N-1}). \tag{23}$$

The proof of this theorem is given in Appendix B.

The importance of Theorem 4.1 lies in showing the fact that ν can be obtained by solving N model matching problems. It should also be noted that the infimum in (22) is not taken over $\mathcal{W}_N$ but over the larger class $\mathcal{C} \cap \mathcal{T}$. Hence it is easy to check that the model matching problem (22) is equivalent to the discrete LTI H_∞ problem

$$\nu_k = \inf_{K_k^L:\text{causal}} \|\mathcal{F}_l(P_k^L, K_k^L)\|, \tag{24}$$

where P_k^L is defined by

$$\begin{aligned} P_k^L &= \begin{bmatrix} P_{k11}^L & P_{k12}^L \\ P_{k21}^L & P_{k22}^L \end{bmatrix} \\ &= \begin{bmatrix} L_N(S_k(P_{11})) & L_N(S_k(P_{12})) \\ L_N(S_k(P_{21})) & L_N(S_k(P_{22})) \end{bmatrix}. \end{aligned} \tag{25}$$

Thus we are led to the following corollary, in which the advantage of our main result is exploited more explicitly.

COROLLARY 4.1. *The optimal cost ν is given by*

$$\nu = \max(\nu_0,\ \nu_1, \ldots,\ \nu_{N-1}), \tag{26}$$

where ν_k $(k = 0, 1, \ldots, N-1)$ is

$$\nu_k = \inf_{K_k^L:\text{causal}} \|\mathcal{F}_l(P_k^L, K_k^L)\|. \tag{27}$$

Although Corollary 4.1 holds even for the general four-block H_∞ problem, it is difficult to calculate the optimal cost for such problems analytically. For this reason, in many practical situations, we consider the suboptimal H_∞ problem to find controllers that make $\mathcal{F}_l(P, K) < \gamma$ for a given γ. Thus we will rewrite Corollary 4.1 to meet such situations.

COROLLARY 4.2. *There exists a stabilizing controller K such that*

$$\|\mathcal{F}_l(P,K)\| < \gamma, \quad K : \text{causal}, \tag{28}$$

if and only if there exists K_k^L $(k = 0, 1, \ldots, N-1)$ such that

$$\|\mathcal{F}_l(P_k^L, K_k^L)\| < \gamma, \quad K_k^L : \text{causal}. \tag{29}$$

REMARK 4.1. *Note that this corollary holds* without *Assumption* 3.1. *This is because Assumption* 3.1 *corresponds to the conditions of invariant zeros on the unit circle in the standard H_∞ problem, and therefore it can be avoided in the case of a suboptimal problem as* (28) *by using a similar method to those used in* [17, 10, 12].

Since the H_∞ problem (27) or (29) is of the LTI system P_k^L and no causality constraint is imposed on it, it can be solved by existing algorithms for the LTI H_∞ problem. Therefore, the optimal cost ν (or its upper bound γ) for the original problem can be easily calculated. As mentioned in the introduction, this point is the primary advantage of our result and the difference from the results in [5] and [6], though our method gives only the optimal cost for the LPTV H_∞ problem and therefore does not give any explicit knowledge of the structure of optimal controllers.

The following example is meant to illustrate the usefulness of our main result.

Example. Consider the 2-periodic system P whose state space realization is given by

$$P = \left[\begin{array}{c|cc} A(\cdot) & B_1(\cdot) & B_2(\cdot) \\ \hline C_1(\cdot) & D_{11}(\cdot) & D_{12}(\cdot) \\ C_2(\cdot) & D_{21}(\cdot) & 0 \end{array}\right], \tag{30}$$

where

$$\begin{aligned} &\left[\begin{array}{c|cc} A(2l+m) & B_1(2l+m) & B_2(2l+m) \\ \hline C_1(2l+m) & D_{11}(2l+m) & D_{12}(2l+m) \\ C_2(2l+m) & D_{21}(2l+m) & 0 \end{array}\right] \\ &\quad = \left[\begin{array}{c|cc} 2(m+1) & 1 & 1 \\ \hline 1 & 1 & m+1 \\ -m-1 & m+2 & 0 \end{array}\right] \quad \forall\ l,\ m = 0, 1. \end{aligned} \tag{31}$$

In view of (25), let us define

$$P_k = \left[\begin{array}{cc} S_k(P_{11}) & S_k(P_{12}) \\ S_k(P_{21}) & S_k(P_{22}) \end{array}\right], \quad k = 0, 1. \tag{32}$$

Then we have $P_0 = P$, and P_1 is given by

$$P_1 = \left[\begin{array}{c|cc} A'(\cdot) & B_1'(\cdot) & B_2'(\cdot) \\ \hline C_1'(\cdot) & D_{11}'(\cdot) & D_{12}'(\cdot) \\ C_2'(\cdot) & D_{21}'(\cdot) & 0 \end{array}\right], \tag{33}$$

where

$$\begin{aligned} &\left[\begin{array}{c|cc} A'(2l+m) & B_1'(2l+m) & B_2'(2l+m) \\ \hline C_1'(2l+m) & D_{11}'(2l+m) & D_{12}'(2l+m) \\ C_2'(2l+m) & D_{21}'(2l+m) & 0 \end{array}\right] \\ &\quad = \left[\begin{array}{c|cc} A(2l+m+1) & B_1(2l+m+1) & B_2(2l+m+1) \\ \hline C_1(2l+m+1) & D_{11}(2l+m+1) & D_{12}(2l+m+1) \\ C_2(2l+m+1) & D_{21}(2l+m+1) & 0 \end{array}\right]. \end{aligned} \tag{34}$$

Compare this with $P_0 = P$ given in (31) to see the role of the input/output-shift operator.

Now, computing the transformations given in (25) (namely, applying the lifting technique), we obtain the state space realizations of P_0^L and P_1^L as

$$P_0^L = \left[\begin{array}{c|cc:cc} 8 & 4 & 1 & 4 & 1 \\ \hline 1 & 1 & 0 & 1 & 0 \\ 2 & 1 & 1 & 1 & 2 \\ \hdashline -1 & 2 & 0 & 0 & 0 \\ -4 & -2 & 3 & -2 & 0 \end{array}\right], \quad P_1^L = \left[\begin{array}{c|cc:cc} 8 & 2 & 1 & 2 & 1 \\ \hline 1 & 1 & 0 & 2 & 0 \\ 4 & 1 & 1 & 1 & 1 \\ \hdashline -2 & 3 & 0 & 0 & 0 \\ -4 & -1 & 2 & -1 & 0 \end{array}\right]. \tag{35}$$

Since the optimal costs for the H_∞ problems of P_0^L and P_1^L, which are easily calculated, are 6.83 and 14.68, respectively, $\gamma_{\min}$ for the H_∞ problem of P is 14.68.

5. Application to LTI systems: The performance limitation with noncausal controllers. In the preceding section, we showed Theorem 4.1 (or, its explicit forms Corollaries 4.1 and 4.2), which gives the optimal cost for the discrete LPTV H_∞ problem in terms of the optimal costs for discrete LTI H_∞ problems. Although these results are primarily for the discrete LPTV H_∞ problem, they can be used to show a property in regard to the optimal cost for the discrete H_∞ problem of LTI systems when we use noncausal controllers. In this section, we show such a property and give an example of how we can apply it to practical situations.

If P is an LTI system, we can take any positive integer N and regard P as N-periodic. Then P_k^L defined by (25) satisfies

$$P_0^L = P_1^L = \cdots = P_{N-1}^L = P^L, \tag{36}$$

where P^L is

$$P^L = \begin{bmatrix} L_N(P_{11}) & L_N(P_{12}) \\ L_N(P_{21}) & L_N(P_{22}) \end{bmatrix}. \tag{37}$$

Hence, by applying Corollary 4.1 to P, we obtain

$$\nu = \inf_{K:\text{causal}} \|\mathcal{F}_l(P,K)\| = \inf_{K^L:\text{causal}} \|\mathcal{F}_l(P^L,K^L)\|. \tag{38}$$

The claim of (38) is summarized as follows.

COROLLARY 5.1. *The optimal cost for the LTI H_∞ problem of P and that for the LTI H_∞ problem of P^L, the lifted system of P, are identical. In other words, the optimal cost cannot be improved even if we use any* possibly noncausal *N-periodic controller K* such that the lifted system K^L of K is causal.

It is shown in [4, 9, 20] that we cannot improve the optimal cost even if we consider causal time-varying controllers in the case of the LTI H_∞ problem. Thus it is a matter of course that *causal* N-periodic controllers do not improve the optimal cost. However, Corollary 5.1 claims more than that: even if we use such controllers that belong to the class of *noncausal* N-periodic systems *whose lifted systems are causal*, the optimal cost cannot be improved. In spite of the facts that N can be taken arbitrarily large and that the above class gets larger as N gets larger (by an integer multiple), this does not lead to the conclusion that no noncausal controller improves the optimal cost, however. Indeed, in most cases, the optimal cost improves with such noncausal controllers that can use information of one step in the future, but such controllers never belong to the above class since every controller in that class is

incapable of using future information at $t = Nk - 1$. Although there is no qualitative explanation about the difference between these two types of noncausal controllers yet, we can show an interesting property on the H_∞ problem of a certain system by using Corollary 5.1.

Consider a discrete LTI system P, and let us consider the following two situations:

(a) The measurement output is periodically delayed by one step. That is, for some N and for some fixed q $(0 \le q \le N-1)$, the measurement output y at $t = Np + q$ $(p = 0, 1, \ldots)$ is not available by the controller until the next step $t = Np + q + 1$ because of, for example, a periodic delay in the circuit of the sampler. More specifically, we consider a sampler whose input sequence is the sequence of the (ideal) measurement output $y_0, y_1, \ldots$, and whose output sequence is

$$y_0, \ldots, y_{q-1}, \phi, \begin{bmatrix} y_q \\ y_{q+1} \end{bmatrix}, y_{q+2}, \ldots, y_{N+q-1}, \phi, \begin{bmatrix} y_{N+q} \\ y_{N+q+1} \end{bmatrix}, \ldots, \tag{39}$$

where ϕ denotes the empty set.

(b) The measurement output is always delayed by one step. In other words, we consider a sampler whose input sequence is the sequence of the (ideal) measurement output $y_0, y_1, \ldots$, and whose output sequence is

$$\phi, y_0, y_1, \ldots. \tag{40}$$

In these two situations, we are to solve the H_∞ problems and compare the optimal costs.

Intuitively, the optimal cost ν_a in (a) seems smaller than the optimal cost ν_b in (b) because, at any time instant, more information is available by controllers in (a) than in (b). However, in reality, the optimal costs in both situations are identical. It can be shown by applying Corollary 5.1, as we do in the following.

First, we consider the situation (b). Since P, together with the sampler (40), is modeled as

$$P' = \begin{bmatrix} P'_{11} & P'_{12} \\ P'_{21} & P'_{22} \end{bmatrix} = \begin{bmatrix} P_{11} & P_{12} \\ \Lambda_1 P_{21} & \Lambda_1 P_{22} \end{bmatrix}, \tag{41}$$

our problem here can be converted into the LTI H_∞ problem to find $K' = K\Lambda_1^{-1}$ for P'. Therefore, in the lifted space, we are to consider the problem to find $K'^L = L_N(K') = L_N(K\Lambda_1^{-1})$ for $P'^L = L_N(P')$ such that $\widehat{K}'^L(0)$ satisfies the causality constraint; i.e., $\widehat{K}'^L(0)$ has the block lower triangular form

$$\widehat{K}'^L(0) = \begin{bmatrix} k'_{00} & 0 & & & \\ k'_{10} & k'_{11} & & \mathbf{0} & \\ \vdots & \vdots & \ddots & \ddots & \\ k'_{N-2,0} & k'_{N-2,1} & \cdots & k'_{N-2,N-2} & 0 \\ k'_{N-1,0} & k'_{N-1,1} & \cdots & k'_{N-1,N-2} & k'_{N-1,N-1} \end{bmatrix}. \tag{42}$$

Namely, we have

$$\nu_b = \inf_{K':\text{causal}} \|\mathcal{F}_l(P', K')\| = \inf_{\substack{K'^L:\text{causal and} \\ \text{satisfying (42)}}} \|\mathcal{F}_l(P'^L, K'^L)\|. \tag{43}$$

From Corollary 5.1,

$$\nu_b = \inf_{K':\text{causal}} \|\mathcal{F}_l(P', K')\| = \inf_{K'^L:\text{causal}} \|\mathcal{F}_l(P'^L, K'^L)\|, \tag{44}$$

and hence by (43) we have

$$\inf_{\substack{K'^L:\text{causal and}\\ \text{satisfying (42)}}} \|\mathcal{F}_l(P'^L, K'^L)\| = \inf_{K'^L:\text{causal}} \|\mathcal{F}_l(P'^L, K'^L)\|. \tag{45}$$

Next we consider the situation (a). If, as in (b), we treat P together with the sampler (39) as a plant, it becomes an N-periodic system. Therefore, we can assume without loss of generality that $q = N - 1$ from Lemma B.2, and, as mentioned in the preceding section, we can restrict the class of the controllers to that of N-periodic systems. On the other hand, if we treat the delay of the measurement output as a constraint on the controller, we are to consider K such that the elements K_{ij} $(i, j = 0, 1, 2, \ldots)$ of its matrix representation satisfy

$$K_{ij} = 0, \quad i < j \text{ or } i = j = Np + q. \tag{46}$$

As a consequence, the problem that we should consider here is to find $K^L = L_N(K)$ for $P^L = L_N(P)$ such that the throughput term $\widehat{K}^L(0)$ of the transfer function of K^L has the form

$$\widehat{K}^L(0) = \begin{bmatrix} k_{00} & 0 & & \multicolumn{2}{c}{\multirow{2}{*}{0}} \\ k_{10} & k_{11} & & \\ \vdots & \vdots & \ddots & & \\ k_{N-2,0} & k_{N-2,1} & \cdots & k_{N-2,N-2} & 0 \\ k_{N-1,0} & k_{N-1,1} & \cdots & k_{N-1,N-2} & 0 \end{bmatrix}. \tag{47}$$

By using P' given by (41) and $P'^L = L_N(P')$, this problem can be further converted into the problem to find $K'^L = L_N(K\Lambda_1^{-1})$ for P'^L such that the throughput term $\widehat{K}'^L(0)$ of the transfer function of K'^L has the form

$$\widehat{K}'^L(0) = \begin{bmatrix} k'_{00} & k'_{01} & & \multicolumn{2}{c}{\multirow{2}{*}{0}} \\ k'_{10} & k'_{11} & k'_{12} & & \\ \vdots & \vdots & & \ddots & \\ k'_{N-2,0} & k'_{N-2,1} & \cdots & k'_{N-2,N-2} \\ k'_{N-1,0} & k'_{N-1,1} & \cdots & k'_{N-1,N-1} \end{bmatrix}. \tag{48}$$

Namely,

$$\nu_a = \inf_{\substack{K'^L:\text{causal and}\\ \text{satisfying (48)}}} \|\mathcal{F}_l(P'^L, K'^L)\|. \tag{49}$$

Since, as a matter of course,

$$\begin{aligned} &\{K'^L|\ \text{causal and satisfying (42)}\} \\ &\quad \subset \{K'^L|\ \text{causal and satisfying (48)}\} \\ &\quad \subset \{K'^L|\ \text{causal}\}, \end{aligned} \tag{50}$$

it follows from (45) and (49) that $\nu_a = \nu_b$. This implies that the optimal costs in both situations (a) and (b) are identical.

This example shows that when we consider the H_∞ problem of LTI systems, there is no advantage in trying to use the available current output if there exists even a *single* sequence of the measurement output $y_p,\ y_{N+p}, \ldots$ that is delayed by one step. What is worse, trying to do so rather becomes a disadvantage compared with the case that *all* the measurement output is delayed by one step, since in such a case we can restrict the class of controllers to that of LTI controllers and do not need to take time-varying controllers into account.

6. Conclusion. In this paper, we proposed a new method to obtain the optimal cost for the discrete LPTV H_∞ problem. Our method reduces the original LPTV H_∞ problem to the ordinary LTI H_∞ problems in the sense that no causality constraint is imposed. This point is the primary advantage of our method over other existing methods of solving the same problem. Another advantage is that by applying our method to LTI systems, not to LPTV systems, it can be shown that we cannot improve the H_∞ performance of LTI systems even if we use a class of noncausal controllers. We demonstrated by an example that we can apply this result in order to discuss the performance limitation of practical problems.

Appendix A. Derivation of (19). In [14], it is shown that a doubly coprime factorization of an N-periodic operator F can be obtained by a doubly coprime factorization of the LTI system $L_N(F)$ and that each factor which appears in the doubly coprime factorization of $L_N(F)$ satisfies the causality constraint. By applying this result to P_{22} in our problem, we obtain

$$P_{22} = N_r D_r^{-1} = D_l^{-1} N_l, \tag{A1}$$

$$\begin{bmatrix} X_l & -Y_l \\ -N_l & D_l \end{bmatrix} \begin{bmatrix} D_r & Y_r \\ N_r & X_r \end{bmatrix} = I, \tag{A2}$$

where N_r, D_r, X_r, Y_r, N_l, D_l, X_l, and Y_l all belong to $\mathcal{C} \cap \mathcal{P}_N$. From (A2), all controllers that internally stabilize the closed-loop system are parametrized by

$$\begin{aligned} K &= (Y_r - D_r Q)(X_r - N_r Q)^{-1} \\ &= (X_l - Q N_l)^{-1}(Y_l - Q D_l), \end{aligned} \tag{A3}$$

where $Q \in \mathcal{C}$. Substituting (A3) into (17), we obtain

$$\mathcal{F}_l(P, K) = R - SQT, \tag{A4}$$

where

$$R = P_{11} + P_{12} D_r Y_l P_{21} \in \mathcal{C} \cap \mathcal{P}_N, \tag{A5}$$

$$S = P_{12} D_r \in \mathcal{C} \cap \mathcal{P}_N, \tag{A6}$$

$$T = D_l P_{21} \in \mathcal{C} \cap \mathcal{P}_N. \tag{A7}$$

Therefore, our problem is equivalent to [4, 19]

$$\nu = \inf_{Q \in \mathcal{C}} \|R - SQT\|, \quad R, S, T \in \mathcal{C} \cap \mathcal{P}_N, \tag{A8}$$

where $Q \in \mathcal{C}$ is a free parameter in the parametrization of stabilizing controllers. Furthermore, in [1, 19] it is shown that if R, S, and T are N-periodic, the infimum in the right-hand side of (A8) remains the same if Q is restricted to $\mathcal{C} \cap \mathcal{P}_N$. Namely,

$$\begin{aligned} \nu &= \inf_{Q \in \mathcal{C}} \|R - SQT\| \\ &= \inf_{Q \in \mathcal{C} \cap \mathcal{P}_N} \|R - SQT\|. \end{aligned} \tag{A9}$$

Thus, by applying the lifting technique to (A9), we obtain

$$\nu = \inf_{Q^L \in \mathcal{W}_N} \|R^L - S^L Q^L T^L\|, \tag{A10}$$

$$R^L = L_N(R) \in \mathcal{W}_N,\ S^L = L_N(S) \in \mathcal{W}_N,\ T^L = L_N(T) \in \mathcal{W}_N \tag{A11}$$

or, equivalently,

$$\nu = \inf_{\widehat{Q}^L(\lambda) \in \hat{\mathcal{W}}_N} \|\widehat{R}^L(\lambda) - \widehat{S}^L(\lambda)\widehat{Q}^L(\lambda)\widehat{T}^L(\lambda)\|_\infty, \tag{A12}$$

$$\widehat{R}^L(\lambda),\ \widehat{S}^L(\lambda),\ \widehat{T}^L(\lambda) \in \hat{\mathcal{W}}_N, \tag{A13}$$

where $\widehat{R}^L(\lambda)$, $\widehat{S}^L(\lambda)$, and $\widehat{T}^L(\lambda)$ denote the transfer functions of R^L, S^L, and T^L, respectively.

Under Assumption 3.1, there exist inner-outer factorizations of $\widehat{S}^L(\lambda)$ and $\widehat{T}^L(\lambda)$ in (A12). Therefore, $\widehat{S}^L(\lambda)$ and $\widehat{T}^L(\lambda)$ can be factorized as follows:

$$\widehat{S}^L(\lambda) = \widehat{S}_i^L(\lambda)\widehat{S}_o^L(\lambda), \tag{A14}$$

$$\widehat{T}^L(\lambda) = \widehat{T}_o^L(\lambda)\widehat{T}_i^L(\lambda), \tag{A15}$$

$$\widehat{S}_i^L(\lambda) \in H_\infty, \quad \widehat{S}_i^{L\sim}(\lambda)\widehat{S}_i^L(\lambda) = I, \tag{A16}$$

$$\widehat{T}_i^L(\lambda) \in H_\infty, \quad \widehat{T}_i^L(\lambda)\widehat{T}_i^{L\sim}(\lambda) = I, \tag{A17}$$

$$\widehat{S}_o^L(\lambda),\ \widehat{S}_o^{L^{-1}}(\lambda),\ \widehat{T}_o^L(\lambda),\ \widehat{T}_o^{L^{-1}}(\lambda) \in H_\infty. \tag{A18}$$

In [2], it is shown that $\widehat{S}_o^L(\lambda)$, $\widehat{T}_o^L(\lambda)$ can always be chosen so as to belong to $\hat{\mathcal{W}}_N$. In this case, the mapping from $\widehat{Q}^L(\lambda)$ to $\widehat{S}_o^L(\lambda)\widehat{Q}^L(\lambda)\widehat{T}_o^L(\lambda)$ is surjective on $\hat{\mathcal{W}}_N$. Therefore, (A12) becomes

$$\nu = \inf_{\widehat{Q}^L(\lambda) \in \hat{\mathcal{W}}_N} \|\widehat{R}^L(\lambda) - \widehat{S}_i^L(\lambda)\widehat{Q}^L(\lambda)\widehat{T}_i^L(\lambda)\|_\infty. \tag{A19}$$

By multiplying

$$\begin{bmatrix} \widehat{S}_i^{L\sim}(\lambda) \\ I - \widehat{S}_i^L(\lambda)\widehat{S}_i^{L\sim}(\lambda) \end{bmatrix} \tag{A20}$$

from the left and

$$\left[\widehat{T}_i^{L\sim}(\lambda) \quad I - \widehat{T}_i^{L\sim}(\lambda)\widehat{T}_i^{L}(\lambda)\right] \tag{A21}$$

from the right of (A19), we obtain

$$\|\widehat{R}^L(\lambda) - \widehat{S}_i^L(\lambda)\widehat{Q}^L(\lambda)\widehat{T}_i^L(\lambda)\|_\infty = \left\| \begin{bmatrix} \widehat{X}^L(\lambda) - \widehat{Q}^L(\lambda) & \widehat{Y}^L(\lambda) \\ \widehat{Z}^L(\lambda) & \widehat{U}^L(\lambda) \end{bmatrix} \right\|_\infty, \tag{A22}$$

where $\widehat{X}^L(\lambda)$, $\widehat{Y}^L(\lambda)$, $\widehat{Z}^L(\lambda)$, and $\widehat{U}^L(\lambda)$ are defined by

$$\begin{aligned} &\begin{bmatrix} \widehat{X}^L(\lambda) & \widehat{Y}^L(\lambda) \\ \widehat{Z}^L(\lambda) & \widehat{U}^L(\lambda) \end{bmatrix} \\ &\quad = \begin{bmatrix} \widehat{S}_i^{L\sim}(\lambda) \\ I - \widehat{S}_i^L\widehat{S}_i^{L\sim}(\lambda) \end{bmatrix} R^L(\lambda) \ \left[\widehat{T}_i^{L\sim}(\lambda) \quad I - \widehat{T}_i^{L\sim}(\lambda)\widehat{T}_i^L(\lambda)\right]. \end{aligned} \tag{A23}$$

Thus our problem (A12) can be rewritten as

$$\nu = \inf_{\widehat{Q}^L(\lambda)\in\widehat{\mathcal{W}}_N} \left\| \begin{bmatrix} \widehat{X}^L(\lambda) - \widehat{Q}^L(\lambda) & \widehat{Y}^L(\lambda) \\ \widehat{Z}^L(\lambda) & \widehat{U}^L(\lambda) \end{bmatrix} \right\|_\infty \tag{A24}$$

or

$$\nu = \inf_{Q^L\in\mathcal{W}_N} \left\| \begin{bmatrix} X^L - Q^L & Y^L \\ Z^L & U^L \end{bmatrix} \right\|. \tag{A25}$$

Let

$$\begin{aligned} X = L_N^{-1}(X^L), \quad Y = L_N^{-1}(Y^L), \\ Z = L_N^{-1}(Z^L), \quad U = L_N^{-1}(U^L). \end{aligned} \tag{A26}$$

Then $X, Y, Z, U \in \mathcal{P}_N$, and (19) is obtained.

Appendix B. Proof of Theorem 4.1. To prove Theorem 4.1, we use the following lemmas.

LEMMA B.1 (see [4]). *Suppose that X, Y, Z, and U belong to $\mathcal{B}$ and that μ is given by*

$$\mu = \inf_{Q\in\mathcal{C}} \left\| \begin{bmatrix} X - Q & Y \\ Z & U \end{bmatrix} \right\|. \tag{B1}$$

Then

$$\mu = \sup_{k\geq -1} \|\Gamma_k\|, \tag{B2}$$

where

$$\Gamma_k = \begin{bmatrix} \Pi_k & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} X & Y \\ Z & U \end{bmatrix} \begin{bmatrix} I - \Pi_k & 0 \\ 0 & I \end{bmatrix}. \tag{B3}$$

REMARK B.1. *The reader may think that this lemma cannot be applied directly to* (19)*, because Q in* (19) *is taken over $\mathcal{C}\cap\mathcal{P}_N$, while Q in* (B1) *is taken over $\mathcal{C}$.*

However, as mentioned before, the right-hand side of (B1) *remains the same even if* Q *is restricted to* $\mathcal{C}\cap\mathcal{P}_N$ *provided that* X, Y, Z, *and* U *belong to* $\mathcal{P}_N$. *Therefore, the above lemma implies that* ν *in* (19) *is given by the right-hand side of* (B2):

$$\nu = \sup_{k\geq -1} \|\Gamma_k\|. \tag{B4}$$

LEMMA B.2. *Assume that* X, Y, Z, *and* U *belong to* $\mathcal{P}_N$. *Then, for any* $i, j = 0, 1, \ldots, N-1$,

$$\inf_{Q\in\mathcal{C}\cap\mathcal{P}_N} \left\| \begin{bmatrix} S_i(X)-Q & S_i(Y) \\ S_i(Z) & S_i(U) \end{bmatrix} \right\| = \inf_{Q\in\mathcal{C}\cap\mathcal{P}_N} \left\| \begin{bmatrix} S_j(X)-Q & S_j(Y) \\ S_j(Z) & S_j(U) \end{bmatrix} \right\|. \tag{B5}$$

Proof of Lemma B.2. Without loss of generality, we assume that $i > j$. Since the mapping $S_k(\cdot)$ is bijective on $\mathcal{C}\cap\mathcal{P}_N$,

$$\begin{aligned} &\inf_{Q\in\mathcal{C}\cap\mathcal{P}_N} \left\| \begin{bmatrix} S_k(X)-Q & S_k(Y) \\ S_k(Z) & S_k(U) \end{bmatrix} \right\| \\ &\quad = \inf_{Q\in\mathcal{C}\cap\mathcal{P}_N} \left\| \begin{bmatrix} S_k(X)-S_k(Q) & S_k(Y) \\ S_k(Z) & S_k(U) \end{bmatrix} \right\| \\ &\quad = \inf_{Q\in\mathcal{C}\cap\mathcal{P}_N} \left\| \begin{bmatrix} S_k(X-Q) & S_k(Y) \\ S_k(Z) & S_k(U) \end{bmatrix} \right\|. \end{aligned} \tag{B6}$$

Therefore, it suffices to show that

$$\|S_i(F)\| = \|S_j(F)\| \quad \forall\ F\in\mathcal{P}_N, \tag{B7}$$

for all $i, j = 0, 1, \ldots, N-1$. From the definition of $S_k(\cdot)$,

$$\begin{aligned} \|S_i(F)\| &= \|S_{i-j}(S_j(F))\| \\ &= \|\Lambda_{-i+j} S_j(F) \Lambda_{i-j}\| \\ &\leq \|S_j(F)\Lambda_{i-j}\| \\ &\leq \|S_j(F)\|. \end{aligned} \tag{B8}$$

Since (B8) is also true for $S_j(F)$ and $S_{N+i}(F)$,

$$\|S_j(F)\| \leq \|S_{N+i}(F)\| = \|S_i(F)\|. \tag{B9}$$

This completes the proof. □

Now we are in a position to prove Theorem 4.1.

Proof of Theorem 4.1. For simplicity, we consider only the 2-periodic case ($N = 2$); that is, we are to prove that

$$\nu = \max(\nu_0,\ \nu_1), \tag{B10}$$

where

$$\begin{aligned} \nu_0 &= \inf_{Q_0^L\in\mathcal{C}\cap\mathcal{T}} \left\| \begin{bmatrix} X_0^L - Q_0^L & Y_0^L \\ Z_0^L & U_0^L \end{bmatrix} \right\| \\ &= \inf_{Q^L\in\mathcal{C}\cap\mathcal{T}} \left\| \begin{bmatrix} X^L - Q^L & Y^L \\ Z^L & U^L \end{bmatrix} \right\|, \end{aligned} \tag{B11}$$

$$\nu_1 = \inf_{Q_1^L \in \mathcal{C} \cap \mathcal{T}} \left\| \begin{bmatrix} X_1^L - Q_1^L & Y_1^L \\ Z_1^L & U_1^L \end{bmatrix} \right\|. \tag{B12}$$

It would be evident that the following arguments can be extended to the general case.

Applying Lemma B.1 to (B11), ν_0 is given by

$$\nu_0 = \sup_{k \geq -1} \|\Gamma_k^L\|, \tag{B13}$$

where

$$\Gamma_k^L = \begin{bmatrix} \Pi_k & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} X^L & Y^L \\ Z^L & U^L \end{bmatrix} \begin{bmatrix} I - \Pi_k & 0 \\ 0 & I \end{bmatrix}. \tag{B14}$$

From (A26) and (B3), $\|\Gamma_k^L\|$ can be expressed in terms of Γ_k as

$$\|\Gamma_k^L\| = \|\Gamma_{2k+1}\|. \tag{B15}$$

Thus, from (B13) and (B15),

$$\nu_0 = \sup_{k=-1,1,\ldots} \|\Gamma_k\|. \tag{B16}$$

Similarly,

$$\nu_1 = \sup_{k=-1,1,\ldots} \|\Gamma_k'\|, \tag{B17}$$

where

$$\Gamma_k' = \begin{bmatrix} \Pi_k & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} S_1(X) & S_1(Y) \\ S_1(Z) & S_1(U) \end{bmatrix} \begin{bmatrix} I - \Pi_k & 0 \\ 0 & I \end{bmatrix}. \tag{B18}$$

Let the matrix representation of

$$\begin{bmatrix} X & Y \\ Z & U \end{bmatrix} \tag{B19}$$

be

$$\left[\begin{array}{ccccc|ccccc} X_{00} & X_{1,-1} & X_{0,-2} & X_{1,-3} & \cdots & Y_{00} & Y_{1,-1} & Y_{0,-2} & Y_{1,-3} & \cdots \\ X_{01} & X_{10} & X_{0,-1} & X_{1,-2} & \cdots & Y_{01} & Y_{10} & Y_{0,-1} & Y_{1,-2} & \cdots \\ X_{02} & X_{11} & X_{00} & X_{1,-1} & \cdots & Y_{02} & Y_{11} & Y_{00} & Y_{1,-1} & \cdots \\ X_{03} & X_{12} & X_{01} & X_{10} & \cdots & Y_{03} & Y_{12} & Y_{01} & Y_{10} & \cdots \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \\ \hline Z_{00} & Z_{1,-1} & Z_{0,-2} & Z_{1,-3} & \cdots & U_{00} & U_{1,-1} & U_{0,-2} & U_{1,-3} & \cdots \\ Z_{01} & Z_{10} & Z_{0,-1} & Z_{1,-2} & \cdots & U_{01} & U_{10} & U_{0,-1} & U_{1,-2} & \cdots \\ Z_{02} & Z_{11} & Z_{00} & Z_{1,-1} & \cdots & U_{02} & U_{11} & U_{00} & U_{1,-1} & \cdots \\ Z_{03} & Z_{12} & Z_{01} & Z_{10} & \cdots & U_{03} & U_{12} & U_{01} & U_{10} & \cdots \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \end{array}\right]. \tag{B20}$$

Then the matrix representation of Γ_k $(k = 1, 3, \ldots)$ is

$$\left[\begin{array}{c|ccc||ccc} & X_{0,-k-1} & X_{1,-k-2} & \cdots & Y_{00} & \cdots & Y_{0,-k-1} & \cdots \\ 0 & \vdots & \vdots & & \vdots & & \vdots & \\ & X_{0,-1} & X_{1,-2} & \cdots & Y_{0k} & \cdots & Y_{0,-1} & \cdots \\ \hline 0 & \multicolumn{3}{c||}{0} & \multicolumn{3}{c}{0} \\ \hline & Z_{0,-k-1} & Z_{1,-k-2} & \cdots & U_{00} & \cdots & U_{0,-k-1} & \cdots \\ & \vdots & \vdots & & \vdots & & \vdots & \\ 0 & Z_{0,-1} & Z_{1,-2} & \cdots & U_{0k} & \cdots & U_{0,-1} & \cdots \\ & \vdots & \vdots & & \vdots & & \vdots & \end{array}\right] \tag{B21}$$

Also, the matrix representation of Γ'_k $(k = 1, 3, \ldots)$ is

$$\left[\begin{array}{c|ccc||ccc} & X_{1,-k-1} & X_{0,-k-2} & \cdots & Y_{10} & \cdots & Y_{1,-k-1} & \cdots \\ 0 & \vdots & \vdots & & \vdots & & \vdots & \\ & X_{1,-1} & X_{0,-2} & \cdots & Y_{1k} & \cdots & Y_{1,-1} & \cdots \\ \hline 0 & \multicolumn{3}{c||}{0} & \multicolumn{3}{c}{0} \\ \hline & Z_{1,-k-1} & Z_{0,-k-2} & \cdots & U_{10} & \cdots & U_{1,-k-1} & \cdots \\ & \vdots & \vdots & & \vdots & & \vdots & \\ 0 & Z_{1,-1} & Z_{0,-2} & \cdots & U_{1k} & \cdots & U_{1,-1} & \cdots \\ & \vdots & \vdots & & \vdots & & \vdots & \end{array}\right] \tag{B22}$$

Since the matrix representation of Γ_{k-1} $(k = 1, 3, \ldots)$ is

$$\left[\begin{array}{c|ccc||ccc} & X_{1,-k} & X_{0,-k-1} & \cdots & Y_{00} & \cdots & Y_{1,-k} & \cdots \\ 0 & \vdots & \vdots & & \vdots & & \vdots & \\ & X_{1,-1} & X_{0,-2} & \cdots & Y_{0,k-1} & \cdots & Y_{1,-1} & \cdots \\ \hline 0 & \multicolumn{3}{c||}{0} & \multicolumn{3}{c}{0} \\ \hline & Z_{1,-k} & Z_{0,-k-1} & \cdots & U_{00} & \cdots & U_{1,-k} & \cdots \\ & \vdots & \vdots & & \vdots & & \vdots & \\ 0 & Z_{1,-1} & Z_{0,-2} & \cdots & U_{0,k-1} & \cdots & U_{1,-1} & \cdots \\ & \vdots & \vdots & & \vdots & & \vdots & \end{array}\right], \tag{B23}$$

Γ_{k-1} can be expressed by

$$\Gamma_{k-1} = \begin{bmatrix} \Lambda_{-1} & 0 \\ 0 & \Lambda_{-1} \end{bmatrix} \Gamma'_k \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_1 \end{bmatrix}. \tag{B24}$$

Therefore,

$$\|\Gamma'_k\| \geq \|\Gamma_{k-1}\| \quad \forall\ k = 1, 3, \ldots, \tag{B25}$$

and hence

$$\begin{aligned}\nu &= \max\left(\sup_{k=-1,1,\ldots}\|\Gamma_k\|,\ \sup_{k=0,2,\ldots}\|\Gamma_k\|\right)\\ &\le \max\left(\sup_{k=-1,1,\ldots}\|\Gamma_k\|,\ \sup_{k=-1,1,\ldots}\|\Gamma'_k\|\right)\\ &= \max(\nu_0,\ \nu_1)\end{aligned}\tag{B26}$$

by (B4), (B16), and (B17). Furthermore, from Lemmas B.1 and B.2 and (B18),

$$\begin{aligned}\nu &= \inf_{Q\in\mathcal{C}\cap\mathcal{P}_2}\left\|\begin{bmatrix} X-Q & Y\\ Z & U\end{bmatrix}\right\|\\ &= \inf_{Q\in\mathcal{C}\cap\mathcal{P}_2}\left\|\begin{bmatrix} S_1(X)-Q & S_1(Y)\\ S_1(Z) & S_1(U)\end{bmatrix}\right\|\\ &= \sup_{k\ge -1}\|\Gamma'_k\|\\ &\ge \sup_{k=-1,1,\ldots}\|\Gamma'_k\| = \nu_1.\end{aligned}\tag{B27}$$

Since $\nu \ge \nu_0$ by (B4) and (B16), this, together with (B26), implies that

$$\nu = \max(\nu_0,\ \nu_1). \qquad \square \tag{B28}$$

REFERENCES

[1] H. Chapellat, M. Dahleh, and S. Bhattacharyya, *Structure and optimality of multivariable periodic controllers*, IEEE Trans. Automat. Control, 38 (1990), pp. 1300–1303.

[2] T. Chen and L. Qiu, *$\mathcal{H}^\infty$ design of general multirate sampled-data control systems*, Automatica J. IFAC, 30 (1994), pp. 1139–1152.

[3] V. Dragan, A. Halanay, and V. Ionescu, *Infinite horizon disturbance attenuation for discrete-time systems. A Popov-Yakubovich approach*, Integral Equations Operator Theory, 19 (1994), pp. 153–215.

[4] A. Feintuch and B. A. Francis, *Uniformly optimal control of linear feedback systems*, Automatica J. IFAC, 21 (1985), pp. 563–574.

[5] A. Feintuch, P. Khargonekar, and A. Tannenbaum, *On the sensitivity minimization problem for linear time-varying periodic systems*, SIAM J. Control Optim., 24 (1986), pp. 1076–1085.

[6] T. T. Georgiou and P. P. Khargonekar, *A constructive algorithm for sensitivity optimization of periodic systems*, SIAM J. Control Optim., 25 (1987), pp. 334–340.

[7] H. Katayama and A. Ichikawa, *H_∞-control with output feedback for time-varying discrete systems*, Internat. J. Control, 63 (1996), pp. 1167–1178.

[8] P. P. Khargonekar, K. Poolla, and A. Tannenbaum, *Robust control of linear time-invariant plants using periodic compensation*, IEEE Trans. Automat. Control, 30 (1985), pp. 1088–1096.

[9] P. P. Khargonekar and K. Poolla, *Uniformly optimal control of linear time-invariant plants: Nonlinear time-varying controllers*, Systems Control Lett., 6 (1986), pp. 303–308.

[10] P. P. Khargonekar, I. R. Petersen, and K. Zhou, *Robust stabilization of uncertain linear systems: Quadratic stabilizability and H^∞ control theory*, IEEE Trans. Automat. Control, 35 (1990), pp. 356–361.

[11] R. A. Meyer and C. S. Burrus, *A unified analysis of multirate and periodically time-varying digital filters*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 22 (1975), pp. 162–168.

[12] T. Mita, *H_∞ Control*, Shoko-do, Tokyo, 1994 (in Japanese).

[13] L. Qiu and T. Chen, *Multirate sampled-data systems: All $\mathcal{H}^\infty$ suboptimal controllers and the minimum entropy controller*, IEEE Trans. Automat. Control, 44 (1999), pp. 537–550.

[14] R. Ravi, P. P. Khargonekar, K. D. Minto, and C. N. Nett, *Controller parametrization for time-varying multirate plants*, IEEE Trans. Automat. Control, 35 (1990), pp. 1259–1262.

[15] M. F. Sågfors, H. T. Toivonen, and B. Lennartson, *State-space solution to the periodic multirate H_∞ problem: A lifting approach*, in Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, CA, 1997, pp. 2061–2066.

[16] M. F. Sågfors, H. T. Toivonen, and B. Lennartson, *H_∞ control of multirate sampled-data systems: A state-space approach*, Automatica J. IFAC, 34 (1998), pp. 415–428.

[17] M. Sampei, T. Mita, and M. Nakamichi, *An algebraic approach to H_∞ output feedback control problems*, Systems Control Lett., 14 (1990), pp. 13–24.

[18] J. M. A. Scherpen and M. H. G. Verhaegen, *$\mathcal{H}_\infty$ output feedback control for linear discrete time-varying systems via the bounded real lemma*, Internat. J. Control, 65 (1996), pp. 963–993.

[19] P. G. Voulgaris, M. A. Dahleh, and L. S. Valavani, *$\mathcal{H}^\infty$ and $\mathcal{H}^2$ optimal controllers for periodic and multirate systems*, Automatica J. IFAC, 30 (1994), pp. 251–263.

[20] C. Zhang, J. Zhang, and K. Fukata, *Analysis of H_2 and H_∞ performance of discrete periodically time-varying controllers*, Automatica J. IFAC, 33 (1997), pp. 619–634.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 2, pp. 380–403

SECOND ORDER SUFFICIENT CONDITIONS FOR OPTIMAL CONTROL PROBLEMS WITH FREE FINAL TIME: THE RICCATI APPROACH*

HELMUT MAURER† AND HANS JOACHIM OBERLE‡

Abstract. Second order sufficient conditions (SSC) for control problems with control-state constraints and free final time are presented. Instead of deriving such SSC from first principles, we transform the control problem with free final time into an augmented control problem with fixed final time for which well-known SSC exist. SSC are then expressed as a condition on the positive definiteness of the second variation. A convenient numerical tool for verifying this condition is based on the Riccati approach, where one has to find a bounded solution of an associated Riccati equation satisfying specific boundary conditions. The augmented Riccati equations for the augmented control problem are derived, and their modifications on the boundary of the control-state constraint are discussed. Two numerical examples, (1) the classical Earth-Mars orbit transfer in minimal time and (2) the Rayleigh problem in electrical engineering, demonstrate that the Riccati equation approach provides a viable numerical test of SSC.

Key words. optimal control, control-state constraints, free final time, second order sufficient conditions, Riccati equation, Earth-Mars orbit transfer, Rayleigh problem

AMS subject classifications. 49K15, 49K40, 65L10, 70M20, 94C99

PII. S0363012900377419

1. Introduction.

In the last three decades, one can find an extensive literature on second order sufficient conditions (SSC) for optimal control problems with control and state constraints; cf. [2, 4, 6, 7, 8, 9, 10, 20, 25, 26, 31, 32, 35] and further literature cited in these papers. SSC have shown to be of fundamental importance for stability and sensitivity analysis of parametric optimal control problems; cf., e.g., [2, 5, 9, 10, 18, 19, 21, 23, 24, 33, 34].

SSC are usually expressed in terms of the positiveness of a quadratic form on a certain critical cone which is obtained through linearization of equality and inequality constraints. In general, such conditions are far too abstract to lend themselves to numerical verification. A practical test for SSC can be devised on the basis of a matrix-valued Riccati equation [23, 25, 37]. The main ideas underlying this approach are already exposed in the book of Bryson and Ho [1] for unconstrained control problems. This test requires the construction of a bounded solution to a Riccati equation which has to satisfy additional boundary conditions. An inherent difficulty arises from the fact that the coefficients of the Riccati equation depend on the accurate solution for state, control, and adjoint variables.

Most of the cited papers deal with control problems on a *fixed* time interval. Extensions of the results to problems with *free* final time or with nonfixed time intervals have been discussed in [1, 4, 12, 26, 31, 32]. The method in Bryson and Ho [1, Chapter 6] uses heuristic arguments and also suffers from the drawback that, e.g.,

*Received by the editors August 31, 2000; accepted for publication (in revised form) December 4, 2001; published electronically June 18, 2002. This work was supported by Deutsche Forschungsgemeinschaft under grant MA 891-3.

http://www.siam.org/journals/sicon/41-2/37741.html

†Westfälische Wilhelms-Universität Münster, Institut für Numerische Mathematik, Einsteinstrasse 62, D-48149 Münster, Germany (maurer@math.uni-muenster.de).

‡Universität Hamburg, Fachbereich Mathematik, Bundesstr. 55, D-20146 Hamburg, Germany (oberle@math.uni-hamburg.de).

time-optimal control problems are not tractable via this approach. More precisely, the function α defined in (6.6.13) of [1] is identically zero, and hence the quantity in (6.6.16) is not defined. A rigorous proof of the SSC in [1] may be found in Chamberland and Zeidan [4], where extensions of the results to control problems with mixed control-state constraints are also given. Again, however, the time-optimal case is not covered by these conditions. A remedy for this deficiency has been proposed in Hull [12] for unconstrained control problems. These conditions have been tested in a numerically unchallenging situation. We emphasize that a general approach for SSC on nonfixed time intervals has been developed by Osmolovskii [26, 31, 32], but the author does not offer any practical device to test his conditions numerically.

The aim of the present paper is to develop verifiable SCC for control problems with free final time and mixed control-state constraints. Our approach is rather straightforward in the sense that it uses the well-known idea (cf., e.g., [11]) of reducing the *free* final time case to the *fixed* final time case by treating the free final time as an *augmented* state variable. It is not surprising that this procedure will lead to an augmented set of Riccati equations and boundary conditions. For unconstrained control problems, this derivation has already been described in [22].

The organization of the paper is as follows. In section 2, we recall known SSC for control problems with fixed final time [25, 37]. Section 3 describes the effect of the time transformation on the augmented variables and functions of the problem. A straightforward calculation shows that the Riccati equation for the augmented problem splits into three separate parts. A salient feature of the approach is that it suffices to solve a reduced form of the Riccati equation on the boundary of the control-state constraint. The boundary conditions for the Riccati equation are worked out in some cases of practical interest. In particular, we derive additional sign conditions of the Riccati solution at the initial and final time which turn out to be crucial in the numerical test.

In sections 4 and 5, we apply the numerical methodology to two practical and challenging examples. A highly accurate numerical solution to both examples is obtained via the multiple shooting method [3, 29]. The classical problem of a planar Earth-Mars orbit transfer in minimal time [14, 16, 17] is treated in section 4. The augmented Riccati equation test succeeds in confirming the optimality of the numerical solution. Section 5 presents a modification of the Rayleigh problem, which has been solved in [13, 23, 36] on fixed time intervals. Surprisingly, when no control constraints are imposed, the free final time problem has several local minima and one local maximum. The augmented Riccati test is capable of proving optimality for both local minima. Then the Rayleigh problem, subject to control constraints, is studied. We derive the reduced Riccati equations on the boundary of the constraint and compute a bounded solution which satisfies the extra boundary condition.

Let us mention two further applications and extensions. First, on the basis of SSC, it is rather straightforward to perform a computational *sensitivity analysis* for both examples. The numerical techniques in [21, 24, 23, 33, 34] indicate that sensitivity differentials of optimal solutions with respect to parameters can be obtained through the solution of an additional linear boundary value problem (BVP). The second extension concerns optimal control problems with *pure* state constraints to which the techniques of this paper apply as well.

2. Second order conditions for control problems with fixed final time. We consider the following autonomous control problem (CP) subject to mixed control-state constraints: for a given final time $T > 0$, determine a control

function $u \in L^\infty(0,T;\mathbb{R}^m)$ and a state function $x \in W^{1,\infty}(0,T;\mathbb{R}^n)$ that minimize the
functional

$$F(x,u) = g(x(0),x(T)) + \int_0^T L(x(t),u(t))dt \tag{2.1}$$

subject to

$$\dot{x}(t) = f(x(t),u(t)) \quad \text{for a.e. } t \in [0,T]\,, \tag{2.2}$$

$$\varphi(x(0),x(T)) = 0\,, \tag{2.3}$$

$$C(x(t),u(t)) \le 0 \quad \text{for a.e. } t \in [0,T]\,. \tag{2.4}$$

It is assumed that the functions $g : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, $\varphi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^r$, $0 \le r \le 2n$, and $C : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^k$ are C^2-functions on appropriate open sets. In this section, the *final time* T is supposed to be specified. Further, we assume that there exists a feasible pair of functions $(x_0,u_0) \in W^{1,\infty}(0,T;\mathbb{R}^n) \times L^\infty(0,T;\mathbb{R}^m)$ satisfying the constraints (2.2)–(2.4).

The first order necessary conditions for an optimal pair (x_0,u_0) are well known in the literature [11, 28]. The *unconstrained* Hamiltonian function H^0, respectively, the *augmented* Hamiltonian H, are defined as

$$H^0(x,u,\lambda) = L(x,u) + \lambda^* f(x,u)\,, \quad H(x,u,\lambda,\mu) = H^0(x,u,\lambda) + \mu^* C(x,u), \tag{2.5}$$

where $\lambda \in \mathbb{R}^n$ denotes the adjoint variable and $\mu \in \mathbb{R}^k$ is the multiplier associated with the control-state constraint (2.4); the asterisk denotes the transpose. Henceforth, partial derivatives will often be denoted by subscripts. In the following, we shall make the hypothesis that first order conditions are satisfied in *normal form* with a nonzero cost multiplier. Hence we assume that there exist Lagrange multipliers (for convenience, we shall drop the lower subscript zero)

$$(\lambda,\mu,\rho) \in W^{1,\infty}(0,T;\mathbb{R}^n) \times L^\infty(0,T;\mathbb{R}^k) \times \mathbb{R}^r$$

such that the following first order necessary conditions hold for a.e. $t \in [0,T]$:

$$\dot{\lambda}(t) = -H_x(x_0(t),u_0(t),\lambda(t),\mu(t))^*\,, \tag{2.6}$$

$$(-\lambda(0),\lambda(T)) = \nabla_{(x(0),x(T))}\,(g + \rho^*\varphi)(x_0(0),x_0(T))\,, \tag{2.7}$$

$$H_u(x_0(t),u_0(t),\lambda(t),\mu(t)) = 0, \tag{2.8}$$

$$\mu(t) \ge 0 \quad \text{and} \quad \mu(t)^* C(x_0(t),u_0(t)) = 0, \tag{2.9}$$

$$H^0(x_0(t),u_0(t),\lambda(t)) \equiv \text{const.} \tag{2.10}$$

We shall use the notation $[t]$ to denote arguments of functions at the reference solution $x_0(t),u_0(t),\lambda(t),\mu(t)$. To introduce regularity assumptions, we consider for $\beta \ge 0$ the set of β-active constraints

$$I_\beta(t) := \{i \in \{1,\dots,k\} \mid C^i[t] \ge -\beta\}\,,$$

where C^i denotes the ith component of the vector C. In particular, for $\beta = 0$, we obtain the set of active indices

$$I_0(t) = \{i \in \{1,\dots,k\} \mid C^i[t] = 0\}\,.$$

The following regularity assumption concerns the *linear independence* of gradients for active constraints; cf. [18, 19, 21, 25, 37].

(A1) For some $\beta > 0$, the gradients $C_u^i[t]$ are uniformly linear independent for all $i \in I_\beta(t)$ a.e. on $[0,T]$.

Further, we consider a margin $\delta \geq 0$ and define the set of indices

$$J_\delta(t) := \{i \in \{1,\dots,k\} \mid \mu_i(t) > \delta\}, \quad j_\delta(t) := \operatorname{card}(J_\delta(t)),$$

where μ_i denotes the ith component of the multiplier μ. It is obvious that $J_\delta(t) \subset I_0(t)$ holds for all $\delta \geq 0$. In particular, the *strict complementarity condition* $\mu^i(t) > 0$ is valid for all indices $i \in J_\delta(t)$. It will be convenient to introduce the notation

$$C^\delta[t] = (C^i[t])_{\,i \in J_\delta(t)}\,.$$

We assume that the following *modified strict Legendre–Clebsch condition* [18, 19, 21, 25, 37] is satisfied, where $|\,.\,|$ denotes the euclidean norm.

(A2) For some $\delta > 0$, there exists $c > 0$ such that, for all $t \in [0,T]$, the estimate $v^* H_{uu}[t] v \geq c|v|^2$ holds for all $v \in \mathbb{R}^m$ satisfying $C_u^\delta[t]v = 0$.

SSC can now be derived by studying the behavior of the second variation on the variational system associated with (2.2)–(2.4). In what follows, we shall use the abbreviation $\varphi[\,0,T] = \varphi(x_0(0), x_0(T))$ and similar notation. The *variational system* of equations (2.2)–(2.4) is the set of functions $(y,v) \in W^{1,2}(0,T;\mathbb{R}^n) \times L^2(0,T;\mathbb{R}^m)$ satisfying

$$\dot y(t) = f_x[t]y(t) + f_u[t]v(t), \quad \text{a.e. } t \in [0,T]\,, \tag{2.11}$$

$$D_{x(0)}\varphi[0,T]y(0) + D_{x(T)}\varphi[0,T]y(T) = 0\,, \tag{2.12}$$

$$C_x^i[t]y(t) + C_u^i[t]v(t) = 0 \quad \forall\ i \in J_\delta(t)\,, \text{ a.e. } t \in [0,T]\,. \tag{2.13}$$

Moreover, we introduce the function

$$G(x(0),x(T)) := g(x(0),x(T)) + \rho^*\,\varphi(x(0),x(T))$$

and define the $(2n,2n)$-matrix

$$\Gamma[0,T] := D^2_{(x(0),x(T))}\,G(x_0(0),x_0(T)) = \begin{pmatrix} G_{00}[0,T] & G_{0T}[0,T] \\ G_{T0}[0,T] & G_{TT}[0,T] \end{pmatrix} \tag{2.14}$$

with obvious notation $G_{00}[0,T] = D^2_{(x(0),x(0))}G[0,T]$, $G_{0T}[0,T] = D^2_{((x(0),x(T))}G[0,T]$, etc. Then the so-called *second variation* is given by the quadratic form

$$\begin{aligned} J^2(y,v) = \frac{1}{2}\int_0^T (y(t)^*, v(t)^*) \begin{pmatrix} H_{xx}[t] & H_{xu}[t] \\ H_{ux}[t] & H_{uu}[t] \end{pmatrix} \begin{pmatrix} y(t) \\ v(t) \end{pmatrix} dt \\ + \frac{1}{2}\,(y(0)^*, y(T)^*)\Gamma[0,T] \begin{pmatrix} y(0) \\ y(T) \end{pmatrix}. \end{aligned} \tag{2.15}$$

The next theorem summarizes the SSC for a weak local minimum which are to be found in [21, 25, 35, 37].

THEOREM 2.1 (SSC for control problems with fixed final time). *Let (x_0,u_0) be admissible for problem* (CP). *Suppose that there exist multipliers $(\lambda,\mu,\rho) \in W^{1,\infty}(0,T;\mathbb{R}^n) \times L^\infty(0,T;\mathbb{R}^k) \times \mathbb{R}^r$ such that the following conditions hold:*

(1) *the necessary conditions* (2.6)–(2.10) *are satisfied;*
(2) *assumptions* (A1) *and* (A2) *hold;*
(3) *there exist* $\gamma_0 > 0$ *such that the quadratic form in* (2.15) *can be estimated from below as*

$$J^2(y,v) \geq \gamma_0 \left(||y||^2_{1,2} + ||v||^2_2 \right) \tag{2.16}$$

for all variations $(y,v) \in W^{1,2}(0,T;\mathbb{R}^n) \times L^2(0,T;\mathbb{R}^m)$ *satisfying the variational system* (2.11)–(2.13);
(4) *if* u_0 *is continuous, then one may choose* $\beta = 0$ *and* $\delta = 0$ *in assumptions* (A1) *and* (A2) *and in condition* (3).

Then for all constants $0 < \gamma < \gamma_0$ *with* γ_0 *as in* (2.16) *there exists* $\alpha > 0$ *such that*

$$F(x,u) \geq F(x_0,u_0) + \gamma \left(||x - x_0||^2_{1,2} + ||u - u_0||^2_2 \right)$$

holds for all admissible (x,u) *with* $||x - x_0||_{1,\infty} + ||u - u_0||_\infty \leq \alpha$. *In particular,* (x_0,u_0) *provides a strict weak local minimum for problem* (CP).

The SSC in the previous theorem usually are not suitable for a *direct numerical verification* in practical control problems. Let us mention that, for a *discretized* version of the control problem (CP), optimization techniques have been developed that allow us to check the positiveness condition by computing the reduced Hessian; cf. [2]. In order to obtain *verifiable sufficient conditions* for the control problem in function spaces, the SSC in Theorem 2.1 are strengthened in the following way.

Consider a symmetric matrix function $Q \in W^{1,\infty}(0,T;M_{n\times n})$. For every variation $y(t)$ satisfying the linearized state equation (2.11), we have $y(t)^*Q(t)(\dot{y}(t) - f_x[t]y(t) - f_u[t]v(t)) \equiv 0$. Adding the last identity to the second variation $J^2(y,v)$ in (2.15) and performing a partial integration, we find that the definiteness condition (2.16) in Theorem 2.1 holds if the following two conditions (a) and (b) are satisfied.

Condition (a). There exist a symmetric matrix $Q \in W^{1,\infty}(0,T;M_{n\times n})$ and $\gamma > 0$ such that

$$\begin{aligned}(y^*, v^*) &\begin{pmatrix} \dot{Q}(t) + Q(t)f_x[t] + f_x[t]^*Q(t) + H_{xx}[t] & H_{xu}[t] + Q(t)f_u[t] \\ H_{ux}[t] + f_u[t]^*Q(t) & H_{uu}[t] \end{pmatrix} \begin{pmatrix} y \\ v \end{pmatrix} \\ &\geq \gamma\, |(y,v)|^2 \end{aligned} \tag{2.17}$$

holds *uniformly* in $t \in [0,T]$ for all vectors $(y,v) \in \mathbb{R}^n \times \mathbb{R}^m$ with

$$C^i_x[t]y + C^i_u[t]v = 0 \quad \forall\, i \in J_\delta(t)\,. \tag{2.18}$$

Condition (b). The boundary condition

$$(\xi_0^*, \xi_1^*) \begin{pmatrix} G_{00}[0,T] + Q(0) & G_{0T}[0,T] \\ G_{T0}[0,T] & G_{TT}[0,T] - Q(T) \end{pmatrix} \begin{pmatrix} \xi_0 \\ \xi_1 \end{pmatrix} > 0 \tag{2.19}$$

is valid for all $(\xi_0,\xi_1) \in \mathbb{R}^n \times \mathbb{R}^n \setminus \{0\}$ satisfying

$$D_{x(0)}\varphi[0,T]\xi_0 + D_{x(T)}\varphi[0,T]\xi_1 = 0\,. \tag{2.20}$$

A first consequence is that the definiteness condition (2.16) holds if the matrix in (2.17) is positive definite on the *whole* space $\mathbb{R}^n \times \mathbb{R}^m$ and if conditions (2.19)

and (2.20) are satisfied. First, this leads to the requirement that the strict Legendre–Clebsch condition

$$H_{uu}[t] \geq c \cdot I_m \quad \forall\, t \in [0,T]\,, \quad c > 0,$$

is valid on the *whole* interval $[0,T]$. Second, by evaluating the Schur complement of this matrix and using the continuous dependence of ODEs on system data, the estimate (2.16) follows from the following assumption: there exists a solution of the Riccati equation

$$\dot{Q} = -Qf_x[t] - f_x[t]^*Q - H_{xx}[t] + (H_{xu}[t] + Qf_u[t])H_{uu}[t]^{-1}(H_{xu}[t] + Qf_u[t])^*, \tag{2.21}$$

for a.e. $t \in [0,T]$ such that the matrix function $Q(t)$ is *bounded* on $[0,T]$ and satisfies the boundary conditions (2.19) and (2.20); cf. [25, Theorem 5.2].

However, in some applications, these conditions are too strong since the Riccati equation (2.21) may fail to have a *bounded* solution; cf. the Rayleigh problem in [23]. A weaker condition can be obtained by introducing the following *modified* or *reduced* Riccati equation. For $\delta \geq 0$, recall the definition of the vector

$$C^\delta[t] = (C^i[t])_{i \in J_\delta(t)}\,, \quad j_\delta(t) = ca\ rd(J_\delta(t))\,.$$

Then the matrix $C_u^\delta[t]$ of partial derivatives has dimension $j_\delta(t) \times m$. For simplicity, the time argument will be omitted in what follows. The *pseudoinverse* of the matrix C_u^δ is given by the $(m \times j_\delta)$-matrix

$$(C_u^\delta)^+ := (C_u^\delta)^*\,(C_u^\delta (C_u^\delta)^*)^{-1},$$

which exists in view of the linear independence assumption (A1). Furthermore, let $(C_u^\delta)^\perp$ denote an $(m \times (m - j_\delta))$-matrix whose column vectors form an orthogonal basis of the kernel $\mathrm{Ker}(C_u^\delta)$. Consider then the following matrices (cf. [9, 10, 25, 37]):

$$D^\delta := -(C_u^\delta)^+ C_x^\delta\,, \quad P^\delta := (C_u^\delta)^\perp, \quad A^\delta := f_x + f_u D^\delta, \tag{2.22}$$

$$\mathcal{H}_{xx}^\delta := H_{xx} + H_{xu}D^\delta + (D^\delta)^* H_{ux} + (D^\delta)^* H_{uu} D^\delta, \tag{2.23}$$

$$\mathcal{H}_{xu}^\delta := H_{xu} + (D^\delta)^* H_{uu}\,, \tag{2.24}$$

$$(\mathcal{H}_{uu}^\delta)^{(-1)} := P^\delta(\,(P^\delta)^*\, H_{uu} P^\delta\,)^{-1}(P^\delta)^*\,. \tag{2.25}$$

Note that the $(m \times m)$-matrix $(\mathcal{H}_{uu}^\delta)^{(-1)}$ in (2.25) is well defined by virtue of assumption (A2). It follows that the estimate (2.16) holds if there exists a symmetric matrix function $Q(t)$ that solves the Riccati equation

$$\dot{Q} = -QA^\delta - (A^\delta)^* Q - \mathcal{H}_{xx}^\delta + (\mathcal{H}_{xu}^\delta + Qf_u)(\mathcal{H}_{uu}^\delta)^{(-1)}(\mathcal{H}_{xu}^\delta + Qf_u)^* \tag{2.26}$$

for a.e. $t \in [0,T]$ such that $Q(t)$ is *bounded* on $[0,T]$ and satisfies the boundary conditions (2.19), (2.20).

In general, it is rather tedious to elaborate this Riccati equation explicitly. To facilitate the numerical treatment in practical applications, we discuss special cases in more detail. On *interior arcs* with $C[t] < 0$, we have $j_\delta(t) = 0$, and thus the Riccati equation (2.26) reduces to the one introduced in (2.21). Consider now a *boundary arc* with $j_\delta(t) = m$, where we have as many control components as active constraints. Due to assumption (A1), the pseudoinverse is given by $(C_u^\delta)^+ = (C_u^\delta)^{-1}$, and hence

the matrix $P^\delta = (C_u^\delta)^\perp = 0$ in (2.22) vanishes. Then the matrices in (2.22)–(2.25) become

$$D^\delta = -(C_u^\delta)^{-1} C_x^\delta, \quad A^\delta = f_x - f_u (C_u^\delta)^{-1} C_x^\delta, \quad (\mathcal{H}_{uu}^\delta)^{(-1)} = 0, \tag{2.27}$$

$$\mathcal{H}_{xx}^\delta = H_{xx} - H_{xu}(C_u^\delta)^{-1} C_x^\delta + [(C_u^\delta)^{-1} C_x^\delta]^* [H_{uu}(C_u^\delta)^{-1} C_x^\delta - H_{ux}], \tag{2.28}$$

and thus the Riccati equation (2.26) reduces to the *linear* ODE

$$\dot{Q} = -QA^\delta - (A^\delta)^* Q - \mathcal{H}_{xx}^\delta . \tag{2.29}$$

Another special case arises for *pure control* constraints where we have $C_x \equiv 0$. Then (2.28) and (2.29) simplify to the linear ODE

$$\dot{Q} = -Qf_x - f_x^* Q - H_{xx} . \tag{2.30}$$

The Rayleigh problem in [23] provided an illustrative application of this approach.

Let us also evaluate the boundary conditions (2.19) and (2.20) in a special case of practical interest. Suppose that the boundary conditions are separated and that some components for the initial and final state are fixed according to

$$x_k(0) = a_k \ \text{ for } \ k \in K_0 \subset \{1, \dots, n\}, \quad x_k(T) = b_k \ \text{ for } \ k \in K_T \subset \{1, \dots, n\}, \tag{2.31}$$

whereas the other components are free. Denote the complements of the index sets by $K_0^c = \{1, \dots, n\} \setminus K_0$, $K_T^c = \{1, \dots, n\} \setminus K_T$. Then it is easy to see that the boundary conditions (2.19), (2.20) are satisfied if the following submatrices are positive definite:

$$[\, Q(0) \,]_{(i,j) \in K_0^c \times K_0^c} > 0, \quad [\, -Q(T) \,]_{(i,j) \in K_T^c \times K_T^c} > 0 . \tag{2.32}$$

By virtue of the continuous dependence of solutions to ODEs on systems data, one of these definiteness conditions can be relaxed. For example, it suffices to require only positive semidefiniteness

$$[\, Q(0) \,]_{(i,j) \in K_0^c \times K_0^c} \geq 0 .$$

This relaxation will be convenient for the numerical verification of SSC applied to the examples in sections 4 and 5.

3. SSC for control problems with free final time. We consider again the control problem (CP) in (2.1)–(2.4), but in this section the final time will *not* be specified. It will always be assumed that the final time $T > 0$ is positive. The first order necessary conditions for problem (CP) with free final time are well known [11, 28] and extend those given in the last section.

Let H be the Hamiltonian defined in (2.5). Then it is assumed that there exist multipliers in normal form,

$$(\lambda, \mu, \rho) \in W^{1,\infty}(0, T; \mathbb{R}^n) \times L^\infty(0, T; \mathbb{R}^k) \times \mathbb{R}^r,$$

which satisfy (2.6)–(2.10). In addition, the following transversality condition associated with the free final time T holds:

$$H[T] = 0, \quad \text{i.e.,} \quad H[t] \equiv 0 \quad \forall\, t \in [0, T] . \tag{3.1}$$

These conditions can be obtained by transforming problem (CP) with *free* final time T into a problem ($\widetilde{\text{CP}}$) with *fixed* final time $\tilde{T} = 1$. The transformation proceeds by

augmenting the state dimension and by introducing the free final time as an additional state variable. Indeed, it is this transformation that will allow us to develop SSC for the free final time case on the basis of the SSC in Theorem 2.1 for fixed final time. Now define the *new time variable* $\tau \in [0,1]$ by

$$t = \tau \cdot T\,, \quad 0 \le \tau \le 1\,. \tag{3.2}$$

We shall use the same notation $x(\tau) := x(\tau \cdot T)$ and $u(\tau) := u(\tau \cdot T)$ for the state and the control variable with respect to the new time variable τ. The *augmented* state

$$\tilde{x} := \begin{pmatrix} x \\ x_{n+1} \end{pmatrix} \in \mathbb{R}^{n+1}, \quad x_{n+1} := T\,,$$

satisfies the differential equations

$$dx/d\tau = T \cdot f(x(\tau), u(\tau))\,, \qquad dx_{n+1}/d\tau \equiv 0\,. \tag{3.3}$$

As a result of this time transformation, we consider the following augmented control problem ($\widetilde{\text{CP}}$) on the fixed time interval $[0,1]$: minimize the functional

$$F(\tilde{x}, u) = F(x, T, u) = \tilde{g}(\tilde{x}(0), \tilde{x}(1)) + \int_0^1 \tilde{L}(\tilde{x}(\tau), u(\tau))\, d\tau \tag{3.4}$$

subject to

$$d\tilde{x}/d\tau = \tilde{f}(\tilde{x}(\tau), u(\tau)), \quad \text{a.e. } \tau \in [0,1]\,, \tag{3.5}$$

$$\tilde{\varphi}(\tilde{x}(0), \tilde{x}(1)) = 0\,, \tag{3.6}$$

$$\tilde{C}(\tilde{x}(\tau), u(\tau)) \le 0, \quad \text{a.e. } \tau \in [0,1]\,. \tag{3.7}$$

The functions herein are given by

$$\tilde{g}(\tilde{x}(0), \tilde{x}(1)) := g(x(0), x(1))\,, \quad \tilde{L}(\tilde{x}, u) := T \cdot L(x, u)\,, \tag{3.8}$$

$$\tilde{f}(\tilde{x}, u) := \begin{pmatrix} T \cdot f(x, u) \\ 0 \end{pmatrix}, \tag{3.9}$$

$$\tilde{\varphi}(\tilde{x}(0), \tilde{x}(1)) := \varphi(x(0), x(1))\,, \quad \tilde{C}(\tilde{x}, u) := C(x, u)\,. \tag{3.10}$$

The transformed problem ($\widetilde{\text{CP}}$) on the fixed time interval $[0,1]$ falls into the category of control problems treated in the preceding section. Thus we are able to obtain SSC for the transformed problem ($\widetilde{\text{CP}}$) by evaluating the SSC in Theorem 2.1 for the augmented state variable $\tilde{x} = (x, T)$.

First we relate the multipliers for problem ($\widetilde{\text{CP}}$) to those of problem (CP) on the time interval $[0, T]$. The Hamiltonian for problem ($\widetilde{\text{CP}}$) becomes

$$\begin{aligned} \tilde{H}(\tilde{x}, u, \tilde{\lambda}, \tilde{\mu}) &= \tilde{L}(\tilde{x}, u) + \tilde{\lambda}^* \tilde{f}(\tilde{x}, u) + \tilde{\mu}^* \tilde{C}(\tilde{x}, u) \\ &= T \cdot [\, L(x, u) + \lambda^* f(x, u)\,] + \tilde{\mu}^* C(x, u), \end{aligned} \tag{3.11}$$

where $\tilde{\lambda}^* = (\lambda^*, \lambda_{n+1}) \in \mathbb{R}^{n+1}$ denotes the augmented adjoint variable and $\tilde{\mu} \in \mathbb{R}^k$ is the multiplier for the constraint (3.7). Introducing the scaled multiplier $\mu := \tilde{\mu}/T$, the Hamiltonian $\tilde{H}$ is related to the Hamiltonian H in (2.5) as follows:

$$\tilde{H}(\tilde{x}, u, \tilde{\lambda}, \tilde{\mu}) = T \cdot [\, L(x, u) + \lambda^* f(x, u) + \mu^* C(x, u)\,] = T \cdot H(x, u, \lambda, \mu)\,. \tag{3.12}$$

According to (2.6)–(2.10), the multipliers for problem ($\widetilde{\text{CP}}$),

$$(\tilde\lambda, \tilde\mu, \rho) \in W^{1,\infty}(0,1;\mathbb{R}^{n+1}) \times L^\infty(0,1;\mathbb{R}^k) \times \mathbb{R}^r\,, \quad \tilde\lambda = \begin{pmatrix} \lambda \\ \lambda_{n+1} \end{pmatrix},$$

satisfy the following necessary conditions for a.e. $\tau \in [0,1]$:

$$d\lambda/d\tau = -\tilde H_x[\tau]^* = -T \cdot H_x[\tau]^*\,, \quad d\lambda_{n+1}/d\tau = -\tilde H_T[\tau] = -H[\tau]\,, \tag{3.13}$$

$$(-\lambda(0), \lambda(1)) = \nabla_{(x(0),x(1))}(g + \rho^*\varphi)(x_0(0), x_0(1))\,, \tag{3.14}$$

$$\lambda_{n+1}(0) = \lambda_{n+1}(1) = 0\,, \tag{3.15}$$

$$\tilde H_u[\tau] = 0\,, \tag{3.16}$$

$$\tilde\mu(\tau) \ge 0 \quad \text{and} \quad \tilde\mu(\tau)^* C[\tau] = 0, \tag{3.17}$$

$$\tilde H[\tau] \equiv \text{const.} \quad \forall\, \tau \in [0,1]. \tag{3.18}$$

Relations (3.13), (3.15), and (3.18) immediately yield

$$0 = \tilde H[1] = T \cdot H[1]\,, \tag{3.19}$$

which proves the transversality condition (3.1). To check the Legendre–Clebsch condition in assumption (A2), one has to observe the scaling

$$\tilde H_{uu}[\tau] = T \cdot H_{uu}[\tau]\,. \tag{3.20}$$

In order to apply the SSC in Theorem 2.1 to problem ($\widetilde{\text{CP}}$), we have to evaluate all terms in relations (2.11)–(2.32) for the tilde quantities defined in (3.8)–(3.10). Recalling the scaled multiplier $\mu_0 = \tilde\mu_0/T$ in (3.12), we obtain the transformed quantities

$$\tilde f_{\tilde x} = \begin{pmatrix} T \cdot f_x & f \\ 0 & 0 \end{pmatrix}, \quad \tilde f_u = \begin{pmatrix} T \cdot f_u \\ 0 \end{pmatrix}, \quad \tilde C_{\tilde x} = (C_x, 0)\,, \quad \tilde C_u = C_u\,, \tag{3.21}$$

$$\tilde H_{\tilde x \tilde x} = \begin{pmatrix} T \cdot H_{xx} & (H^0_x)^* \\ H^0_x & 0 \end{pmatrix}, \quad \tilde H_{xu} = \begin{pmatrix} T \cdot H_{xu} \\ H^0_u \end{pmatrix}, \quad \tilde H_{uu} = T \cdot H_{uu}\,. \tag{3.22}$$

Observe that the last relations make use of the identity $\tilde H_T = H^0$, which can be seen from (3.11) with H^0 denoting the unconstrained Hamiltonian. The preceding considerations lead to the following SSC for problem ($\widetilde{\text{CP}}$) with free final time.

THEOREM 3.1 (SSC for control problems with free final time). *Let (x_0, T_0, u_0) with $T_0 > 0$ be admissible for problem ($\widetilde{\text{CP}}$). Suppose that there exist multipliers $(\lambda, \mu, \rho) \in W^{1,\infty}(0,1;\mathbb{R}^n) \times L^\infty(0,1;\mathbb{R}^k) \times \mathbb{R}^r$ such that the following conditions hold:*

1. *the necessary conditions* (3.13)–(3.19) *are satisfied;*
2. *assumptions* (A1) *and* (A2) *hold with respect to the time interval* $[0,1]$;
3. *there exists $\gamma_0 > 0$ such that the quadratic form* (2.15) *expressed in terms of the transformed functions* (3.21), (3.22) *can be estimated from below as*

$$J^2(\tilde y, \tilde v) \ge \gamma_0 \left(||\tilde y||^2_{1,2} + ||\tilde v||^2_2 \right)$$

for all variations $(\tilde y, \tilde v) \in W^{1,2}(0,1;\mathbb{R}^{n+1}) \times L^2(0,1;\mathbb{R}^m)$ which satisfy the variational system (2.11)–(2.13) *on the time interval* $[0,1]$.

(4) *if* u_0 *is continuous, then one may choose* $\beta = 0$ *and* $\delta = 0$ *in assumptions* (A1) *and* (A2) *and in condition* (3).

Then for all constants $0 < \gamma < \gamma_0$ *there exists* $\alpha > 0$ *such that*

$$F(x,T,u) \geq F(x_0,T_0,u_0) + \gamma\,(\,||x-x_0||_{1,2}^2 + |T-T_0|^2 + ||u-u_0||_2^2\,)$$

holds for all admissible (x,T,u) *with* $||x-x_0||_{1,\infty} + |T-T_0| + ||u-u_0||_\infty \leq \alpha$. *In particular,* (x_0,T_0,u_0) *provides a strict weak local minimum for problem* $(\widetilde{\mathrm{CP}})$.

Note that this theorem immediately yields SSC for the original problem (CP) since we have identified the pair of state and control functions $(x(t),u(t)) = (x(\tau\cdot T), u(\tau\cdot T))$ on the interval $[0,T]$ with the pair $(x(\tau),u(\tau))$ on the interval $[0,1]$. It is apparent that these conditions are not very handy in practical applications. Again, we may resort to the Riccati equations and boundary conditions developed in (2.21)–(2.32).

Now we consider the augmented $(n+1,n+1)$-matrix

$$\tilde{Q} = \begin{pmatrix} Q & R \\ R^* & q_T \end{pmatrix}, \tag{3.23}$$

where Q is a symmetric $(n\times n)$-matrix, R is an n-vector, and q_T is a scalar. Inserting the transformed quantities (3.21) and (3.22) into the Riccati equation (2.21) for the matrix $\tilde{Q}$, we obtain a Riccati equation for Q, a linear equation for R, and a direct integration for q_T on the interval $[0,1]$; the argument τ is omitted for simplicity:

$$dQ/d\tau = T\cdot[\,-Qf_x - f_x^*Q - H_{xx} + (H_{xu}+Qf_u)(H_{uu})^{-1}(H_{xu}+Qf_u)^*\,], \tag{3.24}$$

$$dR/d\tau = -Qf - Tf_x^*R - (H_x^0)^* + (H_{xu}+Qf_u)(H_{uu})^{-1}(H_u^0 + Tf_u^*R)^*, \tag{3.25}$$

$$dq_T/d\tau = -2R^*f + \frac{1}{T}\cdot(H_u^0 + Tf_u^*R)(H_{uu})^{-1}(H_u^0+Tf_u^*R)^*\,. \tag{3.26}$$

Clearly, the Riccati equation (3.24) evaluated on the interval $[0,1]$ agrees with the Riccati equation (2.21) on the interval $[0,T]$. Note again that $H_u^0 \equiv 0$ holds on totally interior arcs with $C[t] < 0$. We wish to draw attention to the fact that (3.25) and (3.26) are *not* identical to corresponding equations in Bryson and Ho [1, sections 6.6, 6.7] and Chamberland and Zeidan [4, formulae (28)–(30)], or Hull [12].

The *modified* Riccati equation (2.26) can be worked out on the time interval $[0,1]$ in a similar way using the transformed quantities (3.21) and (3.22). However, since this procedure is quite cumbersome, we restrict the discussion to the special case $m = j_\delta(t)$, which was considered already in (2.27)–(2.29). Upon computing the matrices in (2.27) and (2.28),

$$A^\delta = f_x - f_u(C_u^\delta)^{-1}C_x^\delta\,,$$
$$\mathcal{H}_{xx}^\delta = H_{xx} - H_{xu}(C_u^\delta)^{-1}C_x^\delta + [(C_u^\delta)^{-1}C_x^\delta]^*[H_{uu}(C_u^\delta)^{-1}C_x^\delta - H_{ux}]\,,$$

for the tilde quantities (3.21), (3.22), we recognize that the Riccati equation (2.29) splits into the following three equations:

$$dQ/d\tau = T\cdot[\,-QA^\delta - (A^\delta)^*Q - \mathcal{H}_{xx}^\delta\,], \tag{3.27}$$

$$dR/d\tau = -Qf - T\cdot(A^\delta)^*R - (H_x^0)^* - [H_u^0(C_u^\delta)^{-1}C_x^\delta]^*, \tag{3.28}$$

$$dq_T/d\tau = -2R^*f\,. \tag{3.29}$$

These formulas simplify considerably if $C_x \equiv 0$ holds, i.e., if the constraint is a *pure control* constraint. Then we get $A^\delta = f_x$, and the last equations yield

$$\begin{aligned} &dQ/d\tau = T \cdot [\, -Qf_x - f_x^* Q - H_{xx}\,], \quad dR/d\tau = -Qf - T \cdot f_x^* R - (H_x^0)^*\,, \\ &dq_T/d\tau = -2R^* f\,. \end{aligned} \tag{3.30}$$

These equations will provide a convenient test for SSC when applied to the Rayleigh problem in section 5.

It is rather tedious to write out the boundary conditions (2.19) and (2.20) in the general case. We shall only discuss the important case in which the initial and final states are fixed; i.e., $x(0) = x_0$ and $x(1) = x_1$ hold with prescribed $x_0, x_1 \in \mathbb{R}^n$. In this situation, the positive definiteness condition (2.32) evaluated for the augmented matrix $\tilde{Q}$ reduces to the following boundary conditions:

$$q_T(0) > 0 \quad \text{and} \quad q_T(1) < 0\,. \tag{3.31}$$

These conditions constitute extra conditions for the free final time case and will turn out to be crucial for the numerical examples discussed in the next two sections. Note that we may relax one of these conditions, e.g., the initial condition, to $q_T(0) \geq 0$ by virtue of the continuous dependence of ODE solutions on initial data.

4. Planar Earth-Mars transfer with minimal flight time. Rocket flights in an inverse square law field have been studied extensively in the literature; see, e.g., Kelley [14, 15], Kenneth and McGill [16], Lawden [17], Moyer and Pinkham [27], and Oberle and Taubert [30]. We consider the classical Earth-Mars orbit transfer with minimal transfer time. The state variables are r: distance of the vehicle to the sun; w: radial component of the velocity; v: horizontal component of the velocity; m: mass of the vehicle. The control variable is ψ: angle of the thrust vector with respect to local horizon. The thrust is always at its maximal value $\beta_{\max}$ since the final time is minimized. All variables are scaled according to the dynamic model treated in Oberle and Taubert [30].

The optimal control problem is to minimize the final time

$$F(x, T, u) = T = \int_0^T 1\, dt \tag{4.1}$$

subject to the equations of motion,

$$\begin{aligned} dr/dt &= w, \\ dw/dt &= \frac{v^2}{r} - \frac{1}{r^2} + \beta_{\max} \frac{c}{m} \sin\psi, \\ dv/dt &= -\frac{w\,v}{r} + \beta_{\max} \frac{c}{m} \cos\psi, \\ dm/dt &= -\beta_{\max}\,, \end{aligned} \tag{4.2}$$

and the boundary conditions for the initial and final state

$$\begin{aligned} &r(0) = 1.0, \qquad w(0) = 0.0, \quad v(0) = 1.0, \qquad m(0) = 1.0, \\ &r(T) = 1.525, \quad w(T) = 0.0, \quad v(T) = 1.0/\sqrt{r(T)}. \end{aligned} \tag{4.3}$$

The constants are given by $c = 1.872$ and $\beta_{\max} = 0.075$. The underlying physical data of the vehicle are given as follows: the initial mass is $m_0 = 679.78$ kg; the (maximal) thrust = 0.56493 N; and the (constant) equivalent exit velocity is $v_c = 55809$ m/s.

4.1. The BVP. Recall now the time transformation $t = \tau \cdot T$ introduced in (3.2). Since there is no control constraint in this problem, the Hamiltonian (3.12) is given by

$$\tilde{H} \;=\; T \cdot \left[1 \;+\; \lambda_r \, w \;+\; \lambda_w \left(\frac{v^2}{r} \;-\; \frac{1}{r^2} \;+\; \beta_{\max} \, \frac{c}{m} \, \sin\psi \right) \right. \\ \left. +\; \lambda_v \left(-\; \frac{w\,v}{r} \;+\; \beta_{\max} \, \frac{c}{m} \, \cos\psi \right) \;-\; \lambda_m \, \beta_{\max} \right]. \tag{4.4}$$

Let us evaluate the first order optimality conditions (3.13)–(3.16) and omit for convenience the subscript zero referring to the optimal solution. The *optimal control* ψ is derived from condition (3.16) and the assumed Legendre–Clebsch condition (A2) as

$$\sin\psi \;=\; -\,\frac{\lambda_w}{\sqrt{\lambda_w^2 \;+\; \lambda_v^2}} \;, \qquad \cos\psi \;=\; -\,\frac{\lambda_v}{\sqrt{\lambda_w^2 \;+\; \lambda_v^2}} \,. \tag{4.5}$$

The *adjoint equations* (3.13) on the normalized time interval $[0,1]$ are given by

$$\begin{aligned} d\lambda_r/d\tau &= T \cdot \left[\lambda_w \left(\frac{v^2}{r^2} \;-\; \frac{2}{r^3} \right) \;-\; \lambda_v \frac{wv}{r^2} \right] , \\ d\lambda_w/d\tau &= T \cdot \left[-\,\lambda_r \;+\; \lambda_v \, \frac{v}{r} \right] , \\ d\lambda_v/d\tau &= T \cdot \left[-\,\lambda_w \, \frac{2v}{r} \;+\; \lambda_v \, \frac{w}{r} \right] , \\ d\lambda_m/d\tau &= T \cdot \left[-\,\frac{\beta_{\max}\, c}{m^2} \, \sqrt{\lambda_w^2 \;+\; \lambda_v^2} \right] . \end{aligned} \tag{4.6}$$

The transversality conditions (3.14) and (3.19) yield

$$\begin{aligned} \lambda_m(1) &= 0, \\ H[1] &= 1 \;-\; \frac{\beta_{\max}\, c}{m(1)} \, \sqrt{\lambda_w(1)^2 \;+\; \lambda_v(1)^2} \;=\; 0 \,. \end{aligned} \tag{4.7}$$

After transforming the state equations (4.2) to the normalized time interval $[0,1]$ according to (3.3), we obtain a two-point BVP consisting of (4.2)–(4.8). This BVP can be further simplified by eliminating the variables $m(\tau)$ and $\lambda_m(\tau)$. The variable $m(\tau)$ is substituted according to

$$m(\tau) \;=\; 1.0 \;-\; \beta_{\max}\, T \cdot \tau, \tag{4.8}$$

and the variable λ_m can be dropped since it does not enter into the first three equations in (4.6). The reduced BVP then comprises the six ODEs with respect to the variables r, w, v, λ_r, λ_w, and λ_v on the fixed time interval $[0,1]$ and the trivial equation $dT/d\tau \equiv 0$ in view of (3.3). The corresponding boundary conditions are the six boundary conditions (4.3) with respect to r, w, and v and the Hamiltonian boundary condition in (4.7).

Once a solution of this BVP has been determined, the adjoint variable λ_m can be obtained through an integration of the last equation in (4.6):

$$\lambda_m(\tau) \;=\; T \cdot \int_\tau^1 \frac{\beta_{\max}\, c}{m(\tau)^2} \, \sqrt{\lambda_w(\tau)^2 \;+\; \lambda_v(\tau)^2} \, d\tau \,.$$

Alternatively, λ_m can be eliminated from the condition $H \equiv 0$.

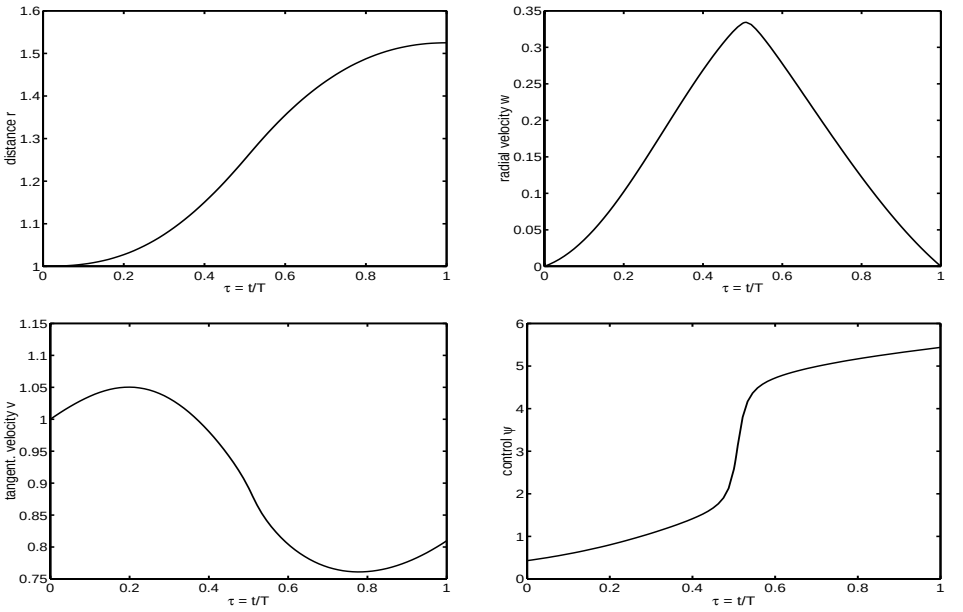


FIG. 4.1. *Earth-Mars transfer: State variables r, w, v and control variable ψ.*

The code BNDSCO in Oberle and Grimm [29] provides the following initial and final values for the adjoint variables and the final time:

$$\begin{aligned} \lambda_r(0) &= -0.52729\,67236 \times 10^1, \quad \lambda_r(1) = -0.37511\,95452 \times 10^1, \\ \lambda_w(0) &= -0.26088\,76037 \times 10^1, \quad \lambda_w(1) = 0.40004\,02548 \times 10^1, \\ \lambda_v(0) &= -0.56884\,53434 \times 10^1, \quad \lambda_v(1) = -0.35509\,23985 \times 10^1, \\ T &= 0.33199\,21219 \times 10^1. \end{aligned} \tag{4.9}$$

Figure 4.1 displays the corresponding state variables r, w, v and the control variable ψ, while Figure 4.2 shows the adjoint variables.

4.2. SSC. Let us first check the strict Legendre–Clebsch condition in assumption (A2), taking into account the scaling $\tilde{H}_{\psi\,\psi}[\tau] = T \cdot H_{\psi\,\psi}[\tau]$ in (3.20). We obtain

$$H_{\psi\,\psi} = \frac{\beta_{\max}\, c}{m} \sqrt{\lambda_w^2 + \lambda_v^2}$$

and find

$$\min\,\{\tilde{H}_{\psi\,\psi}[\tau] \mid \tau \in [\,0,1\,]\,\} = \tilde{H}_{\psi\,\psi}[\tau_0] = 0.07390\,1871 > 0, \quad \tau_0 = 0.50935\,25818\,,$$

which verifies the strict Legendre–Clebsch condition (A2). To prove the SSC in Theorem 3.1, it remains to show that the Riccati equations (3.24)–(3.26) possess a bounded solution such that the sign conditions (3.31) hold.

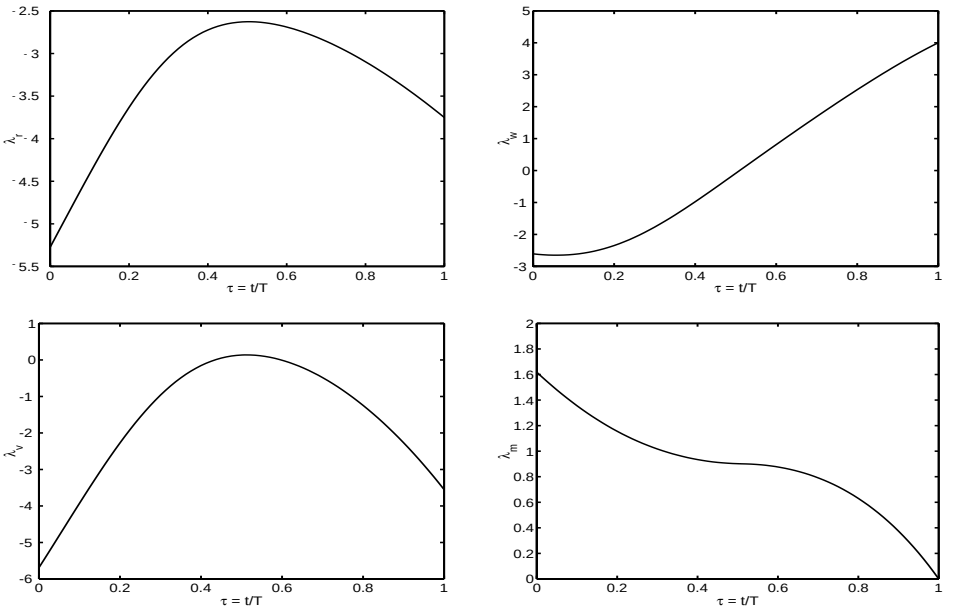


FIG. 4.2. *Earth-Mars transfer: Adjoint variables* λ_r, λ_w, λ_v, λ_m.

The reader is reminded that we have eliminated the state variable m so that the remaining state variables r, w, v have fixed final values. The symmetric Riccati matrix $\tilde{Q}$ in (3.23) is given in the form

$$\tilde{Q} = \begin{pmatrix} Q & R \\ R^* & q \end{pmatrix} = \left(\begin{array}{ccc|c} q_{11} & q_{12} & q_{13} & r_1 \\ q_{12} & q_{22} & q_{23} & r_2 \\ q_{13} & q_{23} & q_{33} & r_3 \\ \hline r_1 & r_2 & r_3 & q_T \end{array}\right).$$

We refrain from writing down the Riccati equations (3.24)–(3.26) explicitly. The evaluation is rather tedious but can be simplified with the help of symbolic computations offered, for example, in the package MAPLE. It should be noted that the coefficients of the Riccati equation are functions of the nominal trajectory characterized by (4.9). We merely indicate how to find appropriate initial values for $\tilde{Q}(0)$ such that the sign conditions (3.31) hold:

$$q_T(0) > 0, \quad q_T(1) < 0.$$

We succeeded using a rather heuristic optimization technique. Starting with initial estimates for $Q(0)$ which allowed the integration of (3.24)–(3.26) on $[0, 1]$, we changed iteratively one component of $Q(0)$ in order to minimize $q_T(1)$. Changing the indices of components in a cyclic way, we got the following initial values:

$$\begin{array}{rcrrcrrcr} q_{11}(0) & = & 1.0, & q_{12}(0) & = & 2.0, & q_{13}(0) & = & 1.0, \\ q_{22}(0) & = & -50.0, & q_{23}(0) & = & -10.0, & q_{33}(0) & = & -100.0, \\ r_1(0) & = & 80.0, & r_2(0) & = & 40.0, & r_3(0) & = & 100.0, \\ q_T(0) & = & 10.0 > 0. & & & & & & \end{array}$$

For these data, a solution of the Riccati equations was found to exist on the whole interval $[0, 1]$ with final value $q_T(1) = -18.090\,44002 < 0$. Hence all assumptions for

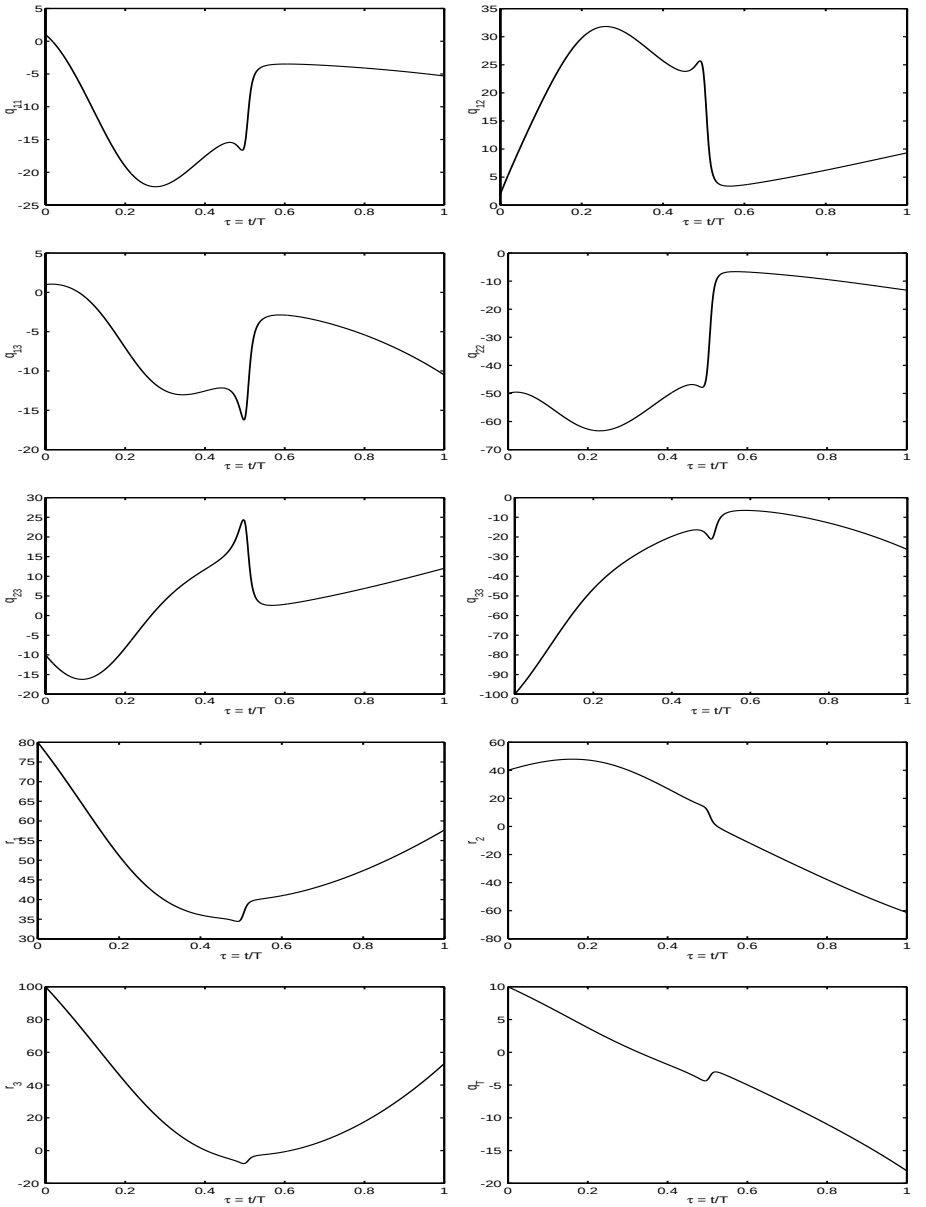


FIG. 4.3. *Solutions of the Riccati equations* (3.24)–(3.26).

Theorem 3.1 are verified, and we draw the conclusion that the trajectory characterized by (4.9) provides a weak local minimum for problem (4.1)–(4.3). The component functions of $\tilde{Q}[\tau]$ are shown in Figure 4.3.

5. Control of current in a tunnel-diode oscillator: Rayleigh problem with control constraints. The following Rayleigh problem has been treated in [13, 36] as a *fixed* final time control problem. SSC and sensitivity analysis for this model have been discussed in Maurer and Augustin [23]. In this section, we investigate a slightly modified problem with *free* final time. Figure 5.1 displays an electric circuit

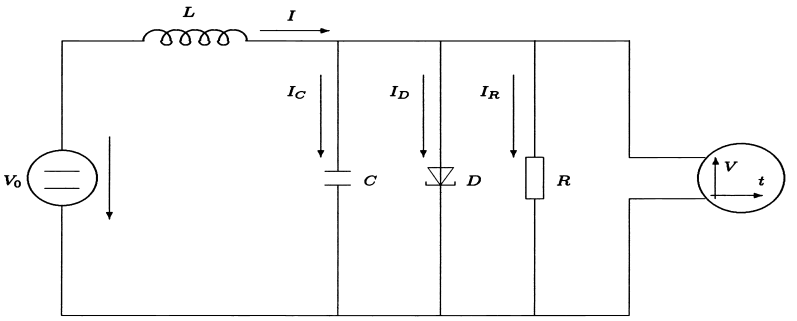


FIG. 5.1. *Tunnel-diode oscillator,* $x_1(t) = I(t)$.

(tunnel-diode oscillator), where L denotes inductivity, C denotes capacity, R denotes resistance, I denotes electric current, and D is a diode. The state variables are the electric current $x_1(t) = I(t)$ at time $t \in [\,0, T\,]$ and $x_2(t) := \dot{x}_1(t)$. The control $u(t)$ is a suitable transformation of the voltage V_0 at the generator.

With an additional parameter $c \geq 0$, the Rayleigh problem with *free* final time is defined as follows: minimize the functional

$$F_c(x, T, u) = c \cdot T + \int_0^T (\, u(t)^2 + x_1(t)^2 \,)\, dt \;=\; \int_0^T (\, c + u(t)^2 + x_1(t)^2 \,)\, dt \tag{5.1}$$

subject to

$$\dot{x}_1(t) = x_2(t)\,,\ \dot{x}_2(t) = -x_1(t) + x_2(t)\,(\,1.4 - 0.14\,x_2(t)^2\,) + 4\,u(t)\,, \tag{5.2}$$

$$x_1(0) = x_2(0) = -5\,, \quad x_1(T) = x_2(T) = 0\,, \tag{5.3}$$

$$|\,u(t)\,| \leq 1 \quad \text{for } t \in [0, T\,]\,. \tag{5.4}$$

The solution of this problem with final time $T = 4.5$ specified and $c = 0$ may be found in [23]. In the following, we denote by $F_c(T)$ the optimal value of the control problem (5.1)–(5.4) for *fixed* final time T. The behavior of this function gives insight into the behavior of optimal solutions for *free* final time.

5.1. Unconstrained optimal solutions. We consider the unconstrained problem with control constraint (5.4) deleted. After applying the time transformation (3.2), the unconstrained Hamiltonian in (3.12) becomes

$$\tilde{H}^0(\tilde{x}, u, \tilde{\lambda}) = T \cdot [\, c + u^2 + x_1^2 + \lambda_1 x_2 + \lambda_2\,(-x_1 + x_2\,(1.4 - 0.14 x_2^2) + 4u)\,]. \tag{5.5}$$

Henceforth, we omit the lower index zero to denote the optimal solution. The control is computed from the equation $\tilde{H}_u^0[\tau] = 0$, which yields

$$u(\tau) = -2\lambda_2(\tau)\,.$$

The transformed state equations (3.3) and adjoint equations (3.13) lead to the following ODEs in $[0, 1]$:

$$\begin{aligned} dx_1/d\tau &= T \cdot x_2, \\ dx_2/d\tau &= T \cdot (-x_1 + 1.4x_2 - 0.14x_2^3 - 8\lambda_2), \\ d\lambda_1/d\tau &= T \cdot (-2x_1 + \lambda_2), \\ d\lambda_2/d\tau &= T \cdot (-\lambda_1 - 1.4\lambda_2 + 0.42x_2^2\lambda_2). \end{aligned} \tag{5.6}$$

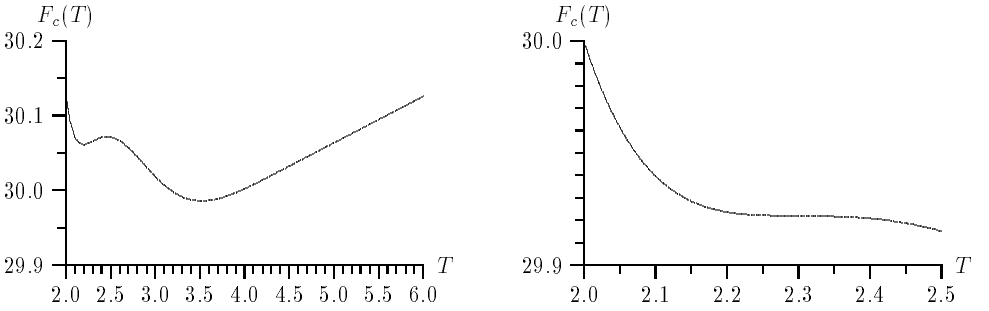


FIG. 5.2. *Graph of* $F_c(T)$ *for* $c = 1/16$ *and* $c = 0$.

To compute the optimal solution of the control problem with *fixed* final time, we resolve the BVP which comprises (5.6) and the boundary conditions

$$x_1(0) = x_2(0) = -5, \quad x_1(1) = x_2(1) = 0\,. \tag{5.7}$$

The BVP is solvable only for final times $T > T^* > 0$, where T^* is a suitable final time. The optimal value function $F_c(T)$ is shown in Figure 5.2 for $T > T^* = 2$ and two different values $c = 1/16$ and $c = 0$. Observe that, for $c = 1/16$, the graph of $F_c(T)$ in Figure 5.2 depicts two local minima and one local maximum with respect to the final time T, whereas no clear minimum can be discerned for $c = 0$.

For *free* final time T, the transversality condition (3.19) and the boundary conditions (5.7) yield $0 = \tilde{H}^0[1] = T \cdot (c + u(1)^2 + 4\lambda_2(1)u(1)) = T \cdot (c - 4\lambda_2(1)^2)$, from which we get the boundary condition

$$\lambda_2(1)^2 = c/4\,. \tag{5.8}$$

Now we solve the BVP (5.6)–(5.8) for the cases $c = 1/16$ and $c = 0$, using again the code BNDSCO in [29].

Case $c = 1/16$. We find three solutions:

Solution 1:	T	=	2.19460 79912,	$F_c(T)$	=	30.06097 62322,
	$\lambda_1(0)$	=	−9.01234 54748,	$\lambda_1(1)$	=	0.97693 36044,
	$\lambda_2(0)$	=	−2.67606 29500,	$\lambda_2(1)$	=	0.125.
Solution 2:	T	=	2.46029 38602,	$F_c(T)$	=	30.07173 02593,
	$\lambda_1(0)$	=	−9.01228 20002,	$\lambda_1(1)$	=	0.95904 74639,
	$\lambda_2(0)$	=	−2.67605 43511,	$\lambda_2(1)$	=	−0.125.
Solution 3:	T	=	3.51535 36980,	$F_c(T)$	=	29.98534 49252,
	$\lambda_1(0)$	=	−9.01085 93855,	$\lambda_1(1)$	=	0.15146 20116,
	$\lambda_2(0)$	=	−2.67586 16249,	$\lambda_2(1)$	=	−0.125.

It is obvious that these three solutions correspond to the two local minima and one local maximum shown in Figure 5.2. Note that $\lambda_2(1)$ changes sign when passing from solution 1 to solutions 2 and 3.

Now let us show that solution 3 indeed provides a local minimum. The respective optimal control, state, and adjoint variables are displayed in Figure 5.3. The

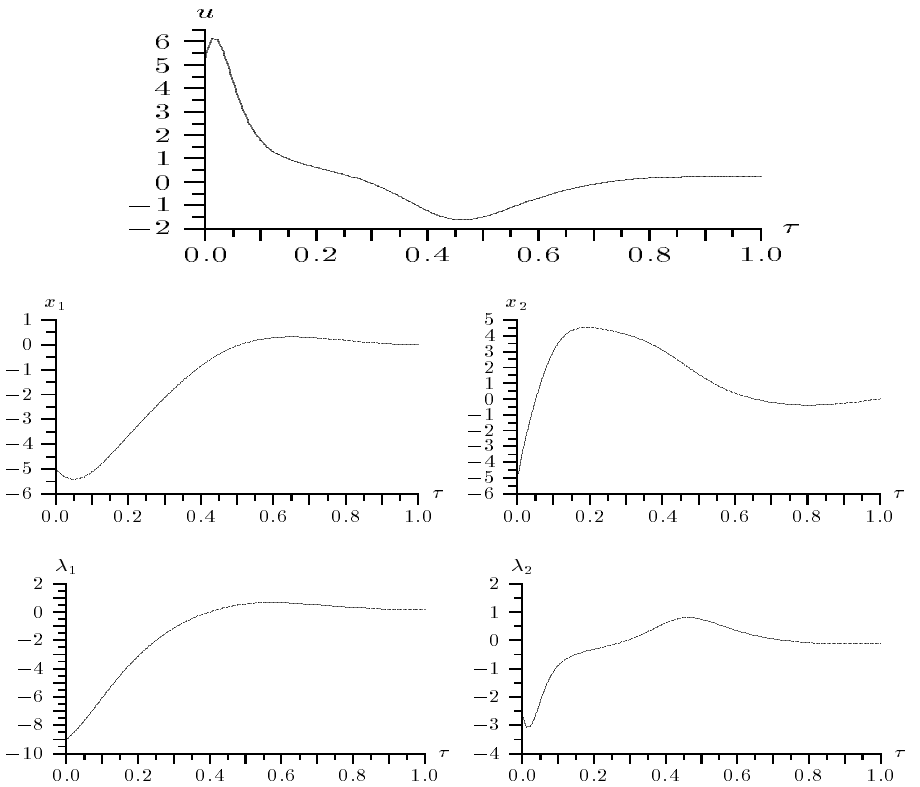


FIG. 5.3. *Optimal control, state, and adjoint variables for* $c = 1/16$ *and* $T = 3.5153536980$.

Legendre–Clebsch condition (A2) trivially holds in view of $\tilde{H}^0_{uu}[\tau] \equiv 2 \cdot T > 0$. Next we verify that the Riccati equations (3.24)–(3.26) have a bounded solution such that the boundary conditions (3.31) hold in the relaxed forms $q_T(0) \geq 0$ and $q_T(1) < 0$. The matrix (3.23) becomes

$$\widetilde{Q} = \begin{pmatrix} Q & R \\ R^* & q \end{pmatrix} =: \left(\begin{array}{cc|c} q_1 & q_2 & r_1 \\ q_2 & q_4 & r_2 \\ \hline r_1 & r_2 & q_T \end{array}\right), \tag{5.9}$$

for which we evaluate the Riccati equations (3.24)–(3.26) as

$$\begin{aligned}
dq_1/d\tau &= T \cdot [\, 2q_2 - 2 + 8q_2^2 \,], \\
dq_2/d\tau &= T \cdot [\, -q_1 - (1.4 - 0.42x_2^2)q_2 + q_4 + 8q_2q_4 \,], \\
dq_4/d\tau &= T \cdot [\, -2(q_2 + (1.4 - 0.42x_2^2)q_4) + 0.84x_2\lambda_2 + 8q_4^2 \,], \\
dr_1/d\tau &= -q_1x_2 - q_2(-x_1 + 1.4x_2 - 0.14x_2^3 - 8\lambda_2) + Tr_2 \\
&\quad -2x_1 + \lambda_2 + 8Tq_2r_2\,, \\
dr_2/d\tau &= -q_2x_2 - q_4(-x_1 + 1.4x_2 - 0.14x_2^3 - 8\lambda_2) - Tr_1 \\
&\quad -T \cdot (1.4 - 0.42x_2^2)r_2 - \lambda_1 - \lambda_2(1.4 - 0.42x_2^2) + 8Tq_4r_2\,, \\
dq_T/d\tau &= -2[\, r_1x_2 + r_2(-x_1 + 1.4x_2 - 0.14x_2^3 - 8\lambda_2) \,] + 8Tr_2^2\,.
\end{aligned} \tag{5.10}$$

It suffices to find a bounded solution of these Riccati equations satisfying $q_T(0) = 0$. After several trials, we were successful with the initial values

$$\begin{array}{lll} q_1(0) = 2.00684\,76891, & q_2(0) = 0.47018\,97048, & q_4(0) = -0.35197\,44265, \\ r_1(0) = 0, & r_2(0) = 0, & q_T(0) = 0, \end{array}$$

for which we get the final values,

$$\begin{array}{lll} q_1(1) = 0, & q_2(1) = 0, & q_4(1) = 0, \\ r_1(1) = -0.125 = \lambda_2(1), & r_2(1) = 0.02353\,79884, & q_T(1) = -r_2(1) < 0, \end{array}$$

and the bound $||\widetilde{Q}(\tau)||_\infty \leq 3$ for all $\tau \in [0,1]$. Hence Theorem 3.1 asserts that solution 3 is indeed a weak local minimum. We mention that the sufficient conditions in Hull [12] can also be checked numerically for solution 3.

In a similar way, we can test the optimality of solution 1 with $T = 2.1946\,079912$. We obtain a bounded solution of the Riccati equation for initial values

$$\begin{array}{lll} q_1(0) = 1.59068\,73787, & q_2(0) = 0.33016\,84322, & q_4(0) = -0.39917\,54076, \\ r_1(0) = 0, & r_2(0) = 0, & q_T(0) = 0 \end{array}$$

and final values

$$\begin{array}{lll} q_1(1) = 0, & q_2(1) = 0, & q_4(1) = 0, \\ r_1(1) = 0.125 = \lambda_2(1), & r_2(1) = -1.15193\,36046, & q_T(1) = r_2(1) < 0. \end{array}$$

These values yield the bound $||\widetilde{Q}(\tau)||_\infty \leq 5$ for all $\tau \in [0,1]$.

The situation is different for solution 2, which provides a *local maximum* with respect to the final time T. All initial values $q_T(0) \geq 0$ that we tested produce a solution of the Riccati equation with $q_T(1) \geq 0$. Though this test does not exclude optimality of the solution, it is a rather strong indication of nonoptimality. Thus we may draw the conclusion that solution 2 behaves like a *saddle point solution*, which is a local minimum with respect to control for every fixed time but a local maximum with respect to final time.

Case $c = 0$. Here the situation is more complicated since Figure 5.2 does not indicate a distinctive local minimum. The code BNDSCO of [29] provides, e.g., the following two solutions:

$$\begin{array}{lrcrrcr} \text{Solution 1:} & T & = & 2.29903\,95815, & F_c(T) & = & 29.9218\,12616, \\ & \lambda_1(0) & = & -9.00409\,78999, & \lambda_1(1) & = & 1.00813\,43679, \\ & \lambda_2(0) & = & -2.67325\,14651, & \lambda_2(1) & = & 0. \end{array}$$

$$\begin{array}{lrcrrcr} \text{Solution 2:} & T & = & 4.50237\,87337, & F_c(T) & = & 29.75107\,51464, \\ & \lambda_1(0) & = & -9.00247\,06599, & \lambda_1(1) & = & -0.0044561\,06307, \\ & \lambda_2(0) & = & -2.67303\,08344, & \lambda_2(1) & = & 0. \end{array}$$

Solving the Riccati equation (5.10), e.g., for the final time $T = 2.2990\ 395815$, we find that the initial value $q_T(0) = 0$ produces the final value $q_T(1) = 0$ in all tested cases. Thus the sign conditions (3.29) cannot be verified, and the Riccati test is not able to detect whether solution 1 is a local minimum.

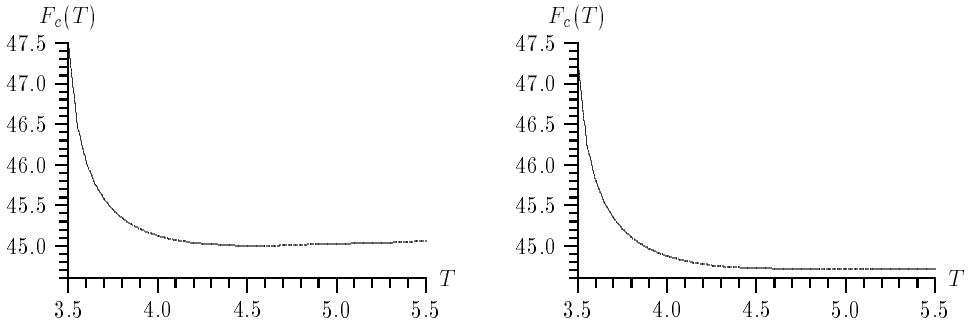


FIG. 5.4. *Graph of* $F_c(T)$ *for control constraint* $|u| \le 1$: *Cases* $c = 1/16$ *and* $c = 0$.

5.2. Constrained optimal solutions. Now we consider solutions satisfying, in addition, the control constraint (5.4),

$$-1 \le u(\tau) \le 1 \quad \forall\ \tau \in [0,1]\,.$$

The augmented Hamiltonian (3.12) becomes, in view of (5.5),

$$\begin{aligned}(5.11)\quad \tilde{H}(\tilde{x},u,\tilde{\lambda},\tilde{\mu}) &= \tilde{H}^0(\tilde{x},u,\tilde{\lambda}) + \tilde{\mu}_1(-u-1) + \tilde{\mu}_2(u-1)\\ &= T\cdot\left[\, c + u^2 + x_1^2 + \lambda_1 x_2 + \lambda_2\left(-x_1 + x_2\,(1.4 - 0.14x_2^2) + 4u\right)\right.\\ &\qquad \left. + \mu_1(-u-1) + \mu_2(u-1)\,\right],\end{aligned}$$

where $\mu_i := \tilde{\mu}/T$, $i = 1,2$, are the scaled multipliers. The state and adjoint equations agree with those given in (5.6). Again, we get the control law $u(\tau) = -2\lambda_2(\tau)$ on interior arcs $|u(\tau)| < 1$. The *unconstrained* control $u(\tau)$ depicted in Figure 5.3 suggests that the *constrained* control has one boundary arc with $u(\tau) \equiv 1$ and one boundary arc with $u(\tau) \equiv -1$. Thus we may assume the following solution structure of the optimal control:

$$(5.12)\qquad u(\tau) = \left\{\begin{array}{rl} 1, & 0 \le \tau \le \tau_1\\ -2\lambda_2(\tau), & \tau_1 \le \tau \le \tau_2\\ -1, & \tau_2 \le \tau \le \tau_3\\ -2\lambda_2(\tau), & \tau_3 \le \tau \le 1\end{array}\right\}.$$

The junction points τ_1, τ_2, τ_3 are implicitly determined through the conditions that the control is *continuous* at these points. This leads to the junction conditions

$$(5.13)\qquad \lambda_2(\tau_1) = -0.5, \quad \lambda_2(\tau_2) = 0.5, \quad \lambda_2(\tau_3) = -0.5\,.$$

Hence, on the interval $[0,1]$, we have to solve the multipoint BVP, which is composed by the state and adjoint equations (5.6) with control substituted from (5.12) as well as the boundary and junction conditions (5.7) and (5.13).

The optimal value function $\tilde{F}_c(T)$ for the constrained problem is depicted in Figure 5.4 for the values $c = 1/16$ and $c = 0$. A distinctive minimum can only be detected in case $c = 1/16$.

Case $c = 1/16$. Again we use the code BNDSCO in [29] and obtain

$$\begin{array}{rclrcl} T &=& 4.54230\,98018, & \tau_1 &=& 0.22932\,06694,\\ \tau_2 &=& 0.37589\,55717, & \tau_3 &=& 0.63465\,63122,\\ \lambda_1(0) &=& -12.70813\,77440, & \lambda_1(1) &=& 0.02860\,41331,\\ \lambda_2(0) &=& -4.59503\,53190, & \lambda_2(1) &=& -0.125,\\ F_c(T) &=& 44.71797\,06589. & & & \end{array}$$

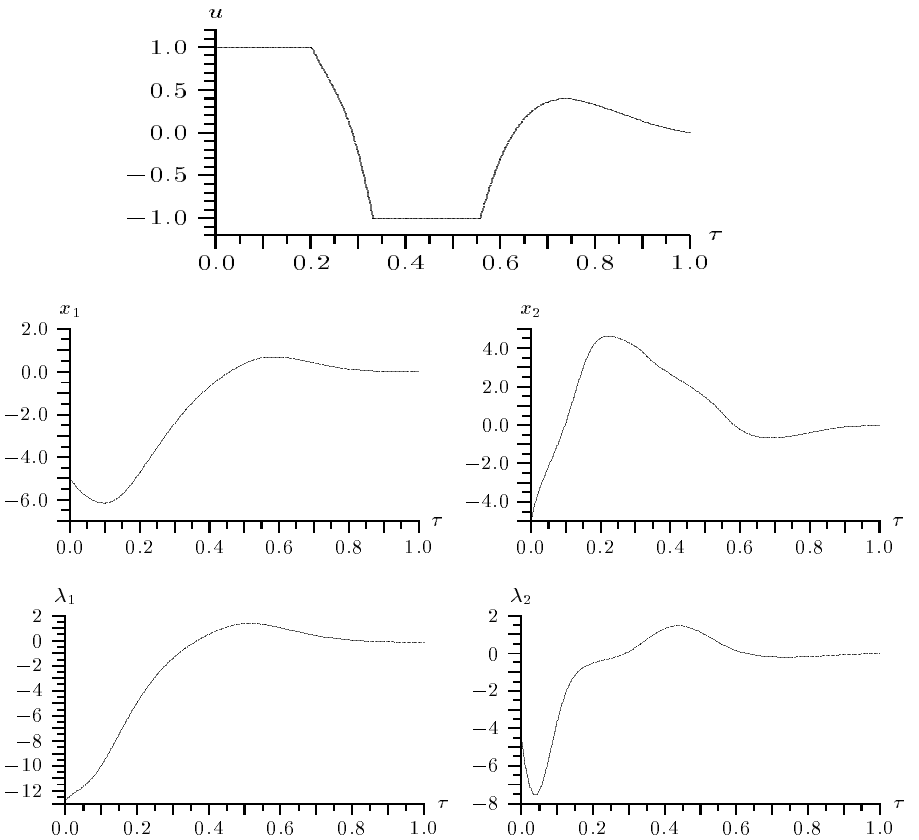


FIG. 5.5. *Optimal control, state, and adjoint variables for* $c = 1/16$ *and* $T = 4.54230\ 98018$.

The corresponding optimal control, state, and adjoint variables are shown in Figure 5.5. In order to check the sufficient conditions in Theorem 3.1, we try to find a *bounded* solution of the Riccati equations (5.10) on the *interior arcs* $[\tau_1, \tau_2]$ and $[\tau_3, 1]$ and the *modified* Riccati equations (3.30) on the *boundary arcs* $[0, \tau_1]$ and $[\tau_2, \tau_3]$. The modified Riccati equations yield the following *linear* equations:

$$
\begin{aligned}
dq_1/d\tau &= T \cdot 2(q_2 - 1)\,, \\
dq_2/d\tau &= T \cdot [\,-q_1 - (1.4 - 0.42x_2^2)q_2 + q_4\,]\,, \\
dq_4/d\tau &= T \cdot [\,-2(q_2 + (1.4 - 0.42x_2^2)q_4) + 0.84x_2\lambda_2\,]\,, \\
dr_1/d\tau &= -q_1x_2 - q_2(-x_1 + 1.4x_2 - 0.14x_2^3 + 4u) + Tr_2 - 2x_1 + \lambda_2\,, \\
dr_2/d\tau &= -q_2x_2 - q_4(-x_1 + 1.4x_2 - 0.14x_2^3 + 4u) - Tr_1 \\
&\quad -T \cdot (1.4 - 0.42x_2^2)r_2 - \lambda_1 - \lambda_2(1.4 - 0.42x_2^2)\,, \\
dq_T/d\tau &= -2\,[\,r_1x_2 + r_2(-x_1 + 1.4x_2 - 0.14x_2^3 + 4u)\,]\,.
\end{aligned}
\tag{5.14}
$$

We wish to find a solution satisfying the terminal values $q_1(1) = q_2(1) = q_4(1) = 0$. A bounded solution of the Riccati equations then is obtained for the initial values

$$
\begin{aligned}
&q_1(0) = 2.39837121, \quad q_2(0) = 0.89021498, \quad q_4(0) = -1.26573031, \\
&r_1(0) = 0, \qquad\qquad r_2(0) = 0, \qquad\qquad q_T(0) = 0,
\end{aligned}
$$

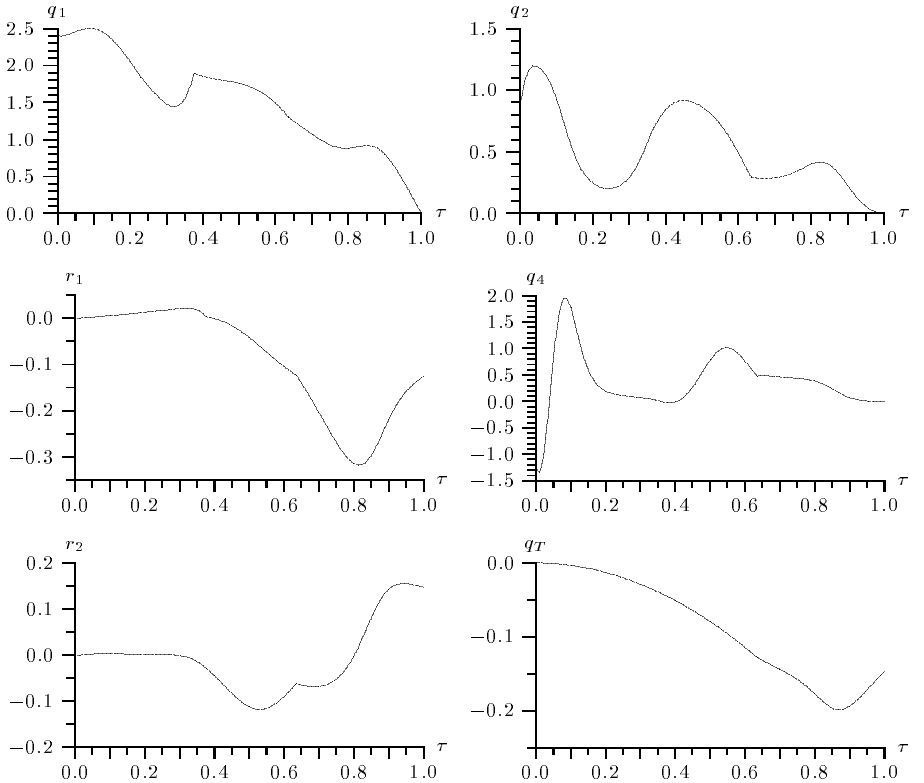


FIG. 5.6. *Solutions* $q_1, q_2, q_4, r_1, r_2, q_T$ *of Riccati equations* (5.10) *and* (5.14) *for* $c = 1/16$ *and* $T = 4.54230\ 98018$.

which produce the desired terminal values for q_1, q_2, q_4 and

$$r_1(1) = -0.125 = \lambda_2(1), \quad r_2(1) = 0.146395866, \quad q_T(1) = -r_2(1) < 0\,.$$

For these values, we get the bound $||\widetilde{Q}(\tau)||_\infty \leq 4$ for all $\tau \in [0,1]$, which can be seen in Figure 5.6. It is interesting to note that it was not possible to obtain a *bounded solution* of the Riccati equation (5.10) on the *whole interval*. Thus the Riccati test developed in section 3 in its weaker form considerably facilitates the numerical check of SSC.

Case $c = 0$. The code BNDSCO yields the solution

$$\begin{aligned}
T &= 5.15173\,31990, & \tau_1 &= 0.20192\,87957,\\
\tau_2 &= 0.33186\,51046, & \tau_3 &= 0.55797\,03824,\\
\lambda_1(0) &= -12.70087\,48310, & \lambda_1(1) &= 0.09649\,19382,\\
\lambda_2(0) &= -4.58996\,79054, & \lambda_2(1) &= 0,\\
\tilde{F}_{c=0}(T) &= 44.70866\,79043.
\end{aligned}$$

Evaluating the Riccati equations (5.10) and (5.14) along this specific solution, we were not able to find a bounded solution satisfying the sign conditions $q_T(0) \geq 0$ and $q_T(1) < 0$. This fact confirms our impression gained from Figure 5.4 that no local minimum can be identified for $c = 0$.

Acknowledgments. We are indebted to D. Augustin, Ch. Beckmann, and J. Strade for numerical assistance with the examples. We would like to thank an anonymous reviewer for helpful remarks.

REFERENCES

[1] A. E. Bryson and Y. C. Ho, *Applied Optimal Control*, revised printing, Hemisphere, New York, 1975.

[2] Ch. Büskens and H. Maurer, *SQP-methods for solving optimal control problems with control and state constraints: Adjoint variables, sensitivity analysis and real-time control*, J. Comput. Appl. Math., 120 (2000), pp. 85–108.

[3] R. Bulirsch, *Die Mehrzielmethode zur numerischen Lösung von nichtlinearen Randwertproblemen und Aufgaben der optimalen Steuerung*, Report of the Carl-Cranz Gesellschaft, Oberpfaffenhofen, Germany, 1971.

[4] M. Chamberland and V. Zeidan, *Second order necessity and sufficiency theory for the free final time problem*, in Proceedings of the 31st IEEE Conference on Decision and Control, Tucson, AZ, 1992, pp. 1518–1525.

[5] A. L. Dontchev, W. W. Hager, A. B. Poore, and B. Yang, *Optimality, stability and convergence in nonlinear control*, Appl. Math. Optim., 31 (1995), pp. 297–326.

[6] J. C. Dunn, *Second-order optimality conditions in sets of L^∞ functions with range in a polyhedron*, SIAM J. Control Optim., 33 (1995), pp. 1603–1635.

[7] J. C. Dunn, *On L^2 sufficient conditions and the gradient projection method for optimal control problems*, SIAM J. Control Optim., 34 (1996), pp. 1270–1290.

[8] J. C. Dunn, *On second order sufficient conditions for structured nonlinear programs in infinite-dimensional function spaces*, in Mathematical Programming with Data Perturbations, A. V. Fiacco, ed., Lecture Notes in Pure and Appl. Math. 195, Marcel Dekker, New York, 1998, pp. 83–107.

[9] U. Felgenhauer, *Diskretisierung von Steuerungsproblemen unter stabilen Optimalitätsbedingungen*, Habilitationsschrift, Department of Mathematics, Technische Universität Cottbus, Cottbus, Germany, 1999.

[10] U. Felgenhauer, *On smoothness properties and approximability of optimal control functions*, Ann. Oper. Res., 101 (2001), pp. 23–42.

[11] M. Hestenes, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.

[12] D. G. Hull, *Sufficient conditions for a minimum of the free-final-time optimal control problem*, J. Optim. Theory Appl., 68 (1991), pp. 275–287.

[13] D. H. Jacobson and D. Q. Mayne, *Differential Dynamic Programming*, American Elsevier Publishing, New York, 1970.

[14] H. J. Kelley, *Gradient theory of optimal flight paths*, ARS Journal, 30 (1960), pp. 947–954.

[15] H. J. Kelley, *Method of gradients*, in Optimization Techniques, G. Leitmann, ed., Academic Press, New York, 1962, pp. 205–254.

[16] P. Kenneth and R. McGill, *Two-point boundary-value problem techniques*, in Advances in Control Systems, Vol. 3, C. T. Leondes, ed., Academic Press, New York, 1966, pp. 69–109.

[17] D. F. Lawden, *Optimal Trajectories for Space Navigation*, Academic Press, New York, 1967.

[18] K. Malanowski, *Two-norm approach in stability and sensitivity analysis of optimization and optimal control problems*, Adv. Math. Sci. Appl., 2 (1993), pp. 397–443.

[19] K. Malanowski, *Stability and sensitivity of solutions to nonlinear optimal control problems*, Appl. Math. Optim., 32 (1994), pp. 111–141.

[20] K. Malanowski, *Sufficient optimality conditions for optimal control subject to state constraints*, SIAM J. Control Optim., 35 (1997), pp. 205–227.

[21] K. Malanowski and H. Maurer, *Sensitivity analysis for parametric control problems with control-state constraints*, Comput. Optim. Appl., 5 (1996), pp. 253–283.

[22] H. Maurer, *Second order sufficient conditions for control problems with free final time*, in Proceedings of 3rd European Control Conference, A. Isidori et al., eds., Rome, Italy, 1995, pp. 3602–3606.

[23] H. Maurer and D. Augustin, *Second order sufficient conditions and sensitivity analysis for the controlled Rayleigh problem*, in Parametric Optimization and Related Topics IV, J. Guddat, H. Th. Jongen, F. Nozicka, G. Still, and F. Twilt, eds., Peter Lang Verlag, Frankfurt, Germany, 1996, pp. 245–259.

[24] H. Maurer and H. J. Pesch, *Solution differentiability for parametric nonlinear control problems with control-state constraints*, J. Optim. Theory Appl., 86 (1995), pp. 285–309.

[25] H. Maurer and S. Pickenhain, *Second order sufficient conditions for optimal control problems with mixed control-state constraints*, J. Optim. Theory Appl., 86 (1995), pp. 649–667.

[26] A. A. Milyutin and N. P. Osmolovskii, *Calculus of Variations and Optimal Control*, Transl. Math. Monogr. 180, AMS, Providence, RI, 1998.

[27] H. G. Moyer and G. Pinkham, *Several trajectory optimization techniques*. II. *Application*, in Computing Methods in Optimization Problems, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1964, pp. 91–105.

[28] L. W. Neustadt, *Optimization: A Theory of Necessary Conditions*, Princeton University Press, Princeton, NJ, 1976.

[29] H. J. Oberle and W. Grimm, *BNDSCO—A Program for the Numerical Solution of Optimal Control Problems*, Internal Report 515–89/22, Institute for Flight Systems Dynamics, DLR, Oberpfaffenhofen, Germany, 1989.

[30] H. J. Oberle and K. Taubert, *Existence and multiple solutions of the minimum-fuel orbit transfer problem*, J. Optim. Theory Appl., 95 (1997), pp. 241–262.

[31] N. P. Osmolovskii, *Quadratic conditions for nonsingular extremals in optimal control (a theoretical treatment)*, Russian J. Math. Phys., 2 (1995), pp. 487–516.

[32] N. P. Osmolovskii, *Second-order conditions for broken extremals*, in Proceedings of the Conference on Calculus of Variations and Optimal Control, A. Ioffe, S. Reich, and I. Shafir, eds., Chapman & Hall/CRC, Boca Raton, FL, 2000, pp. 198–216.

[33] H. J. Pesch, *Real-time computation of feedback controls for constrained optimal control problems*. I. *Neighbouring extremals*, Optimal Control Appl. Methods, 10 (1989), pp. 129–145.

[34] H. J. Pesch, *Real-time computation of feedback controls for constrained optimal control problems*. II. *A correction method based on multiple shooting*, Optimal Control Appl. Methods, 10 (1989), pp. 147–171.

[35] S. Pickenhain, *Sufficiency conditions for weak local minima in multidimensional optimal control problems with mixed control-state restrictions*, Z. Anal. Anwendungen, 11 (1992), pp. 559–568.

[36] T. Tun and T. S. Dillon, *Extensions of the differential dynamic programming method to include systems with state dependent control constraints and state variable inequality constraints*, J. Appl. Sci. Engrg. A, 3 (1978), pp. 171–192.

[37] V. Zeidan, *The Riccati equation for optimal control problems with mixed state-control constraints: Necessity and sufficiency*, SIAM J. Control Optim., 32 (1994), pp. 1297–1321.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 2, pp. 404–424

STOCHASTIC TARGET PROBLEMS, DYNAMIC PROGRAMMING, AND VISCOSITY SOLUTIONS*

H. METE SONER† AND NIZAR TOUZI‡

Abstract. In this paper, we define and study a new class of optimal stochastic control problems which is closely related to the theory of backward SDEs and forward-backward SDEs. The controlled process (X^ν, Y^ν) takes values in $\mathbb{R}^d \times \mathbb{R}$ and a given initial data for $X^\nu(0)$. Then the control problem is to find the minimal initial data for Y^ν so that it reaches a stochastic target at a specified terminal time T. The main application is from financial mathematics, in which the process X^ν is related to stock price, Y^ν is the wealth process, and ν is the portfolio.

We introduce a new dynamic programming principle and prove that the value function of the stochastic target problem is a discontinuous viscosity solution of the associated dynamic programming equation. The boundary conditions are also shown to solve a first order variational inequality in the discontinuous viscosity sense. This provides a unique characterization of the value function which is the minimal initial data for Y^ν.

Key words. stochastic control, dynamic programming, discontinuous viscosity solutions, forward-backward SDEs

AMS subject classifications. Primary, 49J20, 60J60; Secondary, 49L20, 35K55

PII. S0363012900378863

1. Introduction.

Let $(\Omega, \mathcal{F}, P)$ be a probability space, $T > 0$, and let $\{W(t), 0 \le t \le T\}$ be a d-dimensional Brownian motion whose P-completed natural filtration is denoted by $\mathbb{F}$. Given a control process $\nu = \{\nu(t),\ 0 \le t \le T\}$ with values in the control set $\mathcal{U}$, we consider the controlled process $Z_y^\nu = (X_y^\nu, Y_y^\nu) \in \mathbb{R}^d \times \mathbb{R}$ satisfying

$$dZ(t) = \alpha\left(t, Z(t), \nu(t)\right)\, dt + \beta\left(t, Z(t), \nu(t)\right)\, dW(t), \qquad 0 \le t < T, \tag{1.1}$$

together with the initial data $Z^\nu(0) = (X(0), y)$.

For a given real-valued function g, the stochastic target control problem is to minimize the initial data y while satisfying the random constraint $Y_y^\nu(T) \ge g(X_y^\nu(T))$ with probability one, i.e.,

$$v(0, X(0)) := \inf\left\{y \in \mathbb{R} \ :\ \exists\, \nu \in \mathcal{U},\ \ Y_y^\nu(T) \ge\ g\left(X_y^\nu(T)\right) \quad P - \text{a.s.}\right\},$$

which we call the stochastic target problem.

The chief goal of this paper is to obtain a characterization of the value function v as a discontinuous viscosity solution of an associated Hamilton–Jacobi–Bellman (HJB) second order PDE with suitable boundary conditions. We do not address the important uniqueness issue associated to the HJB equation in this paper. We simply refer to Crandall, Ishii, and Lions [5] for some general uniqueness results.

The main step in the derivation of the above-mentioned PDE characterization is a nonclassical dynamic programming principle. To the best of our knowledge, this

*Received by the editors September 26, 2000; accepted for publication (in revised form) November 29, 2001; published electronically June 18, 2002.
http://www.siam.org/journals/sicon/41-2/37886.html

†Department of Mathematics, Koç University, Rumelifeneri Yolu, Sariyer 80910, Istanbul, Turkey (msoner@ku.edu.tr). The work of this author was partially supported by National Science Foundation grant DMS-98-17525. Part of this work was completed during this author's visit to the Feza Gursey Institute for Basic Sciences in Istanbul.

‡CREST, 15 Bd Gabriel Peri, 92245 Malakoff, France (touzi@ensae.fr).

dynamic programming is new; it was only partially used by the authors in a previous paper [23].

This dynamic programming principle is closely related to the theory of viscosity solutions. In the derivation of the supersolution property of the HJB equation, the notion of viscosity solutions is only used to handle the lack of a priori regularity of the value function. However, the use of the notion of viscosity solutions seems necessary in order to derive the subsolution property from our dynamic programming principle, even if the value function were known to be smooth.

This study is mainly motivated by applications to financial mathematics. Indeed, a special specification of the coefficients α and β (see section 6) leads to the so-called superreplication problem; see, e.g., El Karoui and Quenez [11], Cvitanić and Karatzas [6], Broadie, Cvitanić, and Soner [4], Cvitanić, Pham, and Touzi [9], and Cvitanić and Ma [8].

In the financial mathematics literature, the superreplication problem is usually solved via convex duality. In this approach, a classical optimal control problem is derived by first applying the duality; see Jouini and Kallal [15], El Karoui and Quenez [11], Cvitanić and Karatzas [6], and Föllmer and Kramkov [13]. Then, one may use classical dynamic programming to obtain the PDE characterization of the value function v. However, this method cannot be applied to the general stochastic target problem because of the presence of the control ν in the diffusion part of the state process X^ν. The methodology developed in this paper precisely allows us to avoid this step and to obtain the PDE characterization directly from the initial (nonclassical) formulation of the problem without using the duality.

The stochastic target problem is also closely related to the theory of backward SDEs and forward-backward SDEs; see Antonelli [1], Cvitanić, Karatzas, and Soner [7], Hu and Peng [16], Ma, Protter, and Yong [18], Ma and Yong [19], Pardoux [20], and Pardoux and Tang [21]. Indeed, an alternative formulation of the problem is this: find a triple of $\mathbb{F}$-adapted processes (X, Y, ν) satisfying

$$(1.2)\quad (X,Y)\ \text{ solves }\ (1.1)\ \text{ with }\ \nu \in \mathcal{U}, X(0)\ \text{ fixed, }\ Y(T) + A(T) = g(X(T))$$

for some nondecreasing $\mathbb{F}$-adapted process A with $A(0) = 0$ as well as the minimality condition

$$(\tilde{X}, \tilde{Y}, \tilde{\nu}, \tilde{A})\ \text{ satisfies }\ (1.2) \Longrightarrow Y(.)\ \le\ \tilde{Y}(.) \quad P - \text{a.s.}$$

Notice that the nondecreasing process A is involved in the above definition to account for possible constraints on the control ν; see [7]. In financial applications, this connection has been observed by Cvitanić and Ma [8] and El Karoui, Peng, and Quenez [12].

The paper is organized as follows: the definition of the stochastic target problem is formulated in section 2. In section 3, we state the dynamic programming principle. Section 4 studies the HJB equation satisfied by the value function v in the discontinuous viscosity sense. In section 5, the terminal condition of the problem is characterized by a first order variational inequality again in the discontinuous viscosity sense. Finally, in section 6, we apply our results to the problem of superreplication under portfolio constraints in a large investor financial market.

2. Stochastic target problem. In this section, we define a nonstandard stochastic control problem.

Let $T > 0$ be the finite time horizon, and let $W = \{W(t),\ 0 \le t \le T\}$ be a d-dimensional Brownian motion defined on a complete probability space $(\Omega, \mathcal{F}, P)$.

We denote by $\mathbb{F} = \{\mathcal{F}(t),\, 0 \le t \le T\}$ the P-augmentation of the filtration generated by W.

We assume that the control set U is a convex compact subset of $\mathbb{R}^d$ with a nonempty interior, and we denote by $\mathcal{U}$ the set of all progressively measurable processes $\nu = \{\nu(t),\, 0 \le t \le T\}$ with values in U.

The state process is defined as follows: given the initial datum $z = (x, y) \in \mathbb{R}^d \times \mathbb{R}$, an initial time $t \in [0, T]$, and a control process $\nu \in \mathcal{U}$, let the controlled process $Z^\nu_{t,z} = (X^\nu_{t,x}, Y^\nu_{t,z})$ be the solution of the SDE

$$dX^\nu_{t,x}(u) \;=\; \mu\left(u, X^\nu_{t,x}(u), \nu(u)\right) du + \sigma^*\left(u, X^\nu_{t,x}(u), \nu(u)\right) dW(u), \qquad u \in (t, T),$$

$$dY^\nu_{t,x,y}(u) \;=\; b\left(u, Z^\nu_{t,z}(u), \nu(u)\right) du + a^*\left(u, Z^\nu_{t,z}(u), \nu(u)\right) dW(u), \qquad u \in (t, T),$$

with initial data

$$X^\nu_{t,x}(t) \;=\; x, \qquad\qquad Y^\nu_{t,x,y}(t) \;=\; y,$$

where M^* denotes the transpose of the matrix M, and μ, σ, b, a are bounded functions on $[0, T] \times \mathbb{R}^k \times U$ ($k = d$ or $d + 1$) satisfying the usual conditions in order for the process $Z^\nu_{t,z}$ to be well defined.

Throughout the paper, we assume that the matrix $\sigma(t, x, r)$ is invertible and the function

$$r \mapsto \sigma^{-1}(t, x, r) a(t, x, y, r)$$

is one to one for all (t, x, y). Let ψ be its inverse; i.e.,

$$\sigma^{-1}(t, x, r) a(t, x, y, r) \;=\; p \quad \Longleftrightarrow \quad r \;=\; \psi(t, x, y, p). \tag{2.1}$$

This is a crucial assumption which enables us to match the stochastic parts of the X and the Y processes by a judicial choice of the control process ν. Similar assumptions were also utilized in the backward-forward SDEs. See also Remark 2.2.

Now we are in a position to define the "stochastic target" control problem. Let g be a real-valued measurable function defined on $\mathbb{R}^d$. We shall denote by $\mathcal{E}pi(g) := \{(x, y) \in \mathbb{R}^d \times \mathbb{R} :\, y \ge g(x)\}$ the epigraph of g. Let

$$v(t, x) := \inf\left\{y \in \mathbb{R} \;:\; \exists\, \nu \in \mathcal{U}\,,\; Z^\nu_{t,x,y}(T) \;\in\; \mathcal{E}pi(g) \;\; P - \text{a.s.}\right\}. \tag{2.2}$$

In some cases, it is possible to find initial datum and a control so that $Y^\nu_{t,x,y}(T) = g(X^\nu_{t,x}(T))$. In that case, this problem is equivalent to a backward-forward SDE; see the discussion in our introduction. In particular, when $U = \mathbb{R}^d$, the corresponding backward-forward SDE has a solution (see, e.g., [21]), and it is equal to v. However, when the control set U is bounded, in general there is no solution of the backward-forward equation, and v is the natural generalization of the backward-forward SDE. An alternative generalization can be obtained by involving a nondecreasing process, as discussed in the introduction; see [7].

We conclude this section by introducing several sets to simplify the notation. Let

$$\mathcal{A}(t, x, y) := \left\{\nu \in \mathcal{U} \;:\; Z^\nu_{t,x,y}(T) \;\in\; \mathcal{E}pi(g) \;\; P - \text{a.s.}\right\}.$$

Note that $\mathcal{A}(t, x, y)$ may be empty for some initial datum (t, x, y). Next we define

$$\mathcal{Y}(t, x) := \left\{y \in \mathbb{R} \;:\; \mathcal{A}(t, x, y) \neq \emptyset\right\}.$$

Then the stochastic target problem can be written as

$$v(t,x) \;=\; \inf \mathcal{Y}(t,x) = \inf \left\{ y \in \mathbb{R} \;:\; \; y \in \mathcal{Y}(t,x) \right\}.$$

Remark 2.1. The set $\mathcal{Y}(t,x)$ satisfies the following important property:

$$\text{for all } \; y \in \mathbb{R}, \;\; y \in \mathcal{Y}(t,x) \implies [y,\infty) \subset \mathcal{Y}(t,x).$$

This follows from the facts that $X^{\nu}_{t,x}$ is independent of y and $Y^{\nu}_{t,x,y}(T)$ is nondecreasing in y.

Remark 2.2. A more general formulation of this problem, as discussed in our accompanying paper [24], is obtained by defining the *reachability set* of the deterministic target $\mathcal{E}pi(g)$:

$$V(t) := \left\{ z \in \mathbb{R}^{d+1} \;:\; Z^{\nu}_{t,z}(T) \in \mathcal{E}pi(g) \; P-\text{a.s. for some } \nu \in \mathcal{A} \right\}.$$

From the previous remark, the set $V(t)$ is "essentially" characterized as the epigraph of the scalar function $v(t,.)$. A standing assumption in [24] is

$$\mathcal{N}(t,z,p) := \left\{ \nu \in \mathbb{R}^d \;:\; [\sigma|a](t,z,\nu) \begin{bmatrix} p \\ -1 \end{bmatrix} = 0 \right\} \neq \emptyset;$$

i.e., since we wish to hit the deterministic target $\mathcal{E}pi(g)$ with probability one, the diffusion process has to degenerate along certain directions captured by the kernel $\mathcal{N}$. This degeneracy assumption is directly related to our condition (2.1).

3. Dynamic programming. In this section, we introduce a new dynamic programming equation for the stochastic target problem. This will allow us to characterize the value function of the stochastic target problem as a viscosity solution of a nonlinear PDE. For the classical stochastic control problem, this connection between the dynamic programming principle and the PDEs is well known (see, e.g., [14]). The chief goal of this paper is to develop the same tools for this nonstandard target control problem. Namely, we will formulate an appropriate dynamic programming principle and then derive the corresponding nonlinear PDE as a consequence of it.

A discussion of general dynamic programming of this type is the subject of an accompanying paper by the authors [24].

THEOREM 3.1. *Let* $(t,x) \in [0,T] \times \mathbb{R}^d$.

(DP1) *For any* $y \in \mathbb{R}$, *set* $z := (x,y)$. *Suppose that* $\mathcal{A}(t,z) \neq \emptyset$. *Then, for all* $\nu \in \mathcal{A}(t,z)$ *and a* $[t,T]$*-valued stopping time* θ,

$$Y^{\nu}_{t,x,y}(\theta) \geq v\left(\theta, X^{\nu}_{t,x}(\theta)\right) \qquad P-a.s.$$

(DP2) *Set* $y^* := v(t,x)$. *Let* θ *be an arbitrary* $[t,T]$*-valued stopping time. Then, for all* $\nu \in \mathcal{U}$ *and* $\eta > 0$,

$$P\left[Y^{\nu}_{t,x,y^*-\eta}(\theta) \;>\; v\left(\theta, X^{\nu}_{t,x}(\theta)\right) \right] < 1.$$

Proof. We provide only the main idea of the proof. We refer to [24] for the complete argument. Let $z = (x,y)$ and ν be as in the statement of (DP1). By the definition of $\mathcal{A}(t,z)$, $Z^{\nu}_{t,z}(T) \in \mathcal{E}pi(g)$. Since $Z^{\nu}_{t,z}(T) = Z^{\nu}_{\theta, Z^{\nu}_{t,z}(\theta)}(T)$, it follows that

$$\nu(\cdot) \in \mathcal{A}\left(\theta(w), Z^{\nu}_{t,z}(t+\theta(w))\right) \qquad \text{for } P \text{ almost every } w \in \Omega\,.$$

Then, again for P almost every $w \in \Omega$, $Y^\nu_{t,z}(\theta(w)) \in \mathcal{Y}\big(\theta(w), X^\nu_{t,x}(\theta(w))\big)$, and, by the definition of the value function, $v\big(\theta(w), X^\nu_{t,x}(\theta(w))\big) \le Y^\nu_{t,z}(\theta(w))$.

We prove (DP2) by contraposition. So, toward a contradiction, suppose that there exists a $[t,T)$-valued stopping time θ such that

$$Y^\nu_{t,x,y^*-\eta}(\theta) \;>\; v\big(\theta, X^\nu_{t,x}(\theta)\big) \qquad P-\text{a.s.}$$

In view of Remark 2.1, this proves that $Y^\nu_{t,x,y^*-\eta}(\theta) \in \mathcal{Y}\big(\theta, X^\nu_{t,x}(\theta)\big)$. Then there exists a control $\hat\nu \in \mathcal{U}$ such that

$$Y^{\hat\nu}_{\theta, Z^\nu_{t,x,y^*-\eta}(\theta)}(T) \ge g\big(X^{\hat\nu}_{\theta, X^\nu_{t,x}(\theta)}(T)\big) \qquad P-\text{a.s.}$$

Since the process $(X^{\hat\nu}_{\theta,X^\nu_{t,x}(\theta)}, Y^{\hat\nu}_{\theta,Z^\nu_{t,x,y^*-\eta}(\theta)})$ depends on $\hat\nu$ only through its realizations in the stochastic interval $[t,\theta]$, we may chose $\hat\nu$ so that $\hat\nu = \nu$ on $[t,\theta]$. (This is the difficult part of this proof.) Then $Z^{\hat\nu}_{\theta,Z^\nu_{t,x,y^*-\eta}(\theta)}(T) = Z^{\hat\nu}_{t,x,y^*-\eta}(T)$, and therefore $y^*-\eta \in \mathcal{Y}(t,x)$; hence $y^*-\eta \le v(t,x)$. Recall that, by definition, $y^* = v(t,x)$ and $\eta > 0$. □

The dynamic programming principle stated in Theorem 3.1 does not require all of the assumptions made in the first section. Namely, the control set U does not need to be convex or compact, and the function $\sigma^{-1}(t,x,r)a(t,x,y,r)$ is not required to be one to one in the r variable.

For completeness, we mention that the statement of Theorem 3.1 is equivalent to the following, apparently stronger but more natural, dynamic programming principle.

COROLLARY 3.1. *For all $(t,x) \in [0,T)\times\mathbb{R}^d$ and a $[t,T]$-valued stopping time θ, we have*

$$v(t,x) = \inf\big\{y\in\mathbb{R} \;:\; \exists\,\nu\in\mathcal{U},\; Y^\nu_{t,x,y}(\theta) \;\ge\; v\big(\theta, X^\nu_{t,x}(\theta)\big)\;\; P-a.s.\big\}.$$

4. Viscosity property. In this section, we use the dynamic programming principle stated in Theorem 3.1 to prove that the value function of the stochastic target control problem (2.2) is a discontinuous viscosity solution to the corresponding dynamic programming equation.

Following the convention in the viscosity literature, let v_* (resp., v^*) be the lower (resp., upper) semicontinuous envelope of v; i.e.,

$$v_*(t,x) \;:=\; \liminf_{(t',x')\to(t,x)} v(t',x') \quad\text{and}\quad v^*(t,x) \;:=\; \limsup_{(t',x')\to(t,x)} v(t',x').$$

Let δ_U be the support function of the closed convex set U:

$$\delta_U(\zeta) := \sup_{\nu\in U}(\nu^*\zeta), \qquad \zeta\in\mathbb{R}^d.$$

We shall denote by $\tilde U$ the effective domain of δ_U and by $\tilde U_1$ the restriction of $\tilde U$ to the unit circle:

$$\tilde U \;=\; \big\{\zeta\in\mathbb{R}^d \;:\; \delta_U(\zeta)\in\mathbb{R}\big\} \text{ and } \tilde U_1 \;=\; \big\{\zeta\in\tilde U \;:\; |\zeta|=1\big\}$$

so that $\tilde U$ is the closed cone generated by $\tilde U_1$. Under our assumptions, since U is a bounded subset of $\mathbb{R}^d$,

$$\tilde U \;=\; \mathbb{R}^d \text{ and } \tilde U_1 \;=\; \big\{\zeta\in\mathbb{R}^d \;:\; |\zeta|=1\big\}.$$

Remark 4.1. The compactness of U is only needed in order to establish some results which require us to extract convergent subsequences from sequences in U. Therefore, many results contained in this paper hold for a general closed convex subset U. For this reason, we shall keep using the notation $\tilde{U}$ and $\tilde{U}_1$.

Remark 4.2. For later reference, note that the closed convex set U can be characterized in terms of $\tilde{U}$ (see, e.g., [22]):

$$\begin{aligned}\nu \in U \text{ iff } \inf_{\zeta\in\tilde{U}} (\delta_U(\zeta) - \zeta^*\nu) \ &\geq\ 0,\\ \text{iff } \inf_{\zeta\in\tilde{U}_1} (\delta_U(\zeta) - \zeta^*\nu) \ &\geq\ 0;\end{aligned}$$

the second characterization follows from the facts that $\tilde{U}$ is the closed cone generated by $\tilde{U}_1$ and δ_U is positively homogeneous.

Remark 4.3. We shall also use the following characterization of $\mathrm{int}(U)$ in terms of $\tilde{U}_1$:

$$\nu \in \mathrm{int}(U) \text{ iff } \inf_{\zeta\in\tilde{U}_1} (\delta_U(\zeta) - \zeta^*\nu) \ >\ 0.$$

To see this, suppose that the right-hand side infimum is zero. Then, for all $\varepsilon > 0$, there exists some $\zeta_0 \in \tilde{U}_1$ such that $0 \leq \delta_U(\zeta_0) - \zeta_0^*\nu \leq \varepsilon/2$. Then $\delta_U(\zeta_0) - \zeta_0^*(\nu + \varepsilon\zeta_0) < 0$, and therefore $\nu + \varepsilon\zeta_0 \notin U$ by the previous remark. Since $\varepsilon > 0$ is arbitrary, this proves that $\nu \notin \mathrm{int}(U)$. Conversely, suppose that $\ell := \inf_{\zeta\in\tilde{U}_1} (\delta_U(\zeta) - \zeta^*\nu) > 0$. Then, by the Cauchy–Schwarz inequality and the characterization of the previous remark, it is easily checked that the ball around ν with radius ℓ is included in U.

Remark 4.4. Let f be the function defined on $\mathbb{R}^d$ by

$$f(\nu) := \inf_{\zeta\in\tilde{U}_1} (\delta_U(\zeta) - \zeta^*\nu)\,.$$

Then f is continuous. Indeed, since $\tilde{U}_1$ is a compact subset of $\mathbb{R}^d$, the infimum in the above definition of $f(\nu)$ is attained, say, at $\hat{\zeta}(\nu) \in \tilde{U}_1$. Then, for all $\nu, \nu' \in \mathbb{R}^d$,

$$f(\nu') \leq \delta_U(\hat{\zeta}(\nu)) - \hat{\zeta}(\nu)^*\nu + \hat{\zeta}(\nu)^*(\nu - \nu') \ =\ f(\nu) + \hat{\zeta}(\nu)^*(\nu - \nu') \ \leq\ f(\nu) + |\nu - \nu'|$$

by the Cauchy–Schwarz inequality. By symmetry, this proves that f is a contracting mapping.

Finally, we introduce the Dynkin second order differential operator associated to the process X^ν:

$$\mathcal{L}^\nu u(t,x) := \frac{\partial u}{\partial t}(t,x) + \mu(t,x,\nu)^* Du(t,x) + \frac{1}{2}\mathrm{Trace}\left(D^2u(t,x)\sigma^*(t,x,\nu)\sigma(t,x,\nu)\right),$$

where Du and D^2u denote, respectively, the gradient and the Hessian matrix of u with respect to the x variable.

THEOREM 4.1. *Assume that μ, σ, a, b are all bounded and satisfy the usual Lipschitz conditions and that v^*, v_* are finite everywhere. Further assume* (2.1) *and that U has a nonempty interior. Then the value function v of the stochastic target problem is a discontinuous viscosity solution of the equation on $[0,T) \times \mathbb{R}^d$,*

$$\min\left\{-\mathcal{L}^{\nu_0}u(t,x) + b\left(t,x,u(t,x),\nu_0\right);\ H(t,x,u(t,x),Du(t,x))\right\} = 0, \tag{4.1}$$

where

$$\nu_0(t,x) := \psi\left(t,x,u(t,x),Du(t,x)\right), \tag{4.2}$$

$$H(t,x,u(t,x),Du(t,x)) = \inf_{\zeta\in\tilde{U}_1}\left(\delta_U(\zeta) - \zeta^*\nu_0(t,x)\right); \tag{4.3}$$

i.e., v_* and v^* are, respectively, viscosity supersolution and subsolution of (4.1).

Remark 4.5. In view of Remark 4.2, $H \geq 0$ iff $\nu_0 \in U$. Since U has a nonempty interior, it follows from Remark 4.3 that $H > 0$ iff $\nu_0 \in \text{int}(U)$.

The proof of Theorem 4.1 will be completed in the following two subsections. The supersolution part of the claim follows from (DP1) and a classical argument in the viscosity theory which is due to P.-L. Lions. We shall take advantage of the fact that the inequality in (DP1) is in the a.s. sense. This allows for suitable change of measure before taking expectations. The subsolution part is obtained from (DP2) by means of a contraposition argument.

The above result will be completed in Theorem 5.1 by the description of the boundary condition. The reader who is not interested in the technical proof of Theorem 4.1 can go directly to section 5.

4.1. Proof of the viscosity supersolution property. Fix $(t_0,x_0) \in [0,T)\times\mathbb{R}^d$, and let φ be a $C^2([0,T]\times\mathbb{R}^d)$ function satisfying

$$0 \;=\; (v_*-\varphi)(t_0,x_0) = \min_{(t,x)\in[0,T)\times\mathbb{R}^d}(v_*-\varphi)\,.$$

Observe that $v \geq v_* \geq \varphi$ on $[0,T)\times\mathbb{R}^d$.

Step 1. Let $(t_n,x_n)_{n\geq 1}$ be a sequence in $[0,T)\times\mathbb{R}^d$ such that

$$(t_n,x_n) \;\to\; (t_0,x_0) \text{ and } v(t_n,x_n) \;\to\; v_*(t_0,x_0).$$

Set $y_n := v(t_n,x_n)+(1/n)$ and $z_n := (x_n,y_n)$. Then, by the definition of the stochastic target control problem, the set $\mathcal{A}(t_n,z_n)$ is not empty. Let ν_n be any element of $\mathcal{A}(t_n,z_n)$.

For any $[0,T-t_n)$-valued stopping time θ_n (to be chosen later), (DP1) yields

$$Y^{\nu_n}_{t_n,z_n}(t_n+\theta_n) \geq v\left(t_n+\theta_n, X_{t_n,x_n}(t_n+\theta_n)\right) \qquad P-\text{a.s.}$$

Set $\beta_n := y_n - \varphi(t_n,x_n)$. Since, as n tends to infinity, $y_n \to v_*(t_0,x_0)$ and $\varphi(t_n,x_n) \to \varphi(t_0,x_0) = v_*(t_0,x_0)$,

$$\beta_n \to 0\,.$$

Further, since $v \geq v_* \geq \varphi$, we have $v\left(t_n+\theta_n, X_{t_n,x_n}(t_n+\theta_n)\right) \geq \varphi(t_n+\theta_n, X_{t_n,x_n}(t_n+\theta_n))$ P-a.s. Then

$$\beta_n + \left[Y^{\nu_n}_{t_n,z_n}(t_n+\theta_n) - y_n\right] - \left[\varphi\left(t_n+\theta_n, X_{t_n,x_n}(t_n+\theta_n)\right) - \varphi(t_n,x_n)\right] \geq 0 \qquad P-\text{a.s.}$$

By Itô's lemma,

$$\begin{aligned} 0 \leq \beta_n \;+\; & \int_{t_n}^{t_n+\theta_n}\left[b\left(s,Z^{\nu_n}_{t_n,z_n}(s),\nu_n(s)\right) - \mathcal{L}^{\nu_n(s)}\varphi\left(s,X^{\nu_n}_{t_n,x_n}(s)\right)\right]ds \\ & + \int_{t_n}^{t_n+\theta_n}\Big[a\left(s,Z^{\nu_n}_{t_n,z_n}(s),\nu_n(s)\right) \\ & \qquad - \sigma\left(s,X^{\nu_n}_{t_n,x_n}(s),\nu_n(s)\right)D\varphi\left(s,X^{\nu_n}_{t_n,x_n}(s)\right)\Big]^* dW(s). \end{aligned} \tag{4.4}$$

Step 2. For some large constant C, set

$$\theta_n := \inf\left\{s > t_n \ : \ |X^{\nu_n}_{t_n,x_n}(s)| \geq C\right\}.$$

Since U is bounded in $\mathbb{R}^d$ and $(t_n, x_n) \longrightarrow (t_0, x_0)$, one can easily show that

$$\liminf_{n\to\infty} t \wedge \theta_n \ > \ t_0 \text{ for all } t > t_0. \tag{4.5}$$

For $\xi \in \mathbb{R}$, we introduce the probability measure P^ξ_n equivalent to P defined by the density process

$$M^\xi_n(t) := \mathcal{E}\left(-\xi \int_{t_n}^{t\wedge\theta_n} (a - \sigma D\varphi)\left(s, Z^{\nu_n}_{t_n,z_n}(s), \nu_n(s)\right) dW(s)\right), \quad t \geq t_n \ ,$$

where $\mathcal{E}(.)$ is the Doléans–Dade exponential operator. We shall denote by E^ξ_n the conditional expectation with respect to $\mathcal{F}_{t_n}$ under P^ξ_n.

We take the conditional expectation with respect to $\mathcal{F}_{t_n}$ under P^ξ_n in (4.4). The result is

$$\begin{aligned} 0 \leq \beta_n + E^\xi_n \left[\int_{t_n}^{t_n+h\wedge\theta_n} \left(b\left(s, Z^{\nu_n}_{t_n,z_n}(s), \nu_n(s)\right) - \mathcal{L}^{\nu_n(s)}\varphi\left(s, X^{\nu_n}_{t_n,x_n}(s)\right)\right) ds\right] \\ - \xi\, E^\xi_n \left[\int_{t_n}^{t_n+h\wedge\theta_n} \Big|a\left(s, Z^{\nu_n}_{t_n,z_n}(s), \nu_n(s)\right) \right. \\ \left. - \sigma\left(s, X^{\nu_n}_{t_n,x_n}(s), \nu_n(s)\right) D\varphi\left(s, X^{\nu_n}_{t_n,x_n}(s)\right)\Big|^2 ds\right] \end{aligned}$$

for all $h > 0$. We now consider two cases:

- Suppose that the set $\{n \geq 1 : \beta_n = 0\}$ is finite. Then there exists a subsequence, renamed $(\beta_n)_{n\geq 1}$, such that $\beta_n \neq 0$ for all $n \geq 1$. Set $h_n = \sqrt{|\beta_n|}$ and $k_n := \theta_n \wedge (t_n + h_n)$.
- If the set $\{n \geq 1 : \beta_n = 0\}$ is not finite, then there exists a subsequence, renamed $(\beta_n)_{n\geq 1}$, such that $\beta_n = 0$ for all $n \geq 1$. Set $h_n := n^{-1}$ and $k_n := \theta_n \wedge (t_n + h_n)$.

The final inequality still holds if we replace $t \wedge \theta_n$ with k_n. We then divide this inequality by h_n and send n to infinity by using (4.5), the dominated convergence theorem, and the right continuity of the filtration. The result is

$$\begin{aligned} 0 \leq \liminf_{n\to\infty} \frac{1}{h_n} \int_{t_n}^{t_n+h_n} \big[b\left(s, Z^{\nu_n}_{t_n,z_n}(s), \nu_n(s)\right) - \mathcal{L}^{\nu_n(s)}\varphi\left(s, X^{\nu_n}_{t_n,x_n}(s)\right) \\ - \xi\left|a\left(s, Z^{\nu_n}_{t_n,z_n}(s), \nu_n(s)\right) - \sigma\left(s, X^{\nu_n}_{t_n,x_n}(s), \nu_n(s)\right) D\varphi\left(s, X^{\nu_n}_{t_n,x_n}(s)\right)\right|^2\big] ds. \end{aligned}$$

We continue by using the following result, whose proof is given after the proof of the supersolution property.

LEMMA 4.1. *Let $\psi : [0,T] \times \mathbb{R}^{d+1} \times U \to \mathbb{R}$ be locally Lipschitz in (t,z) uniformly in r. Then*

$$\frac{1}{h_n} \int_{t_n}^{t_n+h_n} \left[\psi\left(s, Z^{\nu_n}_{t_n,z_n}(s), \nu_n(s)\right) - \psi\left(t_0, z_0, \nu_n(s)\right)\right] ds \to 0 \qquad P - a.s.$$

along some subsequence.

In view of this lemma,

$$0 \le \liminf_{n\to\infty} \frac{1}{h_n}\int_{t_n}^{t_n+h_n} \big[b\left(t_0,z_0,\nu_n(s)\right) - \mathcal{L}^{\nu_n(s)}\varphi\left(t_0,x_0\right) \\ - \xi\left|a\left(t_0,z_0,\nu_n(s)\right) - \sigma\left(t_0,x_0,\nu_n(s)\right)D\varphi\left(t_0,x_0\right)\right|^2\big]ds.$$

Then, since $h_n^{-1}\int_{t_n}^{t_n+h_n} ds = 1$,

$$\frac{1}{h_n}\int_{t_n}^{t_n+h_n} \big[b\left(t_0,z_0,\nu_n(s)\right) - \mathcal{L}^{\nu_n(s)}\varphi\left(t_0,x_0\right) \\ - \xi\left|a\left(t_0,z_0,\nu_n(s)\right) - \sigma\left(t_0,x_0,\nu_n(s)\right)D\varphi\left(t_0,x_0\right)\right|^2\big]ds \in \bar{\mathrm{co}}\mathcal{V}(t_0,z_0), \tag{4.6}$$

where $\bar{\mathrm{co}}\mathcal{V}(t_0,z_0)$ is the closed convex hull of the set $\mathcal{V}(t_0,z_0)$ defined by

$$\mathcal{V}(t_0,z_0) := \big\{b(t_0,z_0,\nu) - \mathcal{L}^{\nu}\varphi(t_0,x_0) - \xi\left|a(t_0,z_0,\nu) - \sigma(t_0,x_0,\nu)D\varphi(t_0,x_0)\right|^2 \ : \ \nu\in U\big\}.$$

Therefore, it follows from (4.6) that

$$\begin{aligned} 0 &\le \sup_{\phi\in\bar{\mathrm{co}}\mathcal{V}} \phi \\ &= \sup_{\nu\in U}\big\{\xi\left|-a(t_0,z_0,\nu) + \sigma(t_0,x_0,\nu)D\varphi(t_0,x_0)\right|^2 - \mathcal{L}^{\nu}\varphi(t_0,x_0) + b(t_0,z_0,\nu)\big\} \end{aligned} \tag{4.7}$$

for all $\xi\in\mathbb{R}$.

Step 3. For a large positive integer n, set $\xi = -n$. Since U is compact, the supremum in (4.7) is attained at some $\hat{\nu}_n \in U$, and

$$-n\left|a(t_0,z_0,\hat{\nu}_n) - \sigma(t_0,x_0,\hat{\nu}_n)D\varphi(t_0,x_0)\right|^2 - \mathcal{L}^{\hat{\nu}_n}\varphi(t_0,x_0) + b(t_0,z_0,\hat{\nu}_n) \ge 0.$$

By passing to a subsequence, we may assume that there exists $\hat{\nu}\in U$ such that $\hat{\nu}_n \to \nu_0$. Now let n to infinity in the last inequality to prove that

$$\left|a(t_0,z_0,\hat{\nu}_n) - \sigma(t_0,x_0,\hat{\nu}_n)D\varphi(t_0,x_0)\right|^2 \ \to \ 0 \tag{4.8}$$

and

$$-\mathcal{L}^{\nu_0}\varphi(t_0,x_0) + b(t_0,z_0,\nu_0) \ge 0. \tag{4.9}$$

In view of (4.8), we conclude that

$$\nu_0 = \psi\left(t_0,z_0,D\varphi(t_0,x_0)\right). \tag{4.10}$$

Since $\nu_0\in U$, it follows from Remark 4.2 that

$$\inf_{\zeta\in\tilde{U}_1}\left(\delta_U(\zeta) - \zeta^*\nu_0\right) \ge 0. \tag{4.11}$$

The supersolution property now follows from (4.9), (4.10), and (4.11). □

Proof of Lemma 4.1. Since $\psi(t,z,r)$ is locally Lipschitz in (t,z) uniformly in r,

$$\begin{aligned} &\frac{1}{h_n}\int_{t_n}^{t_n+h_n}\left[\psi\left(s,Z^{\nu_n}_{t_n,z_n}(s),\nu_n(s)\right) - \psi\left(t_0,z_0,\nu_n(s)\right)\right]ds \\ &\quad\le K\ \frac{1}{h_n}\int_{t_n}^{t_n+h_n}\left(|s-t_0| + \left|Z^{\nu_n}_{t_n,z_n}(s) - z_0\right|\right)ds \\ &\quad\le K\ \left(h_n + |t_n - t_0| + \sup_{t_n\le s\le t_n+h_n}\left|Z^{\nu_n}_{t_n,z_n}(s) - z_0\right|\right) \end{aligned}$$

for some constant K. Thus, to complete the proof of this lemma, it suffices to show

$$\sup_{t_n \le s \le t_n + h_n} \left| Z^{\nu_n}_{t_n, z_n}(s) - z_0 \right| \longrightarrow 0 \qquad P - \text{a.s.}$$

along a subsequence. Set

$$\gamma(t,x,y,r) := \begin{pmatrix} \mu(t,x,r) \\ b(t,x,y,r) \end{pmatrix} \text{ and } \alpha(t,x,y,r) := \begin{pmatrix} \sigma^*(t,x,r) \\ a^*(t,x,y,r) \end{pmatrix}.$$

Functions α and γ inherit the pointwise bounds from μ, b, σ, and a. We directly calculate that, for $t_n \le s \le t_n + h_n$,

$$Z^{\nu_n}_{t_n, z_n}(s) - z_0 \le |z_n - z_0| + \|\gamma\|_\infty h_n + \left| \int_{t_n}^{s} \alpha\left(s, Z^{\nu_n}_{t_n, z_n}(s), \nu_n(s)\right) dW(s) \right|,$$

and, therefore,

$$\begin{aligned} \sup_{t_n \le s \le t_n + h_n} \left| Z^{\nu_n}_{t_n, z_n}(s) - z_0 \right| \le |z_n - z_0| + \|\gamma\|_\infty h_n \\ + \sup_{t_n \le s \le t_n + h_n} \left| \int_{t_n}^{s} \alpha\left(s, Z^{\nu_n}_{t_n, z_n}(s), \nu_n(s)\right) dW(s) \right|. \end{aligned}$$

The first two terms on the right-hand side converge to zero. We estimate the third term by Doob's maximal inequality for submartingales.

The result is

$$\begin{aligned} E\left[\left(\sup_{t_n \le s \le t_n + h_n} \left| \int_{t_n}^{s} \alpha\left(s, Z^{\nu_n}_{t_n, z_n}(s), \nu_n(s)\right) dW(s) \right| \right)^2 \right] \\ \le 4\, E\left[\int_{t_n}^{t_n + h_n} \alpha\left(s, Z^{\nu_n}_{t_n, z_n}(s), \nu_n(s)\right)^2 ds \right] \\ \le 4\, \|\alpha\|^2_\infty h_n. \end{aligned}$$

This proves that

$$\sup_{t_n \le s \le t_n + h_n} \left| Z^{\nu_n}_{t_n, z_n}(s) - z_0 \right| \to 0 \quad \text{ in } L^2(P),$$

and, therefore, it also converges P-a.s. along some subsequence. □

4.2. Subsolution property. We start with a technical lemma which will be used both in the proof of the subsolution property and also in the next section on the characterization of the terminal data. We first introduce some notation. Given a smooth function $\varphi(t,x)$, we define the open subset of $[0,T] \times \mathbb{R}^d$:

$$\begin{aligned} \mathcal{M}_0(\varphi) := \Big\{ (t,x) \ : \ \inf_{\zeta \in \tilde{U}_1} \left(\delta_U(\zeta) - \zeta^* \nu_0(t,x) \right) > 0 \text{ and} \\ -\mathcal{L}^{\nu_0(t,x)} \varphi(t,x) + b\left(t,x,\varphi(t,x),\nu_0(t,x)\right) > 0 \Big\}, \\ = \{ (t,x) \ : \ \nu_0(t,x) \in \text{int}(U) \text{ and } - \mathcal{L}^{\nu_0(t,x)} \varphi(t,x) \\ + b\left(t,x,\varphi(t,x),\nu_0(t,x)\right) > 0 \}, \\ \text{where} \quad \nu_0(t,x) = \psi\left(t,x,\varphi(t,x),D\varphi(t,x)\right). \end{aligned}$$

LEMMA 4.2. *Let φ be a smooth test function, and let $B = B_R(x_0)$ be the open ball around x_0 with radius $R > 0$. Suppose that there are $t_1 < t_2 \le T$ such that*

$$\mathrm{cl}(\mathcal{M}) \subset \mathcal{M}_0(\varphi), \text{ where } \mathcal{M} := (t_1, t_2) \times B.$$

Then

$$\sup_{\partial_p \mathcal{M}} (v - \varphi) = \max_{\mathrm{cl}(\mathcal{M})} (v^* - \varphi),$$

where $\partial_p \mathcal{M}$ is the parabolic boundary of $\mathcal{M}$; i.e., $\partial_p \mathcal{M} = ([t_1, t_2] \times \partial B) \cup \big(\{t_2\} \times \bar{B}\big)$.

Proof. We shall denote $\overline{\mathcal{M}} := \mathrm{cl}(\mathcal{M})$. Suppose, to the contrary, that

$$\max_{\overline{\mathcal{M}}} (v^* - \varphi) - \sup_{\partial_p \mathcal{M}} (v - \varphi) := 2\beta > 0,$$

and let us work toward a contradiction of (DP2).

Choose $(t_0, x_0) \in \mathcal{M}$ so that $(v - \varphi)(t_0, x_0) \ge -\beta + \max_{\overline{\mathcal{M}}}(v^* - \varphi)$, and

$$(v - \varphi)(t_0, x_0) \ge \beta + \sup_{\partial_p \mathcal{M}} (v - \varphi). \tag{4.12}$$

Step 1. In view of Remark 4.5, $\inf_{\zeta \in \tilde{U}_1} (\delta_U(\zeta) - \zeta^* \nu_0) > 0$ is equivalent to $\nu_0 \in \mathrm{int}(U)$. Set

$$\mathcal{N} := \big\{(t, x, y) \ : \ \hat{\nu}(t, x, y) \in \mathrm{int}(U) \ \text{ and } \ -\mathcal{L}^{\hat{\nu}(t,x,y)} \varphi(t, x) + b\,(t, x, y, \hat{\nu}(t, x, y)) > 0\big\},$$

where $\hat{\nu}(t, x, y) = \psi(t, x, y, D\varphi(t, x))$, and, for $\eta \ge 0$,

$$\mathcal{M}_\eta := \{\ (t, x) \ : \ \ (t, x, \varphi(t, x) - \eta) \in \mathcal{N}\ \}.$$

Note that this definition of $\mathcal{M}_0 := \mathcal{M}_0(\varphi)$ agrees with the previous definition. Moreover, in view of our hypothesis, for all sufficiently small η, $\overline{\mathcal{M}} \subset \mathcal{M}_\eta$. Fix $\eta \le \beta$ satisfying this inclusion.

Step 2. Let η be as in the previous step. Let (X_η, Y_η) be the solution of the state equation with initial data $X_\eta(t_0) = x_0$, $Y_\eta(t_0) = \varphi(t_0, x_0) - \eta$ and the control ν given in the feedback form

$$\nu(t, x) = \psi(t, x, \varphi(t, x) - \eta, D\varphi(t, x)).$$

Set

$$\nu(t) := \nu(t, X_\eta(t))$$

so that

$$(X_\eta, Y_\eta) = Z^\nu_{t_0, x_0, v(t_0,x_0) - \eta} \ = \ \big(X^\nu_{t_0,x_0}, Y^\nu_{t_0,x_0,v(t_0,x_0)-\eta}\big).$$

Set

$$\hat{Y}_\eta(t) := \varphi(t, X_\eta(t)) - \eta + (v - \varphi)(t_0, x_0),$$

and observe that $Y_\eta(0) = \hat{Y}_\eta(0) = v(t_0, x_0) - \eta$. In the next step, we will compare the processes Y_η and $\hat{Y}_\eta$.

Step 3. By Itô's rule,

$$d\hat{Y}_\eta(t) = \mathcal{L}^{\nu(t)}\varphi(t, X_\eta(t))dt + D\varphi(t, X_\eta(t)) \cdot \sigma^*(t, X_\eta(t), \nu(t))dW(t).$$

In view of (2.1) and the definition of $\nu(t)$,

$$D\varphi(t, X_\eta(t)) \cdot \sigma^*(t, X_\eta(t), \nu(t)) = a^*(t, X_\eta(t), \hat{Y}_\eta(t), \nu(t)).$$

Hence

$$d\hat{Y}_\eta(t) = \hat{b}(t)dt + a^*(t, X_\eta(t), \hat{Y}_\eta(t), \nu(t))dW(t),$$

where

$$\hat{b}(t) := \mathcal{L}^{\nu(t)}\varphi(t, X_\eta(t)).$$

Recall that Y_η solves the same SDE with a different drift term:

$$dY_\eta(t) = b(t)dt + a^*(t, X_\eta(t), Y_\eta(t), \nu(t))dW(t),$$

where $b(t) := b(t, X_\eta(t), Y_\eta(t), \nu(t))$.

Let θ be the stopping time

$$\theta := \inf\ \{\ s > 0\ :\ \ (t_0 + s, X_\eta(t_0 + s)) \notin \mathcal{M}\ \}.$$

Since $\mathcal{M}$ is an open set containing (t_0, x_0), the stopping time θ is positive a.s.

Now, from the definition of η, we have $\mathcal{M} \subset \mathcal{M}_\eta$. It follows that, for $t \in [t_0, t_0+\theta)$, $(t, X_\eta(t)) \in \mathcal{M}_\eta$ a.s.; i.e., $(t, X_\eta(t), \hat{Y}_\eta(t)) \in \mathcal{N}$ a.s. by definition of $\mathcal{M}_\eta$. Hence

$$b(t) > \mathcal{L}^{\nu(t)}\varphi(t, X_\eta(t))\ =\ \hat{b}(t), \qquad t \in [t_0, t_0 + \theta), \quad P - \text{a.s.}$$

Since $Y_\eta(0) = \hat{Y}_\eta(0) = v(t_0, x_0) - \eta$, it follows from stochastic comparison that

$$\hat{Y}_\eta(t) \le Y_\eta(t), \qquad t \in [t_0, t_0 + \theta), \quad P - \text{a.s.}$$

Step 4. We now proceed to contradict (DP2). First, observe that, by continuity of the process X_η, $(t_0 + \theta, X_\eta(t_0 + \theta)) \in \partial_p\mathcal{M}$. Also, from inequality (4.12), we have $v \le \varphi - \beta + (v - \varphi)(t_0, x_0)$ on $\partial_p\mathcal{M}$. Therefore,

$$\begin{aligned} Y_\eta(t_0 + \theta) - v(t_0 + \theta, X_\eta(t_0 + \theta)) &\ge \beta + Y_\eta(t_0 + \theta) - \varphi(t_0 + \theta, X_\eta(t_0 + \theta)) \\ &\qquad + (v - \varphi)(t_0, x_0) \\ &= (\beta - \eta) + Y_\eta(t_0 + \theta) - \hat{Y}_\eta(t_0 + \theta) \\ &\ge \beta - \eta\ \ge\ 0 \end{aligned}$$

from step 3. By (4.12) and the definition of (X_η, Y_η), we have $Y_\eta = Y^\nu_{t_0,x_0,v(t_0,x_0)-\eta}$ and $X_\eta = X^\nu_{t_0,x_0}$. Then the previous inequality contradicts (DP2). □

Proof of the subsolution property. Fix $(t_0, x_0) \in [0, T) \times \mathbb{R}^d$, and let φ be a $C^2([0, T] \times \mathbb{R}^d)$ function satisfying

$$(v^* - \varphi)(t_0, x_0) = (\text{strict}) \max_{(t,x)\in[0,T)\times\mathbb{R}^d} (v^* - \varphi).$$

Set $z_0 := (x_0, \varphi(t_0, x_0))$. Let $\mathcal{M}_0 := \mathcal{M}_0(\varphi)$ be as in the previous lemma. Since (t_0, x_0) is a *strict* maximizer of $(v^* - \varphi)$ and since $\mathcal{M}_0$ is an open set, by the previous lemma we conclude that $(x_0, y_0) \notin \mathcal{M}_0$. Then, by the definition of $\mathcal{M}_0$,

$$\min\left\{ \inf_{\zeta\in\tilde{U}_1} \left(\delta_U(\zeta) - \zeta^*\hat{\nu}(t_0, z_0)\right),\ -\mathcal{L}^{\hat{\nu}(t_0,z_0)}\varphi(t_0, x_0) + b\left(t_0, z_0, \hat{\nu}(t_0, z_0)\right) \right\} \le 0,$$

and therefore v^* is a viscosity subsolution. □

5. Terminal condition. To characterize the value function as the unique solution of the dynamic programming equation, we need to specify the terminal data. The definition of the value function implies that

$$v(T,x) = g(x), \qquad x \in \mathbb{R}. \tag{5.1}$$

However, it is known that

$$\underline{G}(x) := \liminf_{t\uparrow T,\ x'\to x} v(t,x')$$

may be strictly larger than $g(x)$ (see, for instance, [4] and Lemma 5.1 below).

In this section, we will characterize $\underline{G}$ as the viscosity supersolution of a first order PDE. We will also study

$$\overline{G}(x) := \limsup_{t\uparrow T,\ x'\to x} v(t,x')$$

and prove that $\overline{G}$ is a viscosity subsolution of the same equation. More precisely, we have the following theorem.

THEOREM 5.1. *Let the assumptions of Theorem* 4.1 *hold, and assume that* $\underline{G}$ *and* $\overline{G}$ *are finite for every* $x \in \mathbb{R}^d$. *Suppose, further, that* $(g_*)^* \geq g$. *Then* $\overline{G}$ *and* $\underline{G}$, *respectively, are viscosity super- and subsolutions of the following equations on* $\mathbb{R}^d$:

$$\min\{\underline{G}(x) - g_*(x);\ H\,(T,x,\underline{G}(x), D\underline{G}(x))\} \geq 0,$$
$$\min\{\overline{G}(x) - g^*(x);\ H\,\big(T,x,\overline{G}(x), D\overline{G}(x)\big)\} \leq 0.$$

In most cases, since a subsolution is not greater than a supersolution, this characterization implies that $\overline{G} \leq \underline{G}$ and therefore that $\overline{G} = \underline{G}$. In the next section, we provide examples for which this holds, and we will also compute $G := \overline{G} = \underline{G}$ explicitly in those examples.

The rest of this section is devoted to the proof of Theorem 5.1.

Remark 5.1. In the definition of $\overline{G}$, we may replace v by v^*:

$$\overline{G}(x) = \limsup_{t\uparrow T,\ x'\to x} v^*(t,x').$$

Similarly,

$$\underline{G}(x) := \liminf_{t\uparrow T,\ x'\to x} v_*(t,x').$$

We start with the following lemma.

LEMMA 5.1. *Suppose that* $\underline{G}(x)$ *and* $\overline{G}(x)$ *are finite for every* $x \in \mathbb{R}^d$. *Then*

$$\underline{G}(x) \ \geq\ g_*(x) \text{ for all } \ x \in \mathbb{R}^d.$$

Proof. Take a sequence $(x_n,t_n) \to (x,T)$ with $t_n < T$. Set $y_n := v(t_n,x_n)+(1/n)$. For each n, there exists a control $\nu_n \in \mathcal{U}$ satisfying

$$Y^{\nu_n}_{t_n,x_n,y_n}(T) \geq g\left(X^{\nu_n}_{t_n,x_n}(T)\right) \qquad \mathrm{P}-\text{a.s.}$$

Since a and b are bounded,

$$E\left[\ Y^{\nu_n}_{t_n,x_n,y_n}(T)\ \right] \leq y_n + \|b\|_\infty(T-t_n) \ =\ v(t_n,x_n) + \frac{1}{n} + \|b\|_\infty(T-t_n).$$

We continue by using the following claim, whose proof will be provided later:

$$\left\{Y^{\nu_n}_{t_n,x_n,y_n}(T)\ ,\ \ n\geq 0\right\} \text{ is uniformly integrable.} \tag{5.2}$$

Then

$$\begin{aligned}\liminf_{n\to\infty} v(t_n,x_n) &\geq \liminf_{n\to\infty} E\left[\ Y^{\nu_n}_{t_n,x_n,y_n}(T)\ \right]\\ &= E\left[\ \liminf_{n\to\infty} Y^{\nu_n}_{t_n,x_n,y_n}(T)\ \right]\\ &\geq E\left[\ \liminf_{n\to\infty} g\left(X^{\nu_n}_{t_n,x_n}(T)\right)\ \right].\end{aligned}$$

Since U is compact and (t_n,x_n) converges to (T,x), $X^{\nu_n}_{t_n,x_n}(T)$ approaches x as n tends to infinity. The required result then follows from the definition of the lower semicontinuous envelope g_* of g.

It remains to prove claim (5.2). Since b is bounded,

$$\begin{aligned}\left|Y^{\nu_n}_{t_n,x_n,y_n}(T)\right| &\leq |y_n| + (T-t_n)\|b\|_\infty + \left|\int_{t_n}^T a\left(u, Z^{\nu_n}_{t_n,x_n,y_n}(u),\nu_n(u)\right)^* dW(u)\right|\\ &\leq T\|b\|_\infty + |v(t_n,x_n)| + \left|\int_{t_n}^T a\left(u, Z^{\nu_n}_{t_n,x_n,y_n}(u),\nu_n(u)\right)^* dW(u)\right|.\end{aligned}$$

Now observe that $\limsup v(t_n,x_n) \leq \limsup v^*(t_n,x_n) \leq \overline{G}(x)$ and $\liminf v(t_n,x_n) \geq \liminf v_*(t_n,x_n) \geq \underline{G}(x)$. This proves that the sequence $v(t_n,x_n)$ is bounded. In order to complete the proof, it suffices to show that the sequence

$$\left\{U_n := \int_{t_n}^T a\left(u, Z^{\nu_n}_{t_n,x_n,y_n}(u),\nu_n(u)\right)^* dW(u),\ n\geq 0\right\}$$

is uniformly integrable. Since a is bounded,

$$\sup_{n\geq 0} E\left[U_n^2\right] \leq \sup_{n\geq 0}(T-t_n)\|a^*a\|_\infty\ \leq\ T\|a^*a\|_\infty.$$

Hence $\{U_n,\ n\geq 0\}$ is bounded in L^2, and, therefore, it is uniformly integrable. □

Next, we will show that $\underline{G}$ is a viscosity supersolution of $H\geq 0$, where H is as in (4.3).

LEMMA 5.2. *Suppose that $\underline{G}(x)$ is finite for every $x\in\mathbb{R}^d$. Then $\underline{G}$ is a viscosity supersolution of*

$$H(T,x,\underline{G}(x),D\underline{G}(x))\geq 0.$$

Proof. By definition, $\underline{G}$ is lower semicontinuous. Let f be a $C^2(\mathbb{R}^d)$ function satisfying

$$0\ =\ (\underline{G}-f)(x_0) = \min_{x\in\mathbb{R}^d}(\underline{G}-f)$$

at some $x_0\in\mathbb{R}^d$. Observe that $\underline{G}\geq f$ on $\mathbb{R}^d$.

Step 1. In view of Remark 5.1, there exists a sequence (s_n,ξ_n) converging to (T,x_0) such that $s_n<T$ and

$$\lim_{n\to\infty}\ v_*(s_n,\xi_n) = \underline{G}(x_0).$$

For a positive integer n, consider the auxiliary test function

$$\varphi_n(t,x) := f(x) - \frac{1}{2}|x-x_0|^2 + \frac{T-t}{(T-s_n)^2}.$$

Let $B := B_1(x_0)$ be the unit open ball in $\mathbb{R}^d$ centered at x_0. Choose $(t_n, x_n) \in [s_n, T] \times \bar{B}$, which minimizes the difference $v_* - \varphi_n$ on $[s_n, T] \times \bar{B}$.

Step 2. We claim that, for sufficiently large n, $t_n < T$, and x_n converges to x_0. Indeed, for sufficiently large n,

$$(v_* - \varphi_n)(s_n, \xi_n) \le -\frac{1}{2(T-s_n)}.$$

On the other hand, for any $x \in \bar{B}$,

$$(v_* - \varphi_n)(T,x) = \underline{G}(x) - f(x) + \frac{1}{2}|x-x_0|^2 \ge \underline{G}(x) - f(x) \ge 0.$$

Comparing the two inequalities leads us to conclude that $t_n < T$ for large n. Suppose that, on a subsequence, x_n converges to x^*. Since $t_n \ge s_n$ and (t_n, x_n) minimizes the difference $(v_* - \varphi_n)$,

$$\begin{aligned}
(\underline{G} - f)(x^*) - (\underline{G} - f)(x_0) &\\
&\le \liminf_{n\to\infty} (v_* - \varphi_n)(t_n, x_n) - (v_* - \varphi_n)(s_n, \xi_n) - \frac{1}{2}|x_n - x_0|^2 \\
&\le \limsup_{n\to\infty} (v_* - \varphi_n)(t_n, x_n) - (v_* - \varphi_n)(s_n, \xi_n) - \frac{1}{2}|x_n - x_0|^2 \\
&\le -\frac{1}{2}|x^* - x_0|^2.
\end{aligned}$$

Since x_0 minimizes the difference $\underline{G} - f$,

$$0 \le (\underline{G} - f)(x^*) - (\underline{G} - f)(x_0) \le -\frac{1}{2}|x^* - x_0|^2.$$

Hence $x^* = x_0$. The above argument also proves that

$$\begin{aligned}
0 &= \lim_{n\to\infty} (v_* - \varphi_n)(t_n, x_n) - (v_* - \varphi_n)(s_n, \xi_n) \\
&= -\underline{G}(x_0) + \lim_{n\to\infty} v_*(t_n, x_n) + \frac{(T-s_n) - (T-t_n)}{(T-s_n)^2} \\
&\ge -\underline{G}(x_0) + \limsup_{n\to\infty} v_*(t_n, x_n).
\end{aligned}$$

This proves that $\limsup_{n\to\infty} v_*(t_n, x_n) \le \underline{G}(x_0)$. Since $\limsup v_*(t_n, x_n) \ge \liminf v_*(t_n, x_n) \ge \underline{G}(x_0)$, by definition of $\underline{G}$, this proves that

$$\lim_{n\to\infty} v_*(t_n, x_n) = \underline{G}(x_0). \tag{5.3}$$

This implies that, for all sufficiently large n, (t_n, x_n) is a local minimizer of the difference $(v_* - \varphi_n)$. In view of the general theory of viscosity solutions (see, for instance, Fleming and Soner [14]), the viscosity property of v_* holds at (t_n, x_n).

Step 3. We now use the viscosity property of v_* in $[0,T) \times \mathbb{R}^d$: for every n,

$$H(t_n, x_n, v_*(t_n, x_n), D\varphi_n(x_n, t_n)) \ge 0.$$

Note that $D\varphi_n(x_n,t_n) = Df(x_n,t_n) - (x_n - x_0)$, and recall that H is continuous; see Remark 4.4. Since (t_n,x_n) tends to (T,x_0), (5.3) implies that

$$H(T,x_0,\underline{G}(x_0),Df(x_0)) \geq 0. \qquad \square$$

These results imply that $\underline{G}$ is a viscosity supersolution of

$$\min\ \{\underline{G}(x) - g_*(x);\ H(T,x,\underline{G}(x),D\underline{G}(x))\} \geq 0, \tag{5.4}$$

proving the first part of Theorem 5.1. The following result concludes the proof of the theorem.

LEMMA 5.3. *Suppose that $\underline{G}(x)$ and $\overline{G}(x)$ are finite for every $x \in \mathbb{R}^d$ and that $(g_*)^* \geq g$. Then $\overline{G}$ is a viscosity subsolution on $\mathbb{R}^d$ of*

$$\min\ \left\{\overline{G}(x) - g^*(x);\ H(T,x,\overline{G}(x),D\overline{G}(x))\right\} \leq 0.$$

Proof. By definition, $\overline{G}$ is upper semicontinuous. Let $x_0 \in \mathbb{R}^d$ and $f \in C^2(\mathbb{R}^d)$ satisfy

$$0 \ = \ (\overline{G} - f)(x_0) = \max_{x\in\mathbb{R}^d}(\overline{G} - f).$$

We need to show that, if $\overline{G}(x_0) > g^*(x_0)$, then

$$H(T,x_0,\overline{G}(x_0),D\overline{G}(x_0)) \leq 0. \tag{5.5}$$

So we assume that

$$\overline{G}(x_0) > g^*(x_0). \tag{5.6}$$

For a positive integer n, set

$$s_n := T - \frac{1}{n^2},$$

and consider the auxiliary test function

$$\varphi_n(t,x) := f(x) + \frac{1}{2}|x - x_0|^2 + n(T-t), \qquad (t,x) \in [s_n,T] \times \mathbb{R}^d.$$

In order to obtain the required result, we shall first prove that the test function φ_n does not satisfy the condition of Lemma 4.2 on $[s_n,T] \times B_R(x_0)$ for some $R > 0$, and then we shall pass to the limit as $n \to \infty$.

Step 1. By definition, $\overline{G} \geq \underline{G}$. From Lemma 5.1, this provides $\overline{G} \geq g_*$ and then $\overline{G} \geq (g_*)^*$ by upper semicontinuity of $\overline{G}$. Hence, by assumption of the lemma,

$$\overline{G} \geq g. \tag{5.7}$$

This proves that $(v - \varphi_n)(T,x) = (g - f)(x) - |x - x_0|^2/2 \leq (\overline{G} - f)(x) \leq 0$ by definition of the test function f. Then, for all $R > 0$,

$$\sup_{B_R(x_0)} (v - \varphi_n)(T,.) \leq 0.$$

Now suppose that there exists a subsequence of (φ_n), still denoted (φ_n), such that

$$\lim_{n\to\infty} \sup_{B_R(x_0)} (v - \varphi_n)(T,.) = 0,$$

and let us work toward a contradiction. For each n, let $(x_n^k)_k$ be a maximizing sequence of $(v-\varphi_n)(T,.)$ on $B_R(x_0)$; i.e.,

$$\lim_{n\to\infty}\lim_{k\to\infty}(v-\varphi_n)(T,x_n^k)=0.$$

Then it follows from (5.7) that $(v-\varphi_n)(T,x_n^k)\le -|x_n^k-x_0|^2/2$, which provides

$$\lim_{n\to\infty}\lim_{k\to\infty}x_n^k=x_0.$$

Therefore,

$$\begin{aligned}0 \;&=\; \lim_{n\to\infty}\lim_{k\to\infty}(v-\varphi_n)(T,x_n^k) \;=\; \lim_{n\to\infty}\lim_{k\to\infty}g(x_n^k)-f(x_0)\\ &\le\; \limsup_{x\to x_0}g(x)-f(x_0) \;=\; (g^*-f)(x_0) \;<\; (G-f)(x_0)\end{aligned}$$

by (5.6). Since $(G-f)(x_0)=0$, this cannot happen since $(G-f)(x_0)=0$. The consequence of this is

$$\limsup_{n\to\infty}\sup_{B_R(x_0)}(v-\varphi_n)(T,.)<0 \qquad \text{for all } R>0. \tag{5.8}$$

Step 2. Let $(t_n,x_n)_n$ be a maximizing sequence of $(v^*-\varphi_n)$ on $[s_n,T]\times\partial B_R(x_0)$. Then, since $T-t_n\le T-s_n=n^{-2}$,

$$\limsup_{n\to\infty}\sup_{[s_n,T]\times\partial B_R(x_0)}(v^*-\varphi_n)\le\limsup_{n\to\infty}(v^*(t_n,x_n)-f(x_n))-\frac{1}{2}R^2.$$

Since $t_n\longrightarrow T$ and, after passing to a subsequence, $x_n\longrightarrow x^*$ for some $x^*\in\partial B_R(x_0)$, we get

$$\limsup_{n\to\infty}\sup_{[s_n,T]\times\partial B_R(x_0)}(v^*-\varphi_n)\le(\overline{G}-f)(x^*)-\frac{1}{2}R^2\;\le\;-\frac{1}{2}R^2.$$

This, together with (5.8), implies that, for all $R>0$, there exists $n(R)$ such that, for all $n>n(R)$,

$$\max\{\;(v-\varphi_n)\;:\;\partial_p\left((s_n,T)\times B_R(x_0)\right)\;\}<0\;=\;(v^*-\varphi_n)(T,x_0).$$

Hence it follows from Lemma 4.2 that

$$(s_n,T)\times B_R(x_0)\text{ is not a subset of }\mathcal{M}_0(\varphi_n)\quad\text{for all }\;n>n(R). \tag{5.9}$$

Step 3. Observe that, for all $\nu\in U$ and (t,x,y),

$$-\mathcal{L}^\nu\varphi_n(t,x)\;=\;n-\mathcal{L}^\nu f(x)-\mu(t,x,\nu)^*(x-x_0)-\frac{1}{2}\mathrm{Trace}[\sigma^*\sigma](t,x,\nu)>b(t,x,y,\nu),$$

provided that n is sufficiently large. Then, for large n,

$$\begin{aligned}&\mathcal{M}_0(\varphi_n)\cap((s_n,T)\times B_R(x_0))\\ &\qquad=\{\;(t,x)\in(s_n,T)\times B_R(x_0)\;:\;H(t,x,\varphi_n(t,x),D\varphi_n(t,x))>0\;\}.\end{aligned}$$

In view of this, it follows from (5.9) that there is a sequence (t_n,x_n) converging to (T,x_0) such that

$$H(t_n,x_n,\varphi_n(t_n,x_n),D\varphi_n(t_n,x_n))\le 0.$$

We now let n tend to infinity to obtain (5.5). □

6. Application: Superreplication problem in finance. Consider a financial market consisting of

- a nonrisky asset with price process $\tilde{X}^0$ normalized to unity,
- a risky asset $\tilde{X}$ defined by a positive price process with dynamics described by an SDE.

A trading strategy is an $\mathbb{F}$-adapted process $\nu = \{\nu(t),\, 0 \le t \le T\}$ valued in the closed interval $[-\ell, u]$ with $\ell, u \in [0, \infty)$ and $\ell + u > 0$. At each time $t \in [0, T]$, $\nu(t)$ represents the proportion of wealth invested in the risky asset $\tilde{X}$. The set of all trading strategies is denoted by $\mathcal{U}$.

Given an initial capital $\tilde{y} > 0$ and a trading strategy ν, the wealth process $\tilde{Y}$ is defined by

$$\tilde{Y}_{\tilde{y}}^{\nu}(0) \;=\; \tilde{y} \text{ and } d\tilde{Y}_{\tilde{y}}^{\nu}(t) \;=\; \tilde{Y}_{\tilde{y}}^{\nu}(t)\nu(t)\,\frac{d\tilde{X}(t)}{\tilde{X}(t)}.$$

We shall consider a "large investor" model in which the dynamics of the risky asset price process may be affected by trading strategies. Namely, given a trading strategy $\nu \in \mathcal{U}$,

$$\begin{aligned} \tilde{X}^{\nu}(0) = e^{X^{\nu}(0)} \;&=\; e^{X(0)}, \quad \tilde{X}^{\nu}(t) \;=\; e^{X^{\nu}(t)}, \\ dX^{\nu}(t) &= \mu\left(t, X^{\nu}(t), \nu(t)\right) dt + \sigma\left(t, X^{\nu}(t), \nu(t)\right) dW(t), \end{aligned}$$

where W is a one-dimensional Brownian motion. Define the log-wealth process:

$$Y_y^{\nu}(0) \;=\; y \;:=\; \ln(\tilde{y}) \text{ and } Y_y^{\nu}(t) \;=\; \ln\left(\tilde{Y}_{\tilde{y}}^{\nu}(t)\right).$$

Then a direct application of Itô's lemma provides

$$dY_y^{\nu}(t) = b\left(t, X^{\nu}(t), \nu(t)\right) dt + \nu(t)\sigma\left(t, X^{\nu}(t), \nu(t)\right) dW(t),$$

where

$$b(t,x,r) = r\left(\mu + \frac{1}{2}\sigma^2\right)(t,x,r) - \frac{1}{2}r^2\sigma^2(t,x,r).$$

Let f be a positive function defined on $[0, \infty)$. The superreplication problem is defined by

$$\tilde{v}(0, X(0)) := \inf\left\{\tilde{y} > 0 \;:\; \exists\, \nu \in \mathcal{U},\; \tilde{Y}_{\tilde{y}}^{\nu}(T) \ge f\left(X^{\nu}(T)\right) \quad P - \text{a.s.}\right\}.$$

Here $f\left(X^{\nu}(T)\right)$ is a contingent claim. The value function of the above superreplication problem is then the minimal initial capital which allows the seller of the contingent claim to face the promised payoff $f\left(X^{\nu}(T)\right)$ through some trading strategy $\nu \in \mathcal{U}$.

To see that the superreplication problem belongs to the general class of stochastic target problems studied in the previous sections, we introduce

$$v(0, X(0)) \;:=\; \ln \tilde{v}(0, X(0)) \text{ and } g \;:=\; \ln f.$$

Then

$$v(0, X(0)) := \inf\left\{y \in \mathbb{R} \;:\; \exists\, \nu \in \mathcal{U},\; Y_y^{\nu}(T) \ge g\left(X^{\nu}(T)\right) \quad P - \text{a.s.}\right\}.$$

Remark 6.1. Assume that function g is bounded. Then the value function v is bounded. Using the notation of previous sections, we also have that v_*, v^*, $\underline{G}$, and $\overline{G}$ are bounded functions.

Let us introduce the support function of the interval $[-\ell^{-1}, u^{-1}]$:

$$h(p) := u^{-1}p^+ \ + \ \ell^{-1}p^-,$$

with the convention $1/0 = +\infty$, and the usual notation $p^+ := p \vee 0$ and $p^- := (-p)^+$. Observe that h is a mapping from $\mathbb{R}$ into $\mathbb{R} \cup \{+\infty\}$. We also denote by $\overline{F}$ and $\underline{F}$ the functions

$$\overline{F} \ := \ e^{\overline{G}} \ = \ \limsup_{t\uparrow T, x' \to x} \tilde{v}(t,x') \text{ and } \underline{F} \ := \ e^{\underline{G}} \ = \ \liminf_{t\uparrow T, x' \to x} \tilde{v}(t,x').$$

Applying Theorems 4.1 and 5.1, we obtain the following characterization of the superreplication problem $\tilde{v}$ by a change of variable.

THEOREM 6.1. *Let μ and σ be bounded Lipschitz functions uniformly in the t variable, and $\sigma > 0$. Suppose further that g is bounded and satisfies $(g_*)^* \geq g$. Then*

(i) *$\tilde{v}$ is a discontinuous viscosity solution of*

$$\min\left\{ -\tilde{v}_t(t,x) - \frac{1}{2}\sigma^2\left(t,x,\tilde{v}_x(t,x)\right)\tilde{v}_{xx}(t,x) \ ; \ \tilde{v}(t,x) - h\left(\tilde{v}_x(t,x)\right) \right\} = 0$$
$$\text{on} \quad [0,T) \times \mathbb{R} \, .$$

(ii) *The terminal value functions $\underline{F}$ and $\overline{F}$ satisfy in the viscosity sense*

$$\min\{ \ \underline{F} - f_* ; \ \underline{F} - h(\underline{F}_x) \ \} \geq 0,$$
$$\min\{ \ \overline{F} - f^* ; \ \overline{F} - h(\overline{F}_x) \ \} \leq 0 \qquad \text{on} \quad \mathbb{R}.$$

The rest of this section is devoted to the characterization of the terminal functions $\overline{F}$ and $\underline{F}$. It is known that the first order variational inequality appearing in part (ii) of the above theorem could fail to have a unique bounded discontinuous viscosity solution: under our condition $(f_*)^* \geq f$, all viscosity discontinuous bounded solutions have the same lower semicontinuous envelope; see Barles [3]. Therefore, we do not have much to say in the case where the payoff function f is not continuous.

We provide a characterization of the terminal condition of the superreplication problem in the case of Lipschitz payoff function f.

PROPOSITION 6.1. *Let the conditions of Theorem* 6.1 *hold. Assume, further, that the payoff function f is Lipschitz on $\mathbb{R}$. Then*

$$\overline{F}(x) \ = \ \underline{F}(x) = \hat{f}(x) \ := \ \sup_{y \in \mathbb{R}} f(x+y)e^{-\delta(y)},$$

where $\delta := \delta_U$ is the support function of the interval $U = [-\ell, u]$.

Proof. From Theorem 6.1, functions $\overline{F}$ and $\underline{F}$ are, respectively, upper and lower semicontinuous viscosity sub- and supersolutions of

$$\text{(VI)} \qquad \min\left\{u - f \ ; \ u - h(u_x)\right\} = 0 \quad \text{on} \quad \mathbb{R}.$$

In order to obtain the required result, we shall first prove that $\hat{f}$ is a (continuous) viscosity supersolution of (VI) (step 1). Then we will prove that $\underline{F} \geq \hat{f}$ (step 2). The proof is then concluded by means of a comparison theorem (Barles [2, Theorem 4.3, p. 93]); since f is Lipschitz, conditions (H1), (H4), and (H11) of this theorem are easily seen to hold. Since $\overline{F} \geq \underline{F}$ by definition, the above claims provide $\hat{f} \geq \overline{F} \geq \underline{F} \geq \hat{f}$.

Step 1. Let us prove that $\hat{f}$ is a continuous viscosity supersolution of (VI).

(i) $\hat{f}$ is a Lipschitz function. To see this, observe that, since δ is a sublinear function, it follows that $\hat{\hat{f}} = \hat{f}$, where $\hat{\hat{f}}$ is defined by the same formula as $\hat{f}$ with $\hat{f}$ substituted to f. Then, since $\hat{f}$ and δ are nonnegative,

$$\begin{aligned}\hat{f}(x+y) - \hat{f}(x) &\le \hat{f}(x+y)(1-e^{-\delta(y)}) \quad \text{for all } \ y \in \mathbb{R} \\ &\le \hat{f}(x+y)\delta(y) \ \le \ \|f\|_\infty \max(u,\ell)|y|.\end{aligned}$$

(ii) $\hat{f}$ is a supersolution of (VI). To see this, let $x_0 \in \mathbb{R}$ and $\varphi \in C^1(\mathbb{R})$ be such that $0 = (\hat{f} - \varphi)(x_0) = \min(\hat{f} - \varphi)$. Observe that $\hat{f} \ge \varphi$. Since $\hat{f} > 0$, we can assume without loss of generality that $\varphi > 0$. By definition, we have $\hat{f}(x_0) \ge f(x_0)$.

It remains to prove that $(\varphi'/\varphi)(x_0) \in [-\ell, u]$. Since $\hat{\hat{f}} = \hat{f}$, we have

$$\varphi(x_0) \ = \ \hat{f}(x_0) \ge \hat{f}(x_0+h)e^{-\delta(h)} \ \ge \ \varphi(x_0+h)e^{-\delta(h)}$$

for all $h \in \mathbb{R}$. Now let h be an arbitrary positive constant. Then

$$\frac{\varphi(x_0+h) - \varphi(x_0)}{h} \le \varphi(x_0+h)\frac{1-e^{-uh}}{h},$$

and, by sending h to zero, we get $\varphi'(x_0) \le u\varphi(x_0)$. Similarly, by considering an arbitrary constant $h < 0$, we see that $\varphi'(x_0) \ge -\ell\varphi(x_0)$.

Step 2. We now prove that $\underline{F} \ge \hat{f}$. From the supersolution property of $\underline{F}$, we have that $\underline{F} \ge f$, and, for all $y \in \mathbb{R}$, $\underline{F}$ satisfies in the viscosity sense

$$\delta(y)\underline{F} - y\underline{F}_x \ge 0.$$

By an easy change of variable, we see that $\underline{G} = \ln \underline{F}$ satisfies in the viscosity sense

$$\delta(y) - y\underline{G}_x \ge 0.$$

This proves that the function $x \longmapsto \delta(y)x - y\underline{G}(x)$ is nondecreasing (see, e.g., Cvitanić, Pham, and Touzi [9]), and therefore

$$\begin{aligned}\delta(y)(x+y) - y\underline{G}(x+y) &\ge \delta(y)x - y\underline{G}(x) \quad \text{for all } \ y > 0, \\ \delta(y)(x+y) - y\underline{G}(x+y) &\le \delta(y)x - y\underline{G}(x) \quad \text{for all } \ y < 0.\end{aligned}$$

Recalling that $\underline{F} \ge f$, this provides

$$\underline{F}(x) \ge \sup_{y\in\mathbb{R}} \underline{F}(x+y)e^{-\delta(y)} \ \ge \ \sup_{y\in\mathbb{R}} f(x+y)e^{-\delta(y)} \ = \ \hat{f}(x). \qquad \square$$

REFERENCES

[1] F. Antonelli, *Backward-forward stochastic differential equations*, Ann. Appl. Probab., 3 (1993), pp. 777–793.

[2] G. Barles, *Solutions de viscosité des équations de Hamilton-Jacobi*, Math. Appl. 17, Springer-Verlag, Paris, 1994.

[3] G. Barles, *Discontinuous viscosity solutions of first-order Hamilton-Jacobi equations: A guided visit*, Nonlinear Anal., 20 (1993), pp. 1123–1134.

[4] M. Broadie, J. Cvitanić, and H. M. Soner, *Optimal replication of contingent claims under portfolio constraints*, The Review of Financial Studies, 11 (1998), pp. 59–79.

[5] M. G. Crandall, H. Ishii, and P. L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[6] J. Cvitanić and I. Karatzas, *Hedging contingent claims with constrained portfolios*, Ann. Appl. Probab., 3 (1993), pp. 652–681.
[7] J. Cvitanić, I. Karatzas, and H. M. Soner, *Backward SDE's with constraints on the gains process*, Ann. Probab., 26 (1998), pp. 1522–1551.
[8] J. Cvitanić and J. Ma, *Hedging options for a large investor and forward-backward SDE's*, Ann. Appl. Probab., 6 (1996), pp. 370–398.
[9] J. Cvitanić, H. Pham, and N. Touzi, *Super-replication in stochastic volatility models under portfolio constraints*, J. Appl. Probab., 36 (1999), pp. 523–545.
[10] N. El Karoui, *Les aspects probabilistes du contrôle stochastique*, in Ninth Saint Flour Probability Summer School—1979, Lecture Notes in Math. 876, Springer-Verlag, New York, 1981, pp. 73–238.
[11] N. El Karoui and M.-C. Quenez, *Dynamic programming and pricing of contingent claims in an incomplete market*, SIAM J. Control Optim., 33 (1995), pp. 29–66.
[12] N. El Karoui, S. Peng, and M.-C. Quenez, *Backward stochastic differential equations in finance*, Math. Finance, 7 (1997), pp. 1–72.
[13] H. Föllmer and D. Kramkov, *Optional decomposition under constraints*, Probab. Theory Related Fields, 109 (1997), pp. 1–25.
[14] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, Heidelberg, Berlin, 1993.
[15] E. Jouini and H. Kallal, *Arbitrage in securities markets with short-sales constraints*, Math. Finance, 3 (1995), pp. 197–232.
[16] Y. Hu and S. Peng, *Solution of forward-backward stochastic differential equations*, Probab. Theory Related Fields, 103 (1995), pp. 273–283.
[17] I. Karatzas and S. Shreve, *Methods of Mathematical Finance*, Springer-Verlag, New York, Heidelberg, Berlin, 1998.
[18] J. Ma, P. Protter, and J. Yong, *Solving forward-backward stochastic differential equations explicitly: A four step scheme*, Probab. Theory Related Fields, 98 (1994), pp. 339–359.
[19] J. Ma and J. Yong, *Solvability of forward-backward SDE's and the nodal set of Hamilton-Jacobi-Bellman equations*, Chinese Ann. Math. Ser. B, 16 (1995), pp. 279–298.
[20] E. Pardoux, *Backward stochastic differential equations and viscosity solutions of semilinear parabolic PDE's of second order*, in Proceedings of Geilo Conference, Geilo, Norway, 1996.
[21] E. Pardoux and S. Tang, *Forward-backward stochastic differential equations and quasilinear parabolic PDE's*, Probab. Theory Related Fields, 114 (1999), pp. 123–150.
[22] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[23] H. M. Soner and N. Touzi, *Superreplication under gamma constraints*, SIAM J. Control Optim., 39 (2000), pp. 73–96.
[24] H. M. Soner and N. Touzi, *Dynamic programming for stochastic target problems and geometric flows*, J. Eur. Math. Soc. (JEMS), to appear.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 2, pp. 425–456

BOUNDARY-VALUE PROBLEMS FOR SYSTEMS OF HAMILTON–JACOBI–BELLMAN INCLUSIONS WITH CONSTRAINTS*

JEAN-PIERRE AUBIN†

Abstract. We study in this paper boundary-value problems for *systems of Hamilton–Jacobi–Bellman* first-order partial differential equations and variational inequalities, the solutions of which are constrained to obey viability constraints. They are motivated by some control problems (such as impulse control) and financial mathematics. We shall prove the *existence and uniqueness of such solutions in the class of closed set-valued maps* by giving a precise meaning to what a solution means in this case. We shall also provide *explicit formulas* for this problem. When we deal with Hamilton–Jacobi–Bellman equations, we obtain the existence and uniqueness of Frankowska contingent episolutions. We shall deduce these results from the fact that the graph of the solution is the viable-capture basin of the graph of the boundary conditions under an auxiliary system and then from their properties and their characterizations proved in [J.-P. Aubin, *SIAM J. Control Optim.*, 40 (2001), pp. 853–881].

Key words. partial differential inclusion, systems of Hamilton–Jacobi–Bellman equations, viability, capture basins, method of characteristics, shocks, impulse control, contingent cone, Marchaud map

AMS subject classifications. 49A52, 49J24, 49K24, 49L25

PII. S0363012900381510

Introduction. It is well known that value functions of optimal control problems are solutions to Hamilton–Jacobi partial differential equations of the form

$$-\frac{\partial}{\partial t}v(t,x) + \inf_{u\in P(x,v(t,x))}\left(\frac{\partial}{\partial x}v(t,x)f(x,v(t,x),u) - g(x,v(t,x),u)\right) \;=\; 0$$

with adequate boundary conditions.

Observe, nevertheless, that, in this equation, the infimum hides two inequalities:

1. there exists $u \in P(x,v(t,x))$ such that

$$-\frac{\partial}{\partial t}v(t,x) + \frac{\partial}{\partial x}v(t,x)f(x,v(t,x),u) - g(x,v(t,x),u) \;\leq\; 0;$$

2. for all $u \in P(x,v(t,x))$,

$$\frac{\partial}{\partial t}v(t,x) - \frac{\partial}{\partial x}v(t,x)f(x,v(t,x),u) + g(x,v(t,x),u) \;\leq\; 0.$$

However, several other problems of control theory lead to the study of *controlled systems* of first-order partial differential equations (or systems of first-order *partial differential inclusions*): Let $P : X \times Y \rightsquigarrow \mathcal{U}$ be a set-valued map associating with any pair (x,y) a feasible set $P(x,y)$ of controls, and let f and g be single-valued maps from $X \times Y \times \mathcal{U}$ to finite dimensional vector spaces X and Y, respectively.

The problem is to find a set-valued map $V : \mathbf{R}_+ \times X \rightsquigarrow Y$ satisfying the following:

*Received by the editors November 21, 2000; accepted for publication (in revised form) November 20, 2001; published electronically June 18, 2002.
http://www.siam.org/journals/sicon/41-2/38151.html

†Réseau de Recherche Viabilité, Jeux, Contrôle 14, rue Domat, F-75005 Paris, France (J.P.Aubin@wanadoo.fr).

1. there exists $u \in P(x, V(t,x))$ such that

$$(1) \quad 0 \in \text{“} - \frac{\partial}{\partial t} V(t,x) + \frac{\partial}{\partial x} V(t,x) f(x, V(t,x), u)\text{”} - g(x, V(t,x), u);$$

2. for all $u \in P(x, v(t,x))$,

$$(2) \qquad 0 \in \text{“} \frac{\partial}{\partial t} V(t,x) - \frac{\partial}{\partial x} V(t,x) f(x, V(t,x), u)\text{”} + g(x, V(t,x), u),$$

where we shall give a meaning to the derivative

$$\text{“} \frac{\partial}{\partial t} V(t,x) - \frac{\partial}{\partial x} V(t,x) f(x, V(t,x), u)\text{”}$$

in Theorem 3.1 below. Indeed, even in the absence of controls, it is well known that such solutions may have shocks, i.e., can be set-valued, and, when they happen to be single-valued, are not even necessarily differentiable in the usual sense.

The definition of solution shall be taken in a generalized sense—the Frankowska solution[1] that we shall define later in the paper.

In order to obtain uniqueness, we have to impose boundary conditions. Furthermore, problems arising in economics, finance, and other fields lead us to *introduce constraints* bearing on both the state and the solution. We shall describe these boundary conditions and constraints by introducing two set-valued maps $\Phi : \mathbf{R}_+ \times X \rightsquigarrow Y$ and $\Psi : \mathbf{R}_+ \times X \rightsquigarrow Y$ such that $\Phi \subset \Psi$. The first one encompasses initial and/or boundary-value conditions, or other conditions as we shall see, and the second one encompasses viability constraints both on the state variables x, which must remain in the domain of Ψ, and on the solution $V(t,x)$.

We shall prove that *there exists a unique* “solution” $(t,x) \rightsquigarrow V(t,x)$ to this general problem (1,2) satisfying the conditions

$$\forall\, (t,x) \in \mathbf{R}_+ \times X, \quad \Phi(t,x) \subset V(t,x) \subset \Psi(t,x)$$

in the class of closed set-valued maps (i.e., set-valued maps with closed graph) that depends continuously of the data Φ (in the “graphical sense,” mapping graphical limits to graphical limits, as is explained later).

Even more, we shall provide an explicit formula when $f(x,u)$ and $P(x)$ do not depend on the variable y and when

$$g(x,y,u) := -M(x,u)y - L(x,u)$$

is affine with respect to y, where

[1] Hélène Frankowska proved that the epigraph of the value function of an optimal control problem—assumed to be only lower semicontinuous—is semipermeable (i.e., invariant and backward viable) under a (natural) auxiliary system. Furthermore, when it is continuous, she proved that its epigraph is viable and its hypograph invariant [43, 44, 46]. By duality, she proved that the latter property is equivalent to the fact that the value function is a viscosity solution of the associated Hamilton–Jacobi equation in the sense of Crandall and Lions. See also [32, 31] for more details. We refer also to [39, 40] for the study of these equations through the characteristics method using the contingent derivative (and not epi- and hypo- derivatives). Such concepts have been extended to solutions of systems of first-order partial differential equations without boundary conditions by Hélène Frankowska and the author (see [19, 20, 21, 22, 23, 24] and chapter 8 of [2] and [7]). See also [16, 17]. This point of view is used here in the case of boundary value problems.

1. M is a continuous matrix-valued function

$$M : (x,u) \in X \times \mathcal{U} \mapsto M(x,u) \in \mathcal{L}(X,Y),$$

2. L is a continuous "vector-Lagrangian"

$$L : (x,u) \in X \times \mathcal{U} \mapsto L(x,u) \in Y.$$

Let us denote by $\mathcal{C} : x \in X \rightsquigarrow \mathcal{C}(x) \subset \mathcal{C}(0,\infty;X) \times L^1(0,\infty;\mathcal{U})$ the set-valued map associating with $x \in X$ the set $\mathcal{C}(x)$ of the pairs $(x(\cdot),u(\cdot))$ of solutions to the control system

$$\left\{ \begin{array}{ll} \text{(i)} & x'(t) \ = \ f(x(t),u(t)), \\ \text{(ii)} & u(t) \ \in \ P(x(t)), \end{array} \right.$$

starting at x at $t = 0$.

In the absence of constraints ($\Psi(t,x) := Y$), we shall prove that, setting

$$\left\{ \begin{array}{l} J_\Phi(t;(x(\cdot),u(\cdot)))(T,x) \\ \\ := \ e^{\int_0^t M(x(s),u(s))ds}\Phi(T-t,x(t)) + \displaystyle\int_0^t e^{\int_0^\tau M(x(s),u(s))ds} L(x(\tau),u(\tau))d\tau, \end{array} \right.$$

the set-valued solution V is defined by

$$V(\Phi)(T,x) \ := \ \bigcup_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)} \ \bigcup_{t\in[0,T]} J_\Phi(t;(x(\cdot),u(\cdot)))(T,x).$$

With an adequate choice of the set-valued map Ψ associated with the set-valued map Φ, we find as a solution the set-valued map defined by

$$W(\Phi)(T,x) \ := \ \bigcup_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)} \ \bigcap_{t\in[0,T]} J_\Phi(t;(x(\cdot),u(\cdot)))(T,x).$$

We shall find as many formulas as pairs (Ψ,Φ) of set-valued maps (see formula (24) of Theorem 5.1 below).

We can read this type of result the other way around: For instance, the set-valued map $V(\Phi)$ defined by

$$\left\{ \begin{array}{l} V(\Phi)(T,x) \ := \ \bigcup_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)} \bigcup_{t\in[0,T]} \\ \left(e^{\int_0^t M(x(s),u(s))ds}\Phi(T-t,x(t)) + \displaystyle\int_0^t e^{\int_0^\tau M(x(s),u(s))ds} L(x(\tau),u(\tau))d\tau \right) \end{array} \right.$$

is the unique "solution" to the Hamilton–Jacobi partial differential inclusion (1,2) satisfying the initial condition

$$V(0,x) \ = \ \Phi(0,x)$$

and

$$\forall\, t \geq 0,\ x \in X, \ \ \Phi(t,x) \ \subset \ V(t,x).$$

They define set-valued analogue of optimal control problems, where the "$\cup$" operation replaces the "inf" operation and the "$\cap$" operation replaces the "sup" operation. Actually, when $Y := \mathbf{R}$ and when we associate with two extended functions $\mathbf{c} : \mathbf{R}_+ \times X \rightsquigarrow \mathbf{R} \cup \{+\infty\}$ and $\mathbf{b} : \mathbf{R}_+ \times X \rightsquigarrow \mathbf{R} \cup \{+\infty\}$ the set-valued maps

$$\left\{ \begin{array}{ll} \text{(i)} & \Phi(t,x) := \mathbf{c}(t,x) + \mathbf{R}_+, \\ \text{(ii)} & \Psi(t,x) := \mathbf{b}(t,x) + \mathbf{R}_+, \end{array} \right.$$

we find problems of dynamic valuation and management of portfolios in mathematical finance, used, in particular, for valuating options as in [55, 30, 56]. For instance, we deduce that

$$\left\{ \begin{array}{l} \inf_{y \in V(\Phi)(T,x)} y \ = \ \inf_{(x(\cdot),u(\cdot)) \in \mathcal{C}(x)} \inf_{t \in [0,T]} \\ \left(e^{\int_0^t M(x(s),u(s))ds} \mathbf{c}(T-t, x(t)) + \displaystyle\int_0^t e^{\int_0^\tau M(x(s),u(s))ds} L(x(\tau), u(\tau)) d\tau \right) \end{array} \right.$$

is the *valuation function* of a stopping time problem (see section 5 below).

These two explicit formulas are given by the *caliber* $V : \mathbf{R}_+ \times X \rightsquigarrow Y$ defined in the following way: y belongs to $V(T,x)$ if there exist a control $t \in [0,T] \mapsto u(t)$ and a time $T^* \in [0,T]$ such that the solution $(x(\cdot), u(\cdot), y(\cdot))$ to the control system

$$\left\{ \begin{array}{lll} \text{(i)} & x'(t) & = f(x(t), y(t), u(t)), \\ \text{(ii)} & y'(t) & = g(x(t), y(t), u(t)), \\ \text{(iii)} & u(t) & \in P(x(t), y(t)) \end{array} \right.$$

starting at $x(0) = x$, $y(0) = y$ satisfies

$$\left\{ \begin{array}{ll} \text{(i)} & \forall\, t \in [0, T^*], \ \ y(t) \in \Psi(T-t, x(t)), \\ \\ \text{(ii)} & y(T^*) \in \Phi(T - T^*, x(T^*)). \end{array} \right.$$

Observe that taking $\Phi(t,x) = \emptyset$ whenever $t > 0$ guarantees that $T^* = T$.

We shall prove that this set-valued map is the unique solution to our problem (1,2). Actually, this is a reformulation dictated by problems arising in dynamic economic theory, finance mathematics, and control theory of the celebrated "method of characteristics."

We shall revisit this method using the tools of set-valued analysis and viability theory which go back to the early 1980's.[2] They find here an unexpected relevance to assert the existence and the uniqueness of the solution to this problem since such solutions may have shocks, i.e., can be set-valued, and even when they happen to be single-valued, they are not differentiable in the usual sense. The tools forged by set-valued analysis and viability theory happen to allow us to prove existence and uniqueness in the class of set-valued maps with closed graph only instead of classes of vector-distributions.[3]

[2] See [18, 2, 60], etc., for instance.

[3] The strong requirement of pointwise convergence of differential quotients can be weakened in (at least) two ways, with each way sacrificing different groups of properties of the usual derivatives:

- *Distributional derivatives.* Fix the direction v, and take the limit of the function $x \mapsto \nabla_h f(x)(v)$ in the weaker sense of distributions. The limit $D_v f$ may then be a distribution and no longer a single-value map. However, it coincides with the usual limit when f is Gâteaux differentiable. Moreover, one can define difference quotients of distributions, take their limit, and thus differentiate distributions.

The basic concept useful in our framework is the concept of the *viable-capture basin* of a "target" $C \subset K$ viable in a constrained subset $K \subset X$ under a differential inclusion $x' \in F(x)$: It is the subset $\mathrm{Capt}_F^K(C)$ of initial states $x_0 \in K$ such that C is reached in finite time before possibly leaving K by at least one solution $x(\cdot) \in \mathcal{S}(x_0)$, where $\mathcal{S}(x_0)$ denotes[4] the set of solutions to the differential inclusion $x' \in F(x)$ starting at x_0.

Then we shall prove that *the graph of the solution* $(t,x) \rightsquigarrow V(t,x)$ *to the above boundary value problem is the viable-capture basin of the graph of the set-valued map* Φ *viable in the graph of the set-valued map* Ψ under the auxiliary differential inclusion

$$\begin{cases} \text{(i)} & \tau'(t) \;=\; -1, \\ \text{(ii)} & x'(t) \;=\; f(x(t),y(t),u(t)), \\ \text{(iii)} & y'(t) \;=\; g(x(t),y(t),u(t)), \\ \text{(iv)} & u(t) \;\in\; P(x(t),y(t)) \end{cases}$$

and that this solution is *unique* among the solutions with closed graph to this boundary value problem.

In some instances, this viable-capture basin can be computed analytically, and we obtain in this case an explicit formula of the solution to the above boundary value problem.

In all cases, the Viability/Capturability Algorithm designed by Patrick Saint-Pierre provides numerically the viable-capture basins and thus the solutions to systems of Hamilton–Jacobi–Bellman equations, bypassing finite-difference methods. (See

---

Distributions are no longer functions or maps defined on $\mathbf{R}^n$, so they lose the pointwise character of functions and maps but retain the linearity of the operator $f \mapsto D_v f$, which is mandatory for using the theory of the linear operator for solving partial differential equations.

- *Graphical derivatives.* Fix the direction x, and take the limit of the function $v \mapsto \nabla_h f(x)(v)$ in the weaker sense of "graphical convergence." (The graph of the graphical limit is the Painlevé–Kuratowski upper limit of the graphs.) The limit $Df(x)$ may then be a set-valued map and no longer a single-value map. However, it coincides with the usual limit when f is Gâteaux differentiable. Moreover, one can define difference quotients of set-valued maps, take their limit, and thus differentiate set-valued maps. These graphical derivatives keep the pointwise character of functions and maps, which is mandatory for implementing the Fermat rule, proving inverse function theorems under constraints or using Lyapunov functions, for instance, but they lose the linearity of the map $f \mapsto Df(x)$.

In both cases, the approaches are similar: They use (different) convergences *weaker than the pointwise convergence* for increasing the possibility for the difference-quotients to converge. But the price to pay is the loss of some properties by passing to these weaker limits (the pointwise character for distributional derivatives, the linearity of the differential operator for graphical derivatives).

[4]When we are studying the viable-capture basins of targets under differential inclusions, we observe that they are not specific to differential inclusions. They involve only a few properties of the solution map $\mathcal{S}$, associating with any initial state x the set $\mathcal{S}(x)$ of solutions $t \mapsto x(t)$ that are solutions to the above differential inclusion starting at x at initial time 0. These properties (translation and concatenation properties as well as continuity properties) of the solution map $x \rightsquigarrow \mathcal{S}(x)$ are common to other control problems, such as

1. control problems with memory (see the contributions of [50, 51, 52], some of them being presented in [2]), before known under the name of functional control problems and now called "path dependent control systems";
2. parabolic (diffusion-reaction) type partial differential inclusions (see the contributions of [64, 65, 66, 67], some of them being presented in [2]), also known as distributed control systems;
3. "mutational equations" governing the evolution in metric spaces, including "morphological equations" governing the evolution of sets (see [4], for instance).

[61, 38, 55, 56] for solutions of Hamilton–Jacobi–Bellman equations derived from mathematical finance.)

This existence and uniqueness result follows from the following three steps:

1. From a characterization proved in [15] stating the viable-capture basin $\mathrm{Capt}^K(C)$ is the *unique* closed subset D between C and K satisfying

$$\text{(3)} \qquad \begin{cases} \text{(i)} & \mathrm{Capt}^D(C) \;=\; D, \\ \text{(ii)} & \mathrm{Capt}^K(D) \;=\; D. \end{cases}$$

2. From a characterization stated below of the viable-capture basin of a target C viable in a closed subset K under a differential inclusion $x' \in F(x)$ proved in [12] (see also [57, 58, 59]): Let us recall that
   (a) $K\backslash C$ is a repeller, meaning that all solutions $x(\cdot) \in \mathcal{S}(x)$ starting from $x \in K\backslash C$ reach C or leave K in finite time;
   (b) the subset $D\backslash C$ is said to be *locally viable* under $\mathcal{S}$ if from any initial state $x \in D\backslash C$ starts at least one solution viable in $D\backslash C$ on a nonempty interval;
   (c) a subset $D \subset K$ is *locally backward invariant relatively to* K under $\mathcal{S}$ if, for every $x \in D$, all backward solutions starting from x and viable in K on an interval $[0, T]$ are viable in D on $[0, T]$.

   THEOREM 0.1. *Let us assume that F is Marchaud*[5] *and that a closed subset $C \subset K$ satisfies the property*

$$\text{(4)} \qquad K\backslash C \;\; \textit{is a repeller under } F.$$

   Then the viable-capture basin $\mathrm{Capt}^K(C)$ *is the* unique *closed subset D satisfying $C \subset D \subset K$ and*

$$\text{(5)} \qquad \begin{cases} \text{(i)} & D\backslash C \textit{ is} \text{ locally viable } \textit{under } \mathcal{S}, \\ \text{(ii)} & D \textit{ is} \text{ locally backward invariant relatively to } K \textit{ under } \mathcal{S}. \end{cases}$$

3. From the viability and invariance theorems that translate the necessary and sufficient conditions (5) in terms of tangential conditions. We recall that the *contingent cone* to a subset K at a point $x \in K$, introduced in the early 1930's independently by Bouligand and Severi, adapts to any subset the concept of tangent space to manifolds: A direction $v \in X$ belongs to $T_K(x)$ if there exist sequences $h_n > 0$ and $v_n \in X$ converging to 0 and v, respectively, such that

$$\forall\, n \geq 0, \;\; x + h_n v_n \;\in\; K.$$

   This means that the contingent cone is the Painlevé–Kuratowski upper limit of the subsets $\frac{K-x}{h}$ when h converges to 0 (see, for instance, [18, 60] for more details).

   We shall use the following statements of [12].

   THEOREM 0.2. *Let us assume that F is Marchaud, that $C \subset K$ and K are closed, and that $K\backslash C$ is a repeller. Then the viable-capture basin* $\mathrm{Capt}_F^K(C)$ *is*

[5] F is a *Marchaud map* if

$$\begin{cases} \text{(i)} & \text{the graph and the domain of } F \text{ are nonempty and closed,} \\ \text{(ii)} & \text{the values } F(x) \text{ of } F \text{ are convex,} \\ \text{(iii)} & \text{the growth of F is linear:} \\ & \exists\, c > 0 \mid \forall\, x \in X, \;\; \|F(x)\| := \sup_{v \in F(x)} \|v\| \;\leq\; c(\|x\| + 1). \end{cases}$$

(a) *the* largest *closed subset D satisfying $C \subset D \subset K$ and*

$$\forall\, x \in D\backslash C,\ \ F(x) \cap T_D(x) \neq \emptyset; \tag{6}$$

(b) *if, furthermore, F is Lipschitz, the* unique *closed subset D satisfying $C \subset D \subset K$ and the* Frankowska properties

$$\begin{cases} \text{(i)} & \forall\, x \in D\backslash C,\ \ F(x) \cap T_D(x) \neq \emptyset, \\ \text{(ii)} & \forall\, x \in D \cap \mathrm{Int}(K),\ \ -F(x) \subset T_D(x), \\ \text{(iii)} & \forall\, x \in D \cap \partial K,\ \ -F(x) \subset T_D(x) \cup T_{X\backslash K}(x). \end{cases} \tag{7}$$

4. From the property

$$T_{\mathrm{Graph}(V)}(t,x,y) = \mathrm{Graph}(DV(t,x,y))$$

of the contingent derivative $DV(t,x,y)$ of the set-valued map $V : (t,x) \rightsquigarrow V(t,x)$ at the point (t,x,y) of its graph introduced in [1]: The graph of the set-valued map $DV(t,x,y)$ from $\mathbf{R} \times X$ to Y is equal to the contingent cone to the graph of V at (t,x,y) (see [1]). This is how Fermat in 1637 defined the derivative of a function as the slope of the tangent to its graph. Leibniz and Newton provided the characterization in terms of limits of difference quotients. Here, too, *the contingent derivative $DV(t,x,y)$ is the upper graphical limit of the difference quotients*, the graph of which being by definition the upper limit of the graphs of the difference quotients $\nabla_h V(t,x,y)$ of V at $(t,x,y) \in \mathrm{Graph}(V)$ defined by

$$(\lambda, f) \mapsto \nabla_h V(t,x,y)(\lambda,f) := \frac{V(t+\lambda h, x+hf) - y}{h}.$$

Indeed, we observe that

$$\mathrm{Graph}(\nabla_h V(t,x,y)) = \frac{\mathrm{Graph}(V) - (t,x,y)}{h}$$

so that the contingent cone to the graph of V, being the upper limit of the graphs of the difference quotients, is equal by definition to the graph of the upper graphical limit of the difference quotients. Consequently, to say that $g \in Y$ belongs to the contingent derivative $DV(t,x,y)\,(\pm 1, f)$ of V at (t,x,y) in the direction $(\pm 1, f) \in \mathbf{R} \times X$ means that

$$\liminf_{h \to 0+,\ f' \to f} \left\| \frac{V(t \pm h, x + hf') - y}{h} - g \right\| = 0.$$

Since the contingent cone is a closed subset, the graph of a contingent derivative is always closed and positively homogenous. (This is what remains of the required linearity of the derivative in classical analysis, but, fortunately, we can survive pretty well without linearity.)

When $u : \mathbf{R} \times X \mapsto Y$ is single-valued, we set $Du(t,x) := Du(t,x,u(t,x))$. We see at once that $Du(t,x)(\pm 1, f) = \pm \frac{\partial u(t,x)}{\partial t} + \frac{\partial u(t,x)}{\partial x} \cdot f$ whenever u is differentiable at (t,x). When u is Lipschitz on a neighborhood of (t,x) and when the dimension of X is finite, the domain of $Du(t,x)$ is not empty. Furthermore, the Rademacher theorem stating that a locally Lipschitz single-valued map is almost everywhere differentiable implies that $x \rightsquigarrow Du(t,x)$ is almost everywhere single-valued.

However, in this case, equality $Du(t,x)(-1,-f) = -Du(t,x)(1,f)$ is not true in general. We refer to [18, 60] for more details.

The above results—which are interesting by themselves for other mathematical models of evolutionary economics (see [3]), population dynamics, dynamical games (see [11]), and epidemiology—can be applied to many other problems. Dealing with subsets, they can be applied to graphs of single-valued maps as well as set-valued maps, to epigraphs and hypographs of (extended) real-valued functions for solving Hamilton–Jacobi–Bellman equations, to graph of "impulse" maps (which take empty values except in discrete sets, useful in the study of hybrid systems or inventory management), etc. (See, for instance, [5, 10, 6, 8, 9, 25, 26, 27, 28, 29, 34, 53, 54, 62, 63] and their bibliographical references.)

Outline. We begin in section 1 with two nonstandard motivations arising in macroeconomic problems faced by central banks (filtering informations on the economy from past informations and future expectations) and in the study of impulse and hybrid systems, leading to systems of Hamilton–Jacobi–Bellman inclusions. This section (which states the problems without solving them, as this task is done in specific articles) can be skipped by true believers in mathematics. We define in section 2 the "caliber" of a pair of set-valued maps (Φ, Ψ) under a control system that appears naturally in some control problems and in economic and financial mathematics. We next prove in section 3 that the caliber is the unique solution to the system of Hamilton–Jacobi–Bellman inclusions satisfying the imposed conditions. In section 4, we provide a useful stability result, stating roughly that the caliber of graphical limits is the graphical limit of calibers. Section 5 deals with the explicit formula of the caliber when the control system is structured and the exosystem is affine with respect to the second variable. We also prove that, in this case, the caliber is the unique solution to a system of "fixed-set equations" and provide interesting "barrier properties" of the boundary of the caliber. We derive in section 6 the usual characterization theorems of the valuation functions of a large class of control and stopping time problems as Frankowska episolutions to the scalar Hamilton–Jacobi–Bellman equations that justify the usefulness of the results they are derived from, and we extend this scalar situation in section 7 to the case of "dynamical vector optimization," where we look for intertemporal Pareto minima. □

1. Motivations. We shall provide two motivations coming from recent issues arising

1. in macroeconomics (filtering informations on the economy from past informations and future expectations), and
2. in hybrid systems and impulse control,

leading to systems of Hamilton–Jacobi–Bellman partial differential inclusions.

We refer to specific articles for more details, since they use the basic theorems of this paper to solve the Hamilton–Jacobi–Bellman partial differential inclusions that appear in those articles.

Further applications to the value functions of optimal control and stopping time problems are given in section 6, and applications to dynamic vector optimization are given in section 7.

1.1. Selector through past informations and future expectations. As a first motivation, we present a problem originating in a research program under current investigation by Noël Bonneuil, Halim Doss, Georges Haddad, Henri Pages, Dominique Pujal, Patrick Saint-Pierre, and the author on macroeconomic problems faced by central banks.

We quote excerpts of the introduction of [68]: *It is a truism that monetary policy*

operates under considerable uncertainty about the state of the economy and the size and nature of the disturbances that hit the economy...But in a more realistic case where important variables are forward-looking [and not only backward-looking] variables, the problem of efficient signal-extraction is inherently more complicated...In the real world, many important indicator variables are forward-looking variables [routinely watched by central banks].

We suggest taking up this issue by using nonlinear continuous evolutionary models controlled by instruments such as interest rates. We keep the problem of extracting the "real evolution," knowing only backward-looking measurements and forward-looking expectations that we shall describe by "expectation tubes."

The use of the Kalman filter for extracting information is replaced by the recent concept of the detector introduced in [13] and [27] in the case of "impulse and hybrid control systems." We adapt this concept of the detector in the case of both backward-looking and forward-looking informations and expectations.

For that purpose, we introduce three time variables for describing the evolution of the system: the current (or present) time T, the past time $t \in [0, T]$, and a prediction time or forward-looking time $s \geq T$, where $a := s - T \in \mathbf{R}_+$ is the prediction horizon used to take into account anticipations and expectations (or make predictions) in the future.

At each past date, the state is measured, or informations on the state are gathered: This is mathematically described by a *detectability tube* (as in [13])

$$t \in \mathbf{R}_+ \rightsquigarrow I(t) \ \subset \ X := \mathbf{R}^n$$

that provides the limited amount of information about the states at time t. We take $I(t) := X$ when no information is recorded at time t. Hence "discrete" measurements are obtained when $I(t) \neq X$ only for a discrete number of instants t_n.

An example of detectability tubes is given by $I(t) := h^{-1}(y(t))$, where $h : X \mapsto Y$ is an observation map (or measurement map) and where $t \mapsto y(t)$ is the evolution of the observed output:

$$\forall\, t \in [0, T], \ \ y(t) \ = \ h(x(t)).$$

The same framework also houses the case when the observation map is set-valued: We set $I(t) := H^{-1}(y(t))$, and the above viability condition reads

$$\forall\, t \in [0, T], \ \ y(t) \ \in \ H(x(t)),$$

so that "tychastic" uncertainties (by opposition to stochastic uncertainties) on the measurements can be incorporated in this framework.

In order to take into account expectations made at each instant t for future dates $s := t + a$, $a \geq 0$, we describe them mathematically by an *expectation tube* $(t, a) \in \mathbf{R}^2_+ \mapsto I(t, a) \subset X$, where we set $I(t, 0) := I(t)$ for obtaining the detectability tube.

We may assume that, if $a_1 \leq a_2$, then $I(t, a_2) \subset I(t, a_1) \subset I(t, 0) =: I(t)$, since the predictions made at time t up to time $s_2 := t+a_2$ are valid up to time $s_1 := t+a_1 \leq s_2$.

Therefore, we associate with any current time T, any horizon $s \geq t$, and any backward-looking time $t \in [0, T]$ the set $I(t, s - t)$ of states measured at time t that depend upon the duration $a := s - t$ of the expectation interval between t and the horizon s.

Hence the information/expectation constraint can be summarized by

$$\forall\, T \geq 0,\ \forall\, s \geq T,\ \forall\, t \in [0,T],\ x(t) \in I(t, s-t). \tag{8}$$

Let $\mathcal{U}$ be a space of controls, regulees, prices, interest-rate instruments, etc. The dynamics of the state are described by a map $f : (t,a,x,u) \in \mathbf{R}^2_+ \times X \times \mathcal{U} \mapsto X$ and by a set-valued map $P : \mathbf{R}_+ \times X \rightsquigarrow \mathcal{U}$ depicting the state-dependent constraints on the controls $u \in P(a,x)$.

For any current time T and horizon $s \geq T$, we assume that the evolution of the state of the system is governed by the control system

$$\forall\, t \in [0,T], \begin{cases} \text{(i)} & x'(t) = f(t, s-t, x(t), u(t)), \\ \text{(ii)} & u(t) \in P(s-t, x(t)). \end{cases} \tag{9}$$

In other words, at each time $t \in [0,T]$, the velocity $x'(t)$ depends upon the time t, the time $s-t$ left to the horizon s, and a control $u(t)$ subjected to constraints depending upon both the expected time $s-t$ left to the horizon and the state $x(t)$ at time t.

We also introduce a set-valued map $a \in \mathbf{R}_+ \rightsquigarrow C(a)$ specifying another constraint on the subsets $C(a)$ of initial states that may depend upon the term $a \geq 0$ satisfying

$$\forall\, a \geq 0,\ C(a) \subset I(0,a).$$

DEFINITION 1.1. *Let us consider a control system (f,P), an expectation tube $I : \mathbf{R}^2_+ \rightsquigarrow X$, and a tube $C : \mathbf{R}_+ \rightsquigarrow X$ satisfying, for all $a \geq 0$, $C(a) \subset I(0,a)$.*

The selector *is the set-valued map $\mathbf{S}_{(I,C)} : \mathbf{R}^2_+ \rightsquigarrow X$ that associates with any current time T and any expectation $a := s-T$ the (possibly empty) subset $\mathbf{S}_{(I,C)}(t,a)$ of states $x \in I(t,a)$ such that there exists a solution $x(\cdot) \in \mathcal{S}(t,a,x)$ to the system*

$$\begin{cases} & \forall\, t \in [0,T],\ \forall\, s := T + a \geq 0, \\ \text{(i)} & x'(t) = f(t, s-t, x(t), u(t)), \\ \text{(ii)} & u(t) \in P(s-t, x(t)) \end{cases} \tag{10}$$

such that

$$\begin{cases} \text{(i)} & x(0) \in C(s), \\ \text{(ii)} & \forall\, t \in [0,T],\ x(t) \in I(t, s-t), \\ \text{(iii)} & x(T) = x. \end{cases}$$

In other words, both the dynamics and the constraints depend upon horizon $s \geq T$ and take into account the informations gathered at any preceding time $t \in [0,T]$ and expectations at time $s-t$ left to the horizon s. The selector is thus a tube associating with any horizon $s \geq T$ the set of states x such that there exists a control $u(\cdot)$ governing the evolution $x(\cdot)$ through control system (9):

$$\forall\, t \in [0,T], \begin{cases} \text{(i)} & x'(t) = f(t, s-t, x(t), u(t)), \\ \text{(ii)} & u(t) \in P(s-t, x(t)), \end{cases}$$

which satisfies for all anterior time $t \in [0,T]$ the expected constraints made at that time t for the future time $s-t$.

We can easily check the following lemma.

LEMMA 1.2. *The graph of the selector $\mathbf{S}_{(I,C)}$ is the capture basin of $\{0\} \times \mathrm{Graph}(C)$ viable in the graph of the tube I under the auxiliary system*

$$\begin{cases} \text{(i)} & \tau'(t) = -1, \\ \text{(ii)} & \alpha'(t) = +1, \\ \text{(iii)} & x'(t) = -f(\tau(t), \alpha(t), x(t)). \end{cases}$$

Proof. Indeed, to say that (T, a, x) belongs to the viable-capture basin of $\{0\} \times \text{Graph}(C)$ viable in $\text{Graph}(I)$ means that there exist an evolution $\widehat{x}(\cdot)$ to $\widehat{x}'(t) \in -f(T-t, a+t, \widehat{x}(t), \widetilde{u}(t))$ starting at $\widehat{x}(0) := x$ and a time $t^* \geq 0$ such that

$$\left\{ \begin{array}{lll} \text{(i)} & \forall\, t \in [0, t^*], & (T-t, a+t, \widehat{x}(t)) \in \text{Graph}(I), \\ \text{(ii)} & & (T-t^*, a+t^*, \widehat{x}(t^*)) \in \{0\} \times \text{Graph}(C). \end{array} \right.$$

The second condition means that $t^* = T$ and that $\widehat{x}(T)$ belongs to $C(a+T)$. The first one means that for every $t \in [0, T]$, $\widehat{x}(t) \in I(T-t, a+t)$. This amounts to saying that the evolution $x(\cdot) := \widehat{x}(T - \cdot)$ is a solution to the control system

$$x'(t) \;=\; f(t, a+T-t, x(t), u(t))$$

starting at $x(0) := \widehat{x}(T) \in C(T+a)$, satisfying $x(T) = x$, and

$$\forall\, t \in [0, T],\ \forall\, a \geq 0,\ \ x(t) \in I(t, a+T-t).$$

This means that $x \in \mathbf{S}_{(I,C)}(T, a)$. □

We shall therefore characterize the selector as a solution to a system of Hamilton–Jacobi–Bellman partial differential inclusions

$$\forall\, x \in V(t,a),\ \exists\, u \in P(a,x) \quad \text{such that} \quad -\frac{\partial V(t,a,x)}{\partial t} + \frac{\partial V(t,a,x)}{\partial a} + f(t,a,x,u) \;=\; 0$$

satisfying the initial condition

$$\forall\, a \geq 0,\ \ V(0,a) \;=\; C(a)$$

and the viability constraints

$$\forall\, (t,a) \in \mathbf{R}^2_+,\ \ V(t,a) \;\subset\; I(t,a).$$

We deduce from the knowledge of the derivatives of the selector the *regulation map* $R : \mathbf{R}^2_+ \times X \rightsquigarrow \mathcal{U}$ providing the controls (or regulees, prices, or interest-rate instruments) that at each time t, for any future date a and any state x, answer the detection/prediction problems. The regulation map associates with any triple (t, a, x) the set $R(t, a, x)$ of controls $u \in P(a, x)$ such that the solutions to the new control system

$$\left\{ \begin{array}{ll} & \forall\, t \in [0, T],\ \forall\, s \geq T, \\ \text{(i)} & x'(t) \;=\; f(t, s-t, x(t), u(t)), \\ \text{(ii)} & u(t) \;\in\; R(t, s-t, x(t)) \end{array} \right. \tag{11}$$

satisfy the constraints

$$\left\{ \begin{array}{ll} \text{(i)} & x(0) \;\in\; C(s), \\ \text{(ii)} & \forall\, t \in [0, T],\ \ x(t) \;\in\; I(t, s-t), \\ \text{(iii)} & x(T) \;=\; x. \end{array} \right.$$

Finally, the Capture Basin Algorithm allows us to compute the selector and the regulation map.

1.2. The substratum of an impulse differential inclusion. *Impulse differential inclusions* are described by two set-valued maps, F—the right-hand side of the differential inclusion $x' \in F(x)$ governing the continuous evolution of an impulse system—and Φ, describing the *reset map* reinitializing the system when required and a constrained set K inside which the evolution of the impulse differential equation must remain. We denote by $\mathcal{S}(x)$ the set of solutions $x(\cdot)$ to the differential inclusion starting at x.

Let us set $x(^-t) := \lim_{\tau \mapsto t-} x(\tau)$ when $x(\cdot)$ is defined on some interval $[t-\eta, t[$, where $\eta > 0$, and, for consistency, $x(s) = x(^-t)$ if $s = t$.

An evolution of the impulse differential inclusion, called a "run" or an "execution" in the hybrid system community, is a finite or infinite sequence $x(\cdot) := \{(\tau_n, x_n, x_n(\cdot))\}_{n \geq 0}$ made of triples of

1. nonnegative *cadences* $\tau_n \in [0, +\infty[$,
2. a sequence of *reinitialized states* x_n,
3. a sequence of *motives* $x_n(\cdot) \in \mathcal{S}(x_n)$ satisfying the end-point condition $x_n(\tau_n) \in \Phi^{-1}(x_{n+1})$,

defining the sequence of impulse times $t_{n+1} := t_n + \tau_n$ and, on each interval $[t_n, t_{n+1}[$,

$$\forall\, t \in [t_n, t_{n+1}[, \;\; x(t) \;:=\; x_n(t - t_n). \tag{12}$$

If the sequence is finite and stops at τ_N, we set $\tau_{N+1} := +\infty$ and take $x_N(\cdot) \in \mathcal{S}(x_N)$. (This definition is taken from [25, 26].)

We say that a run $x(\cdot)$ is viable in K if, for any $t \geq 0$, $x(t) \in K$.

At this stage, a run $x(\cdot)$ can just be a (discrete) sequence of states $x_{n+1} \in \Phi(x_n)$ at a fixed time, or just a (continuous) solution $x(\cdot)$ to the differential inclusion $x' \in F(x)$, or a hybrid of these two modes, the discrete and the continuous.

We just define the concept of a *substratum* of an impulse differential inclusion introduced in [9], which summarizes the salient features of a run, by considering only its sequences of cadences τ_n and of reinitialized states x_n. Knowing them, we can reconstruct the motives of the run by taking solutions $x_n(\cdot) \in \mathcal{S}(x_n)$ satisfying the end-point condition $x_n(\tau_n) \in \Phi^{-1}(x_{n+1})$. The question that arises is to provide an algorithm that provides these sequences of cadences τ_n and of reinitialized states x_n without solving the impulse differential inclusion, but only through such an algorithm. The substratum just does that.

Definition 1.3. *We associate with the dynamics (F, Φ) of the impulse differential inclusion its* substratum $\Gamma : \mathbf{R}_+ \times K \rightsquigarrow K$*, that is the set-valued map associating with any $(T, x) \in \mathbf{R}_+ \times K$ the subset*

$$\Gamma(T, x) \;:=\; \bigcup_{x(\cdot) \in \mathcal{S}_F^K(x)} \Phi(x(T))$$

of the elements $y \in \Phi(x(T))$, where $x(\cdot) \in \mathcal{S}_F^K(x)$ is a solution to the differential inclusion $x' \in F(x)$ starting at x and viable in K until it reaches $x(T) \in C := K \cap \Phi^{-1}(K)$ at time T.

We associate with the substratum Γ

1. *the* cadence map

$$\mathbf{C}(x) \;:=\; \{t \geq 0 \quad \textit{such that} \;\; \Gamma(t, x) \;\neq\; \emptyset\} \;\; \textit{and}$$

2. *the* initialization map $\mathbf{I} : K \rightsquigarrow X$*,*

$$\mathbf{I}(x) \;=\; \bigcup_{t \in \mathbf{C}(x)} \Gamma(t, x).$$

Knowing the substratum Γ of (K, F, Φ) and thus the cadence map $\mathbf{C}$ and the initialization map $\mathbf{I}$, we can *reconstruct* a viable run of the impulse differential inclusion (F, Φ) through the following algorithm: Given the cadence τ_n and the initial state x_n, we take

$$(13)\quad \begin{cases} \text{(i)} & \text{the next cadence } \tau_{n+1} \in \mathbf{C}_{(F,\Phi)}(x_n), \\ \text{(ii)} & \text{the next reinitialized state } x_{n+1} \in \Gamma_{(F,\Phi)}(\tau_{n+1}, x_n) \subset \mathbf{I}(x_n), \\ \text{(iii)} & \text{the next motive } x_n(\cdot) := x(\cdot + t_n), \text{ a solution to } x' \in F(x) \text{ satisfying} \\ & x_n(0) \;=\; x_n \;\text{ and }\; x_n(\tau_{n+1}) \;\in\; \Phi^{-1}(x_{n+1}). \end{cases}$$

Assume for a while that the impulse differential inclusion is actually an impulse differential equation (f, φ), where the maps f and φ are single-valued, and that the substratum is single-valued and differentiable. We define

$$\Phi(t,x) \;:=\; \begin{cases} \varphi(x) & \text{if} \quad x \in C := K \cap \varphi^{-1}(K), \\ \emptyset & \text{if} \quad x \notin C \end{cases}$$

and

$$\Psi(t,x) \;:=\; \begin{cases} K & \text{if} \quad x \in K, \\ \emptyset & \text{if} \quad x \notin K. \end{cases}$$

Then we shall prove that the substratum is a "solution" $v(t,x)$ to the system of first-order partial differential inclusions

$$\forall\, i = 1, \dots n, \quad -\frac{\partial v_i(t,x)}{\partial t} + \sum_{j=1}^{n} \frac{\partial v_i(t,x)}{\partial x_j} f_j(x) \;=\; 0$$

satisfying the "condition"

$$\forall\, x \in C := K \cap \varphi^{-1}(K), \;\; v(0,x) \;=\; \varphi(x),$$

which either is single-valued or takes empty values—and thus is a set-valued map.

Actually, we shall extend this result to general impulse differential inclusions by characterizing the substratum as a generalized (set-valued) solution—a Frankowska solution—to the system of first-order partial differential inclusions

$$\begin{cases} \text{(i)} & -\dfrac{\partial}{\partial t} V(t,x) + \dfrac{\partial V(t,x)}{\partial x} \cdot u \;=\; 0, \\ \\ \text{(ii)} & u \;\in\; F(x) \end{cases}$$

satisfying the "condition"

$$\forall\, x \in C := K \cap \Phi^{-1}(K), \;\; V(0,x) \;=\; \Phi(x)$$

and the constraints $V(t,x) \subset K$.

Indeed, the substratum is a particular case of a caliber with $f(x,y,u) := u$, $g(x,y,u) := 0$, $P(x,y) := F(x)$, $\Phi(0,x) := \Phi(x)$ if $x \in K$, $\Phi(0,x) := \emptyset$ if $x \notin K$, $\Phi(t,x) := \emptyset$ if $t > 0$, $\Psi(t,x) := K$ if $x \in K$, and $\Psi(t,x) := \emptyset$ otherwise.

2. The caliber of dynamical constraints and objective. The purpose of this section is to show how "viability techniques" may be efficient for solving systems of first-order partial differential inclusions arising in different fields of control theory and hybrid systems.

We denote by $L^1(0,\infty;\mathcal{U})$ the set of measurable functions from $[0,+\infty[$ to a vector space $\mathcal{U}$, the *control space*.

We consider a control system of the form

$$\left\{\begin{array}{ll} \text{(i)} & x'(t) \;=\; f(x(t),y(t),u(t)),\\ \text{(ii)} & y'(t) \;=\; g(x(t),y(t),u(t)),\\ \text{(iii)} & u(t) \;\in\; P(x(t),y(t)).\end{array}\right.$$

We denote by $\mathcal{B}(x,y)$ the set of solutions $(x(\cdot),y(\cdot),u(\cdot)) \in \mathcal{C}(0,\infty;X\times Y)\times L^1(0,\infty;\mathcal{U})$ to the above system starting at (x,y) at time 0.

Definition 2.1. *We say that the* control system *is*

1. Marchaud *if the set-valued map $P : X \times Y \rightsquigarrow \mathcal{U}$ is Marchaud and if $f : X \times Y \times \mathcal{U} \mapsto X$ and $g : X \times Y \times \mathcal{U} \mapsto Y$ are continuous and affine,*[6]
2. Lipschitz *if the set-valued map $P : X \rightsquigarrow \mathcal{U}$ is Lipschitz and if $f : X\times Y\times\mathcal{U} \mapsto X$ and $g : X \times Y \times \mathcal{U} \mapsto Y$ are Lipschitz.*

We associate the set-valued map $G : \mathbf{R}_+ \times X \times Y \rightsquigarrow \mathbf{R}_+ \times X \times Y$ defined by

$$G(T,x,y) \;:=\; \{\{-1\}\times\{f(x,y,u)\}\times\{g(x,y,u)\}\}_{u\in P(x,y)}\,, \tag{14}$$

and we denote by $\mathcal{R}$ the set-valued map defined by the formula

$$\mathcal{R}(T,x,y) \;=\; \{(T-\cdot,x(\cdot),y(\cdot),u(\cdot))\}_{(x(\cdot),y(\cdot),u(\cdot))\in\mathcal{B}(x,y)}.$$

We infer that *the set-valued map G is a Marchaud (resp., Lipschitz) map whenever the control system is Marchaud (resp., Lipschitz).*

We introduce now dynamical constraints and objectives defined by

1. a set-valued map $\Phi : \mathbf{R}_+ \times X \rightsquigarrow Y$ defining an *objective*, regarded as an obstacle in problems of unilateral mechanics, for instance;
2. a set-valued map $\Psi : \mathbf{R}_+ \times X \rightsquigarrow Y$ defining dynamical constraints. *State constraints* are involved in the domain

$$\mathrm{Dom}(\Psi) \;:=\; \{(t,x)\in\mathbf{R}_+\times X \mid \Psi(t,x) \;\neq\; \emptyset\}$$

of the set-valued map Ψ.

We shall assume that

$$\left\{\begin{array}{ll} \text{(i)} & \forall\,(t,x)\in\mathbf{R}_+\times X,\quad \Phi(t,x) \;\subset\; \Psi(t,x),\\ \\ \text{(ii)} & \forall\,t<0,\ \forall\,x\in X,\quad \Psi(t,x) \;:=\; \emptyset.\end{array}\right.$$

Definition 2.2. *The two constraint and objective set-valued maps being given, the* caliber $(T,x) \rightsquigarrow V_\Psi(\Phi)(T,x)$ of the pair (Ψ,Φ) under the controlled system *is the set-valued map associating with the pair (T,x) made of the horizon T and the initial state x the set of initial observations y such that there exist a control $t\in[0,T]\mapsto u(t)$*

[6] We actually need only that the values $\{(f(x,y,u),g(x,y,u))\}_{u\in P(x,y)}$ are closed and convex.

and a time $T^* \in [0,T]$ *such that a solution* $(x(\cdot),u(\cdot),y(\cdot)) \in \mathcal{B}(x,y)$ *starting at* $x(0)=x$, $y(0)=y$ *satisfies*

$$\text{(15)} \qquad \begin{cases} \text{(i)} & \forall\, t \in [0,T^*],\ \ y(t) \in \Psi(T-t,x(t)), \\ \text{(ii)} & y(T^*) \in \Phi(T-T^*,x(T^*)). \end{cases}$$

We observe at once the following property: *The caliber satisfies the initial condition*

$$\forall\, x \in X,\ \ V_\Psi(\Phi)(0,x) \ = \ \Phi(0,x).$$

Indeed, condition (15) (ii) with $T=0$ means that $y \in V_\Psi(\Phi)(0,x)$, implying that $T^*=0$ and $(0,x,y) \in \mathrm{Graph}(\Phi)$, i.e., $y \in \Phi(0,x)$. Hence $V_\Psi(\Phi)(0,x) \subset \Phi(0,x) \subset V_\Psi(\Phi)(0,x)$. □

What is the connection between this problem and the basic viability theorems? The answer is simple: The graph of the caliber is the capture basin of the graph of the set-valued map Φ viable in the graph of Ψ under the auxiliary control system G.

Proposition 2.3. *The graph of the caliber* $V_\Psi(\Phi)$ *is equal to the viable-capture basin of* $\mathrm{Graph}(\Phi)$ *viable in* $\mathrm{Graph}(\Psi)$ *under the auxiliary system* $\mathcal{R}$*:*

$$\mathrm{Graph}(V_\Psi(\Phi))) \ = \ \mathrm{Capt}_{\mathcal{R}}^{\mathrm{Graph}(\Psi)}(\mathrm{Graph}(\Phi)).$$

Proof. It is enough to translate conditions (15) in the form

$$\text{(16)} \qquad \begin{cases} \text{(i)} & \forall\, t \in [0,T^*],\ \ (T-t,x(t),y(t)) \in \mathrm{Graph}(\Psi), \\ \text{(ii)} & (T-T^*,x(T^*),y(T^*)) \in \mathrm{Graph}(\Phi) \end{cases}$$

to recognize that

$$\mathrm{Graph}(V_\Psi(\Phi)) \ = \ \mathrm{Capt}_{\mathcal{R}}^{\mathrm{Graph}(\Psi)}(\mathrm{Graph}(\Phi)). \qquad \square$$

This being checked, it will be sufficient to translate the properties of capture basins in terms of caliber.

3. Set-valued solutions to systems of Hamilton–Jacobi–Bellman inclusions. In the case of controlled systems, we shall relate the caliber with the set-valued solution to the controlled Hamilton–Jacobi–Bellman partial differential equations:

1. there exists $u \in P(x,V(t,x))$ such that

$$0 \ \in \ \text{“} - \frac{\partial}{\partial t}V(t,x) + \frac{\partial}{\partial x}V(t,x)f(x,V(t,x),u)\text{”} - g(x,V(t,x),u);$$

2. for all $u \in P(x,v(t,x))$,

$$0 \ \in \ \text{“}\frac{\partial}{\partial t}V(t,x) - \frac{\partial}{\partial x}V(t,x)f(x,V(t,x),u)\text{”} + g(x,V(t,x),u),$$

satisfying the initial condition

$$V(0,x) \ = \ \Phi(0,x)$$

and the constraints

$$\forall\, t \geq 0,\ x \in\ X,\ \ \Phi(t,x) \ \subset\ V(t,x) \ \subset\ \Psi(t,x)$$

in a sense that we make precise in the following theorem.

Theorem 3.1. *Let us assume that the system is Marchaud.*

1. *Then the caliber* $V_\Psi(\Phi)$ *of* (Ψ, Φ) *is the* largest *closed set-valued map* $V : \mathbf{R}_+ \times X \rightsquigarrow Y$ *satisfying*

$$\forall\, (t,x) \in \mathbf{R}_+ \times X,\ \ \Phi(t,x)\ \subset\ V(t,x)\ \subset\ \Psi(t,x) \tag{17}$$

and

$$\begin{cases} \forall\, y \in V(t,x)\backslash\Phi(t,x),\ \exists\, u \in P(x,y) \\ \textit{such that} \\ 0\ \in\ DV(-1, f(x,y,u)) - g(x,y,u). \end{cases} \tag{18}$$

2. *Let us set*

$$\mathbf{R}(t,x,y)\ :=\ \{u \in P(x,y) \mid 0\ \in\ DV_\Psi(\Phi)(t,x,y)(-1, f(x,y,u)) - g(x,y,u)\}.$$

Knowing the caliber, any solution satisfying the constraints (15) (i) *and reaching the objective* (15) (ii) *in finite time is obtained in the following way: Starting from* (x_0, y_0) *such that* $y_0 \in V_\Psi(\Phi)(T, x_0)\backslash\Phi(T, x_0)$*, any solution* $(x(\cdot), y(\cdot), u(\cdot))$ *to the control system: for almost all* $t \in [0,T]$*,*

$$\begin{cases} \text{(i)} & x'(t)\ =\ f(x(t), y(t), u(t)), \\ \text{(ii)} & y'(t)\ =\ g(x(t), y(t), u(t)), \\ \text{(iii)} & u(t)\ \in\ \mathbf{R}(T-t, x(t), y(t)), \end{cases} \tag{19}$$

starting at (x,y) *is a solution satisfying*

$$y(t)\ \in\ V_\Psi(\Phi)(T-t, x(t))$$

until the first time $t^* \in]0,T]$ *when*

$$y(t^*)\ \in\ \Phi(T-t^*, x(t^*)).$$

3. *If we assume, furthermore, that the system is Lipschitz, then the caliber* $V_\Psi(\Phi)$ *of* (Ψ, Φ) *is the* unique *closed set-valued map* $V : \mathbf{R}_+ \times X \rightsquigarrow Y$ *satisfying* (17), (18)*, and*

$$\begin{cases} \forall\, y \in V(t,x) \cap \Psi^\circ(t,x),\ \forall\ u \in P(x,y), \\ 0\ \in\ DV(1, -f(x,y,u)) + g(x,y,u) \\ \textit{and} \\ \forall\, y \in V(t,x) \cap \Psi^\partial(t,x),\ \forall\ u \in P(x,y), \\ 0\ \in\ DV(1, -f(x,y,u)) \cup D\Psi^c(+1, -f(x,y,u)) + g(x,y,u), \end{cases} \tag{20}$$

where $\mathrm{Graph}(\Psi^\circ) := \mathrm{Int}(\mathrm{Graph}(\Psi))$, $\mathrm{Graph}(\Psi^\partial) := \partial\mathrm{Graph}(\Psi)$*, and* $\mathrm{Graph}(\Psi^c) := X\backslash\mathrm{Graph}(\Psi)$*.*

Proof.

1. The graph of the caliber $V_\Psi(\Phi)$ being equal to the capture basin $\mathrm{Capt}_G^{\mathrm{Graph}(\Psi)}(\mathrm{Graph}(\Phi))$ by Proposition 2.3, we can apply the first part of Theorem 0.1 since $\mathrm{Graph}(\Psi) \subset \mathbf{R}_+ \times X \times Y$ is a repeller under G. Hence it is the largest graph of a set-valued map $V : X \rightsquigarrow Y$ between $\mathrm{Graph}(\Phi)$ and $\mathrm{Graph}(\Psi)$ such that, for any $(t,x,y) \in \mathrm{Graph}(V)\backslash\mathrm{Graph}(\Phi))$, i.e., whenever $y \in V(t,x)\backslash\Phi(t,x)$, there exists a control $u \in P(x)$ such that

$$(-1, f(x,y,u), -g(x,y,u))\ \in\ T_{\mathrm{Graph}(V)}(t,x,y). \tag{21}$$

In other words, it is the graph of the largest closed set-valued map V satisfying

$$\Phi(t,x) \ \subset \ V(t,x) \ \subset \ \Psi(t,x)$$

and, whenever $y \in V(t,x)\backslash\Phi(t,x)$, there exists $u \in P(x,y)$ such that

$$0 \ \in \ DV(-1, f(x,y,u)) - g(x,y,u).$$

2. The solutions $(T-\cdot, x(\cdot), y(\cdot))$ viable in the graph of Ψ until they reach the graph of Ψ satisfy: for almost all $t \geq 0$,

$$\left\{ \begin{array}{l} u(t) \ \in \ P(x(t),y(t)) \\ \text{and} \\ (-1, f(x(t),y(t),u(t)), g(x(t),y(t),u(t))) \ \in \ T_{\mathrm{Graph}(V)}(T-t, x(t), y(t)). \end{array} \right.$$

This condition can be rewritten as follows:

$$\text{for almost all } t \geq 0, \ \ u(t) \ \in \ R(T-t, x(t), y(t)).$$

3. Under the Lipschitz conditions, Theorem 0.2 states that the graph of $V_{_\Psi}(\Phi)$ is the unique closed subset $\mathrm{Graph}(V)$ satisfying the Frankowska properties:

$$\left\{ \begin{array}{ll} \text{(i)} & \forall\ (t,x,y) \in \mathrm{Graph}(V)\backslash\mathrm{Graph}(\Phi), \\ & \exists\ u \in P(x,y) \mid (-1, f(x,y,u), g(x,y,u)) \in T_{\mathrm{Graph}(V)}(t,x,y), \\ \text{(ii)} & \forall\ (t,x,y) \in \mathrm{Graph}(V) \cap \mathrm{Int}(\mathrm{Graph}(\Psi)), \forall\ u \in P(x,y), \\ & (1, -f(x,y,u), -g(x,y,u)) \ \in \ T_{\mathrm{Graph}(V)}(t,x,y), \\ \text{(iii)} & \forall\ (t,x,y) \in \mathrm{Graph}(V) \cap \partial(\mathrm{Graph}(\Psi)), \forall\ u \in P(x,y), \\ & (1, -f(x,y,u), -g(x,y,u)) \ \in \ T_{\mathrm{Graph}(V)}(t,x,y) \cup T_{X\backslash\mathrm{Graph}(\Psi)}(t,x,y). \end{array} \right.$$

Using the fact that $\mathrm{Graph}(DV)(t,x,y) = T_{\mathrm{Graph}(V)}(t,x,y)$, we infer the third part of the theorem. □

4. Stability properties. We state the following theorem of [12].

THEOREM 4.1. *Let us consider a sequence of closed subsets C_n satisfying* $\mathrm{Viab}(K) \subset C_n \subset K$ *and*

$$\mathrm{Lim}_{n\to+\infty} C_n \ := \ \mathrm{Limsup}_{n\to+\infty} C_n \ = \ \mathrm{Liminf}_{n\to+\infty} C_n.$$

If the set-valued map F is Marchaud and Lipschitz and if K is closed and backward invariant under F, then

$$\mathrm{Lim}_{n\to+\infty} \mathrm{Capt}^K(C_n) \ = \ \mathrm{Capt}^K(\mathrm{Lim}_{n\to+\infty} C_n). \tag{22}$$

Theorem 4.1 implies "continuity properties" of the caliber regarded as a map $\Phi \mapsto V_\Psi(\Phi)$.

For that purpose, let us recall the following definitions of graphical convergence (see [18] and/or [60], for instance):

1. the upper limit of the graphs

$$\mathrm{Limsup}_{n\to+\infty} \mathrm{Graph}(\Phi_n) \ =: \ \mathrm{Graph}(\mathrm{Lim}^\sharp{}_{n\to+\infty} \Phi_n)$$

is the graph of the *graphical upper limit* $\mathrm{Lim}^\sharp{}_{n\to+\infty}\Phi_n$ defined by

$$(\mathrm{Lim}^\sharp{}_{n\to+\infty}\Phi_n)(T,x) \ = \ \liminf_{n\to+\infty,\ s\to t,\ y\to x} V(\Phi_n)(s,y);$$

2. the lower limit of the graphs

$$\text{Liminf}_{n\to+\infty}\text{Graph}(\Phi_n) \;=:\; \text{Graph}(\text{Lim}^{\flat}{}_{n\to+\infty}\Phi_n)$$

is the graph of the *graphical lower limit*.

Then we derive the following "continuity" properties of the calibers.

THEOREM 4.2. *Let us consider a sequence of nontrivial set-valued maps* $\Psi_n : \mathbf{R}_+ \times X \rightsquigarrow Y \cup \{+\infty\}$ *and* $\Phi_n : \mathbf{R}_+ \times X \rightsquigarrow Y \cup \{+\infty\}$ *such that*

$$\forall\, (T,x) \in \mathbf{R}_+ \times X, \;\; \Phi_n(T,x) \;\subset\; \Psi_n(T,x).$$

The calibers also satisfy the following properties:

1. *Let us assume that the auxiliary control system* F *is Marchaud. Then*

$$\lim^{\sharp}{}_{n\to+\infty} V_{\Psi_n}(\Phi_n)(T,x) \;\subset\; V_{\lim^{\sharp}{}_{n\to+\infty}\Psi_n}(\lim^{\sharp}{}_{n\to+\infty}\Phi_n)(T,x).$$

2. *Let us assume that the auxiliary control system* F *is Marchaud and Lipschitz. Then*

$$\begin{cases} V(\lim^{\flat}{}_{n\to+\infty}\Phi_n)(T,x) \;\subset\; \lim^{\flat}{}_{n\to+\infty} V(\Phi_n)(T,x) \\ \subset\; \lim^{\sharp}{}_{n\to+\infty} V(\Phi_n)(T,x) \;\subset\; V(\lim^{\sharp}{}_{n\to+\infty}\Phi_n)(T,x). \end{cases}$$

Therefore, if the sequence of set-valued maps Φ_n *converges graphically to* Φ*, the caliber of the graphical limit is the graphical limit of the calibers.*

Proof. If the system is Marchaud, Theorem 4.1 implies that

$$\begin{aligned} &\text{Limsup}_{n\to+\infty}\text{Capt}_{\mathcal{R}}^{\text{Graph}(\Psi_n)}(\text{Graph}(\Phi_n)) \\ &\subset\; \text{Capt}_{\mathcal{R}}^{\text{Limsup}_{n\to+\infty}\text{Graph}(\Psi_n)}(\text{Limsup}_{n\to+\infty}\text{Graph}(\Phi_n)). \end{aligned}$$

Hence we deduce from Proposition 2.3, characterizing the graph of the caliber as a viable-capture basin, and from the definitions of graphical limits that

$$\lim^{\sharp}{}_{n\to+\infty} V_{\Psi_n}(\Phi_n)(T,x) \;\subset\; V_{\lim^{\sharp}{}_{n\to+\infty}\Psi_n}(\lim^{\sharp}{}_{n\to+\infty}\Phi_n)(T,x).$$

Under Lipschitz conditions of G, Theorem 4.1 implies that

$$\text{Capt}_{\mathcal{R}}(\text{Liminf}_{n\to+\infty}\text{Graph}(\Phi_n)) \;\subset\; \text{Liminf}_{n\to+\infty}\text{Capt}_{\mathcal{R}}(\text{Graph}(\Phi_n)).$$

Hence we deduce that

$$V(\lim^{\flat}{}_{n\to+\infty}\Phi_n)(T,x) \;\subset\; \lim^{\flat}{}_{n\to+\infty} V(\Phi_n)(T,x).$$

This completes the proof. □

5. Caliber of structured problems with linear exosystems. Of special interest is the particular case when the first differential equation does not depend upon the variable y and when the set-valued map $P : X \rightsquigarrow \mathcal{U}$ does not depend on the observation variable y: We thus obtain a *structured system*—as *age structured systems* in demography, when x plays the role of the age variable—of the form

$$\begin{cases} \text{(i)} & x'(t) \;=\; f(x(t),u(t)), \\ \text{(ii)} & y'(t) \;=\; g(x(t),y(t),u(t)), \\ \text{(iii)} & u(t) \;\in\; P(x(t)), \end{cases}$$

where $y(\cdot)$ is often regarded as an *observation* of the state (see, for instance, [14]). In control theory, the second controlled equation is called the *exosystem*.

We denote by $\mathcal{C} : x \in X \rightsquigarrow \mathcal{C}(x) \in \mathcal{C}(0,\infty;X) \times L^1(0,\infty;\mathcal{U})$ the set-valued map associating with $x \in X$ the set $\mathcal{C}(x)$ of the pairs $(x(\cdot),u(\cdot))$ of solutions to the control system

$$\begin{cases} \text{(i)} & x'(t) \ = \ f(x(t),u(t)), \\ \text{(ii)} & u(t) \ \in \ P(x(t)), \end{cases}$$

starting at x at $t = 0$.

We shall also set $g(x,y,u) := -M(x,u)y - L(x,u)$, where

1. M is a bounded continuous matrix-valued function

$$M : (x,u) \in X \times \mathcal{U} \mapsto M(x,u) \in \mathcal{L}(X,Y), \ \text{ and}$$

2. L is a continuous[7] "vector-Lagrangian"

$$L : (x,u) \in X \times \mathcal{U} \mapsto L(x,u) \in Y$$

with linear growth.

We introduce the map $(y;(x(\cdot),u(\cdot))) \rightsquigarrow \mathcal{S}(y;(x(\cdot),u(\cdot)))$ associating the subset of evolutions

$$y(t) \ = \ e^{-\int_0^t M(x(s),u(s))ds}\left(y - \int_0^t e^{\int_0^\tau M(x(s),u(s))ds} L(x(\tau),u(\tau))d\tau\right) \tag{23}$$

to the linear dynamical system

$$y'(t) \ = \ -M(x(t),u(t))y(t) - L(x(t),u(t))$$

starting at $y \in Y$.

We already know that the caliber is the unique Frankowska solution to the following:

1. there exists $u \in P(x,V(t,x))$ such that

$$0 \ \in \ \text{“} - \frac{\partial}{\partial t}V(t,x) + \frac{\partial}{\partial x}V(t,x)f(x,V(t,x),u)\text{”} - M(x,u)V(t,x) - L(x,u);$$

2. for all $u \in P(x,v(t,x))$;

$$0 \ \in \ \text{“}\frac{\partial}{\partial t}V(t,x) - \frac{\partial}{\partial x}V(t,x)f(x,V(t,x),u)\text{”} + M(x,u)V(t,x) + L(x,u),$$

satisfying the viability constraints

$$\forall\, t \geq 0,\ x \in \ X, \ \ \Phi(t,x) \ \subset \ V(t,x) \ \subset \ \Psi(t,x)$$

and the initial condition

$$V(0,x) \ = \ \Phi(0,x).$$

[7] We could take L to be set-valued map, but we restrict our attention to the single-valued case for simplicity.

In summary, we now deal with a *structured problem* where the exosystem is linear with respect to the observations. In this case, we shall be able to provide an explicit formula of the caliber V.

For that purpose, we introduce the subset

$$\left\{\begin{array}{l} J_\Phi(t;(x(\cdot),u(\cdot)))(T,x) \\ \\ := \ e^{\int_0^t M(x(s),u(s))ds}\Phi(T-t,x(t)) + \displaystyle\int_0^t e^{\int_0^\tau M(x(s),u(s))ds}L(x(\tau),u(\tau))d\tau \end{array}\right.$$

(where t ranges over $[0,T]$). The controls—most often prices or other regulees in economics and portfolios in finance—appear *both* in the matrix M and in the Lagrangian L.

We associate with Ψ the set-valued map J_Ψ defined by

$$\left\{\begin{array}{l} J_\Psi(t;(x(\cdot),u(\cdot)))(T,x) \\ \\ := \ e^{\int_0^t M(x(s),u(s))ds}\Psi(T-t,x(t)) + \displaystyle\int_0^t e^{\int_0^\tau M(x(s),u(s))ds}L(x(\tau),u(\tau))d\tau \end{array}\right.$$

and

$$K_\Psi(t,x;(x(\cdot),u(\cdot))) \ := \ \bigcap_{s\in[0,t]} J_\Psi(s,x;(x(\cdot),u(\cdot))).$$

We next introduce

$$L^\Phi_\Psi(t;(x(\cdot),u(\cdot)))(T,x) \ := \ K_\Psi(t,x;(x(\cdot),u(\cdot))) \cap J_\Phi(t;(x(\cdot),u(\cdot)))(T,x).$$

5.1. Explicit formula of the caliber. We shall prove now that the caliber V of the pair (Ψ,Φ) under $\mathcal{B}$ is equal to the set-valued map $V_\Psi(\Phi)$ defined by

$$V_\Psi(\Phi)(T,x) \ := \bigcup_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)} \ \bigcup_{t\in[0,T]} L^\Phi_\Psi(t;(x(\cdot),u(\cdot)))(T,x). \tag{24}$$

Note that if $\Psi_1 \subset \Psi_2$ and $\Phi_1 \subset \Phi_2$, then $V_{\Psi_1}(\Phi_1) \subset V_{\Psi_2}(\Phi_2)$ and

$$\forall\, t \geq 0, \forall\, x \in X, \ \ \Phi(t,x) \ \subset \ V_\Psi(\Phi)(t,x) \ \subset \ \Psi(t,x).$$

We shall use the fact that its graph is the viable-capture basin of the graph of the cost function Φ viable under the graph of Ψ under the auxiliary system $\mathcal{R}$.

THEOREM 5.1. *Let us assume that the set-valued maps Ψ and Φ are nontrivial. Then the caliber of the pair (Ψ,Φ) is equal to the set-valued map $V_\Psi(\Phi)$ defined by* (24).

Furthermore, the caliber is the unique solution *V to the two following "fixed set equations":*

$$\left\{\begin{array}{rcl} V_V(\Phi)(T,x) & := & \bigcup_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)} \bigcup_{t\in[0,T]} L^\Phi_V(t,x;(x(\cdot),u(\cdot))) \\ = & & V(T,x) \\ = & & V_\Psi(V)(T,x) \ := \ \bigcup_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)} \bigcup_{t\in[0,T]} L^V_\Psi(t,x;(x(\cdot),u(\cdot))). \end{array}\right. \tag{25}$$

Moreover, when F is Marchaud and the set-valued maps Φ and Ψ are closed, the graph of the caliber $V_\Psi(\Phi)$ is closed.

5.2. Examples.

1. We see that the set-valued map defined by

$$V(\Phi)(T,x) \;:=\; \bigcup_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)} \bigcup_{t\in[0,T]} J_\Phi(t;(x(\cdot),u(\cdot)))(T,x) \;=\; V_{\mathbf{Y}}(\Phi)$$

is the caliber of (Ψ,Φ), where the set-valued map $\Psi=\mathbf{Y}$ is defined by

$$\mathbf{Y}(t,x) \;:=\; \begin{cases} Y & \text{if} \quad t\geq 0,\\ \emptyset & \text{if} \quad t<0. \end{cases}$$

Indeed, we observe that, taking $\Psi=\mathbf{Y}$,

$$K_{\mathbf{Y}}(t,x;(x(\cdot),u(\cdot))) \;=\; J_{\mathbf{Y}}(t,x;(x(\cdot),u(\cdot))) \;=\; Y$$

so that, for any set-valued map Φ, we have $L^\Phi_{\mathbf{Y}}=J_\Phi$ and thus $V(\Phi)=V_{\mathbf{Y}}(\Phi)$.

2. We also observe that the set-valued map

$$W(\Psi)(T,x) \;:=\; \bigcup_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)} \bigcap_{t\in[0,T]} J_\Psi(t;(x(\cdot),u(\cdot)))(T,x) \;=\; V_\Psi(\Psi_\emptyset)$$

is the caliber of (Ψ,Φ), where we take $\Phi:=\Psi_\emptyset:\mathbf{R}_+\times X\rightsquigarrow Y$ defined by

$$\Psi_\emptyset(t,x) \;:=\; \begin{cases} \Psi(0,x) & \text{if} \quad t=0,\\ \emptyset & \text{if} \quad t\neq 0. \end{cases}$$

Indeed, we see that $J_{\Psi_\emptyset}(t,x;(x(\cdot),u(\cdot)))=\emptyset$ if $t<T$ and $J_{\Psi_\emptyset}(T,x;(x(\cdot),u(\cdot)))=J_\Psi(T,x;(x(\cdot),u(\cdot)))$. Therefore,

$$L^{\Psi_\emptyset}_\Psi(t,x;(x(\cdot),u(\cdot))) \;:=\; \begin{cases} K_\Psi(T,x;(x(\cdot),u(\cdot))) & \text{if} \quad t=T,\\ \emptyset & \text{if} \quad t<T, \end{cases}$$

and thus $W(\Psi)=V_\Psi(\Psi_\emptyset)$.

3. Let us introduce a time-independent set-valued map $U:X\rightsquigarrow X$. We shall associate with three pairs (Ψ,Φ) of set-valued maps associated with U the three following calibers:

   (a) Taking $\Phi:=U_\emptyset$ and $\Psi=\mathbf{Y}$, we obtain

$$\bigcup_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)} \Bigg(e^{\int_0^T M(x(s),u(s))ds} U(x(T)) + \int_0^T e^{\int_0^\tau M(x(s),u(s))ds} L(x(\tau),u(\tau))d\tau \Bigg).$$

   When $M=0$, the above problem boils down to the set-valued equivalent of the *Bolza map*

$$\bigcup_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)} \Bigg(U(x(T)) + \int_0^T L(x(\tau),u(\tau))d\tau \Bigg)$$

   and the *Mayer map*

$$\bigcup_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)} U(x(T))$$

   when, furthermore, $L=0$. This is the case of the substratum of an impulse differential inclusion defined above (see also [9]).

(b) Taking $\Phi := U_\emptyset$ and $\Psi := U$, we obtain

$$\bigcup_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)} \bigcap_{t\in[0,T]} \left(e^{\int_0^t M(x(s),u(s))ds} U(x(t)) + \int_0^t e^{\int_0^\tau M(x(s),u(s))ds} L(x(\tau),u(\tau))d\tau \right).$$

(c) Taking $\Phi := U$ and $\Psi := \mathbf{Y}$, we obtain

$$\bigcup_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)} \bigcup_{t\in[0,T]} \left(e^{\int_0^t M(x(s),u(s))ds} U(x(t)) + \int_0^t e^{\int_0^\tau M(x(s),u(s))ds} L(x(\tau),u(\tau))d\tau \right).$$

When $M = 0$, we find

$$V(U)(T,x) := \bigcup_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)} \bigcup_{t\in[0,T]} \left(U(x(t)) + \int_0^t L(x(\tau),u(\tau))d\tau \right).$$

This map—the set-valued analogue of the valuation function of "obstacle problems"—involves "max-plus" operations and is the set-valued equivalent of the *mathematical fear or faith* with respect to a Maslov measure introduced in dynamical optimization by Pierre Bernhard (see [35, 36, 37]).

Regarding the caliber as a transform $\Phi \mapsto V(\Phi)$ mapping closed set-valued maps to closed set-valued map, we observe that the caliber satisfies

$$V\left(\bigcup_{i=1,\dots,n} \Phi_i\right) = \bigcup_{i=1,\dots,n} V(\Phi_i),$$

the *extensivity* property, $\Phi \subset V(\Phi)$, and the *monotonicity* property: if $\Phi_1 \subset \Phi_2$, then $V(\Phi_1) \subset V(\Phi_2)$.

We recall that Theorem 4.2 implies that the value transform is also "upper continuous" in the sense that

$$\text{Limsup}_{n\to+\infty,\, s\to t,\, y\to x} V(\Phi_n)(s,y) \subset V\left(\text{Limsup}_{n\to+\infty,\, s\to t,\, y\to x} \Phi_n(s,y)\right)(t,x)$$

and actually, under stronger assumptions, "continuous" in the sense that the caliber of a graphical limit is the graphical limit of calibers.

5.3. Proof of the explicit formula. First, to say that a pair (T,x,y) belongs to the viable-capture basin

$$\text{Graph}(V) := \text{Capt}_{\mathcal{R}}^{\text{Graph}(\Psi)}(\text{Graph}(\Phi) \subset \text{Graph}(\Psi)$$

means that there exist a solution $(T-\cdot, \widetilde{x}(\cdot), \widetilde{u}(\cdot), y(\cdot)) \in \mathcal{R}(T,x,y)$ to the auxiliary control problem and some $t^* \geq 0$ such that

$$\begin{cases} \text{(i)} \quad \forall\, t \in [0,t^*], \ (T-t, \widetilde{x}(t), y(t)) \in \text{Graph}(\Psi), \\ \text{(ii)} \quad (T-t^*, \widetilde{x}(t^*), y(t^*)) \in \text{Graph}(\Phi) \end{cases}$$

or, equivalently, such that

$$\begin{cases} \text{(i)} & \forall\, t \in [0,t^*],\ \ y(t) \ \in\ \Psi(T-t,\widetilde{x}(t)), \\ \text{(ii)} & y(t^*) \ \in\ \Psi(T-t^*,\widetilde{x}(t^*). \end{cases}$$

By the very definition of $\mathcal{R}(T,x,y)$ and by the definition (23) of the auxiliary system that its component $y(\cdot)$ satisfies

$$y(t) \ =\ e^{-\int_0^t M(\widetilde{x}(s),\widetilde{u}(s))ds}\left(y-\int_0^t e^{\int_0^\tau M(\widetilde{x}(s),\widetilde{u}(s))ds}L(\widetilde{x}(\tau),\widetilde{u}(\tau))d\tau\right),$$

this implies that (T,x,y) belongs to $\mathrm{Capt}_{\mathcal{R}}^{\mathrm{Graph}(\Psi)}(\mathrm{Graph}(\Phi)$ if and only if there exists a solution $(\widetilde{x}(\cdot),\widetilde{u}(\cdot))\in\mathcal{C}(x)$ satisfying, for almost all $t\in[0,\overline{t}]$,

$$\begin{cases} \text{(i)} & y \in e^{\int_0^t M(\widetilde{x}(s),\widetilde{u}(s))ds}\Psi(T-t,\widetilde{x}(t),\widetilde{u}(t))+\displaystyle\int_0^t e^{\int_0^\tau M(\widetilde{x}(s),\widetilde{u}(s))ds}L(\widetilde{x}(\tau),\widetilde{u}(\tau))d\tau, \\ \text{(ii)} & y \in e^{\int_0^{t^*} M(\widetilde{x}(s),\widetilde{u}(s))ds}\Phi(T-t^*,\widetilde{x}(t^*))+\displaystyle\int_0^{t^*} e^{\int_0^\tau M(\widetilde{x}(s),\widetilde{u}(s))ds}L(\widetilde{x}(\tau),\widetilde{u}(\tau))d\tau. \end{cases}$$

(26)

Since we set

$$K_\Psi(t,x;(\widetilde{x}(\cdot),\widetilde{u}(\cdot)))\ :=\ \bigcap_{s\in[0,t]} J_\Psi(s,x;(\widetilde{x}(\cdot),\widetilde{u}(\cdot)))$$

and

$$L_\Psi^\Phi(t,x;(\widetilde{x}(\cdot),\widetilde{u}(\cdot)))\ :=\ K_\Psi(t,x;(\widetilde{x}(\cdot),\widetilde{u}(\cdot)))\cap J_\Phi(t,x;(\widetilde{x}(\cdot),\widetilde{u}(\cdot))),$$

this is equivalent to state that y belongs to $V_\Psi(\Phi)(T,x)$.

Formula (3) of [15] and the first part of this theorem implies that the caliber is the unique solution to the system (25).

Finally, the closedness of the graph of the caliber is an immediate consequence of Theorem 0.1 because $\mathrm{Graph}(\Psi)\subset\mathbf{R}_+\times X\times Y$ is a repeller under G, which is Marchaud, and the viable-capture basin $\mathrm{Graph}(V_\Psi)(\Phi))\ :=\ \mathrm{Capt}_{\mathcal{R}}^{\mathrm{Graph}(\Psi)}(\mathrm{Graph}(\Phi))$ is closed. □

5.4. The barrier property of the caliber. We begin with a first form of the *barrier property.*

Proposition 5.2. *Let us consider $y_T\in\partial_{\Psi(T,x)}V_\Psi(\Phi)(T,x)$. Then, any solution $(x(\cdot),u(\cdot))\in\mathcal{C}(x)$ starting from x satisfying the following inclusion: for every $t\in[0,t^*]$,*

$$y_T\ \in\ L_\Psi^\Phi(t;(x(\cdot),u(\cdot)))(T,x) \tag{27}$$

until the first time t^* when

$$y_T\in\Phi(T-t^*,x(t^*)$$

actually satisfies the following: for every $t\in[0,t^*]$,

$$\forall\, y_T\ \in\ \partial L_\Psi^\Phi(t;(x(\cdot),u(\cdot)))(T,x).$$

Proof. Since y_T belongs to $V_\Psi(\Phi)(T,x)$, there exists a solution $(x^*(\cdot),u^*(\cdot))$ and a time $t^* \in [0,T]$ such that

$$y_T \ \in \ L^\Phi_\Psi(t^*,x;(x^*(\cdot),u^*(\cdot))).$$

Since y_T belongs to $\partial_{\Psi(T,x)}V_\Psi(\Phi)(T,x)$, it can be approximated by elements $y_n \in \Psi(T,x)\backslash V_\Psi(\Phi)(T,x)$. We know that, for every t and any solution $(x(\cdot),u(\cdot))$, either y_n does not belong to $K_\Psi(t,x;(x(\cdot),u(\cdot)))$ or it belongs to $K_\Psi(t,x;(x(\cdot),u(\cdot)))\backslash V(T,x)$.

We claim that

$$\left\{\begin{array}{l} \forall\ (x(\cdot),u(\cdot)) \in \mathcal{C}(x),\ \forall\ t \in [0,T], \\ \\ K_\Psi(t,x;(x(\cdot),u(\cdot)))\backslash V(T,x) \ \subset \ X\backslash J_V(t,x;(x(\cdot),u(\cdot))). \end{array}\right. \tag{28}$$

Indeed, take y in $K_\Psi(t,x;(x(\cdot),u(\cdot)))\backslash V(T,x)$. This means that, for every $s \in [0,t]$, y belongs to $J_\Psi(s,x;(x(\cdot),u(\cdot)))$ or, equivalently,

$$\forall\ s \in [0,t],\ \ (T-s,x(s),y(s)) \ \in \ \mathrm{Graph}(\Psi),$$

where

$$y(t) \ := \ e^{-\int_0^t M(x(s),u(s))ds}\left(y - \int_0^t e^{\int_0^\tau M(x(s),u(s))ds}L(x(\tau),u(\tau))d\tau\right).$$

Since (T,x,y) does not belong to the viable capture basin, we infer that every solution in $\mathcal{R}(T,x,y)$ starting at (T,x,y) is viable in the graph $\mathrm{Graph}(\Psi)$ of Ψ before hitting the graph $\mathrm{Graph}(V)$ of V: Therefore, if

$$\forall\ s \in [0,t],\ \ y(s) \ \in \ \Psi(T-s,x(s),u(s)),$$

we have

$$y(t) \ \in \ X\backslash V(T-t,x(t)) \ \subset \ X\backslash\Phi(T-t,x(t)),$$

which can be written in the form

$$y \ \in \ X\backslash J_V(t,x;(x(\cdot),u(\cdot))) \ \subset \ X\backslash J_\Phi(t,x;(x(\cdot),u(\cdot))).$$

Therefore, inclusion (28) holds true, and thus y_n does not belong to $J_V(t,x;(x(\cdot),u(\cdot)))$. Consequently, for any $t \in [0,T]$ and for any $n \geq 0$,

$$y_n \ \in \ X\backslash L^V_\Psi(t,x;(x(\cdot),u(\cdot))) \ \subset \ X\backslash L^\Phi_\Psi(t,x;(x(\cdot),u(\cdot))).$$

This is particularly true for the solution $(x^*(\cdot),u^*(\cdot))$ and for any $t \in [0,t^*]$. Hence, letting y_n converge to y_T, we deduce the conclusion of the proposition. □

We can strengthen this result and prove the following "barrier property." Let us recall that $J_{V_\Psi(\Phi)}(0,x;(x(\cdot),u(\cdot))) = V_\Psi(\Phi)(T,x)$. We shall prove that the boundary condition

$$y_T \ \in \partial V_\Psi(\Phi)(T,x) \ = \ \partial J_{V_\Psi(\Phi)}(0,x;(x(\cdot),u(\cdot)))$$

propagates as long as y_T remains in the interior of $K_\Psi(t,x;(x(\cdot),u(\cdot)))$.

THEOREM 5.3. *Let us consider $y_T \in \partial V_\Psi(\Phi)(T,x) \cap \mathrm{Int}(\Psi(T,x))$. Then, any solution $(x(\cdot),u(\cdot)) \in \mathcal{C}(x)$ starting from x satisfying the inclusion*

$$y_T \ \in \ J_{V_\Psi(\Phi)}(t,x;(x(\cdot),u(\cdot))) \tag{29}$$

until the first time $\overline{t}$ *when*

$$y_T \notin \mathrm{Int}(K_\Psi(\overline{t}, x; (x(\cdot), u(\cdot))))$$

actually satisfies

$$\forall\, t \in [0, \overline{t}],\ \ y \in \partial J_{V_\Psi(\Phi)}(t; (x(\cdot), u(\cdot)))(T, x). \tag{30}$$

Proof. Since y_T belongs to $\partial V_\Psi(\Phi)(T, x)$, it can be approximated by elements $y_n \in \Psi(T, x) \backslash V_\Psi(\Phi)(T, x)$. By assumption, for any $t < \overline{t}$, there exists δ_t such that

$$B(y_T, \delta_t) \subset K_\Psi(t; (x(\cdot), u(\cdot)))(T, x),$$

and thus there exists some N_t such that, for any $n \geq N_t$, y_n belongs to $B(y^T, \delta_t)$, and thus

$$y_n \in K_\Psi(t; (x(\cdot), u(\cdot)))(T, x) \backslash V_\Psi(\Phi)(T, x).$$

However, by (28), we know that, in this case, y_n does not belong to $J_{V_\Psi(\Phi)}(t; (x(\cdot), u(\cdot)))(T, x)$. □

6. Frankowska contingent episolutions to Hamilton–Jacobi–Bellman equations. When $Y := \mathbf{R}$, we can associate with two extended functions $\mathbf{c} : \mathbf{R}_+ \times X \rightsquigarrow \mathbf{R} \cup \{+\infty\}$ and $\mathbf{b} : \mathbf{R}_+ \times X \rightsquigarrow \mathbf{R} \cup \{+\infty\}$ such that

$$\forall\, (t, x) \in \mathbf{R}_+ \times X,\ \ 0 \leq \mathbf{b}(t, x) \leq \mathbf{c}(t, x)$$

the set-valued maps Φ and Ψ defined by

$$\begin{cases} \text{(i)} & \Phi(t, x) := \mathbf{c}(t, x) + \mathbf{R}_+, \\ \text{(ii)} & \Psi(t, x) := \mathbf{b}(t, x) + \mathbf{R}_+ \end{cases}$$

by setting $\Phi(t, x) := \emptyset$ whenever $\mathbf{c}(t, x) = +\infty$. We observe that $\mathrm{Graph}(\Phi) = \mathcal{E}p(\mathbf{c})$ and that

$$D\Phi(t, x, \mathbf{c}(t, x))(\pm 1, v) = D_\uparrow \mathbf{c}(t, x)(\pm 1, v) + \mathbf{R}_+,$$

where $D_\uparrow \mathbf{c}(t, x)$ is the *contingent epiderivative* of $\mathbf{c}$ at (t, x) in the direction $(\pm 1, v)$, defined by

$$D_\uparrow \mathbf{c}(t, x)(\pm 1, v) = \liminf_{h \to 0+, v' \mapsto v} \frac{\mathbf{c}(t \pm h, x + hv') - \mathbf{c}(t, x)}{h}.$$

We set

$$\begin{cases} J_{\mathbf{c}}(t; (x(\cdot), u(\cdot)))(T, x) \\ \\ := e^{\int_0^t M(x(s), u(s))ds} \mathbf{c}(T - t, x(t)) + \displaystyle\int_0^t e^{\int_0^\tau M(x(s), u(s))ds} L(x(\tau), u(\tau))d\tau, \end{cases}$$

and then

$$K_{\mathbf{b}}(t, x; (x(\cdot), u(\cdot))) := \sup_{s \in [0, t]} J_{\mathbf{b}}(s, x; (x(\cdot), u(\cdot))).$$

We next integrate this cumulated cost together with the former cost $J_{\mathbf{c}}(t,x;(x(s),u(s)))$ by introducing the new cost function

$$L^{\mathbf{c}}_{\mathbf{b}}(t;(x(\cdot),u(\cdot)))(T,x) \;:=\; \max(K_{\mathbf{b}}(t,x;(x(\cdot),u(\cdot))),J_{\mathbf{c}}(t;(x(\cdot),u(\cdot)))(T,x).$$

We shall deduce from Theorem 5.1 the following consequence.

THEOREM 6.1. *Let us assume that the extended functions* $\mathbf{b}$ *and* $\mathbf{c}$ *are nontrivial and nonnegative. The viable-capture basin* $\mathrm{Capt}^{\mathcal{E}p(\mathbf{b})}_{\mathcal{R}}(\mathcal{E}p(\mathbf{c}))$ *of* $\mathcal{E}p(\mathbf{c})$ *under* $\mathcal{R}$ *is the epigraph of the valuation function* $V_{\mathbf{b}}(\mathbf{c})$ *defined by*

$$V_{\mathbf{b}}(\mathbf{c})(T,x) \;:=\; \inf_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)}\ \inf_{t\in[0,T]} L^{\mathbf{c}}_{\mathbf{b}}(t;(x(\cdot),u(\cdot)))(T,x).$$

Furthermore, any solution $(x(\cdot),u(\cdot))\in\mathcal{C}(x)$ *starting from* $x\in X$ *satisfying the following inequality: for every* $t\in[0,t^*]$

$$(31)\quad \left\{\begin{array}{l} V_{\mathbf{b}}(\mathbf{c})(T,x) \\ \geq\; e^{\int_0^t M(x(s),u(s))ds}V_{\mathbf{b}}(\mathbf{c})(T-t,x(t)) + \displaystyle\int_0^t e^{\int_0^\tau M(x(s),u(s))ds}L(x(\tau),u(\tau))d\tau \end{array}\right.$$

until the first time t^* *when*

$$V_{\mathbf{b}}(\mathbf{c})(T-t^*,x(t^*)) = \mathbf{c}(T-t^*,x(t^*))$$

is an optimal solution for the optimal time t^*.

Finally, the valuation function is a solution $\mathbf{v}$ *to the two following functional equations stating that the functions* $L^{\mathbf{v}}_{\mathbf{b}}$ *and* $L^{\mathbf{c}}_{\mathbf{v}}$ *have the same infimum as* $L^{\mathbf{c}}_{\mathbf{b}}$:

$$(32)\quad \left\{\begin{array}{l} \inf_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)}\inf_{t\in[0,T]} L^{\mathbf{c}}_{\mathbf{v}}(t,x;(x(\cdot),u(\cdot))) \\ =\; \mathbf{v}(T,x) \\ =\; \inf_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)}\inf_{t\in[0,T]} L^{\mathbf{v}}_{\mathbf{b}}(t,x;(x(\cdot),u(\cdot))). \end{array}\right.$$

Proof. It is enough—and easy—to check that

$$L^{\Phi}_{\Psi}(t;(x(\cdot),u(\cdot)))(T,x) \;=\; L^{\mathbf{c}}_{\mathbf{b}}(t;(x(\cdot),u(\cdot)))(T,x) + \mathbf{R}_+$$

and thus that

$$\inf_{y\in V^{\Phi}_{\Psi}(t,x)} y \;=\; V^{\mathbf{c}}_{\mathbf{b}}(t,x)$$

since $\inf\left(\bigcup_{i\in I}[a_i,\infty[\right) = \inf_{i\in I} a_i$. □

Theorem 5.3 implies the following form of the optimality principle.

THEOREM 6.2. *Let us assume that* $V_{\mathbf{b}}(\mathbf{c})(T,x) > \mathbf{b}(T,x)$. *Then any solution* $(x(\cdot),u(\cdot))\in\mathcal{C}(x)$ *starting from* $x\in\mathrm{Dom}(V_{\mathbf{b}}(\mathbf{c}))$ *satisfying inequalities*

$$(33)\quad \left\{\begin{array}{l} V_{\mathbf{b}}(\mathbf{c})(T,x) \\ \geq\; e^{\int_0^t M(x(s),u(s))ds}V_{\mathbf{b}}(\mathbf{c})(T-t,x(t)) + \displaystyle\int_0^t e^{\int_0^\tau M(x(s),u(s))ds}L(x(\tau),u(\tau))d\tau \end{array}\right.$$

until the first time $t^*\in[0,T]$ *when*

$$V_{\mathbf{b}}(\mathbf{c})(T-t^*,x(t^*) = \mathbf{b}(T-t^*,x(t^*))$$

actually satisfies equality

$$\left\{\begin{array}{l} \forall\, t \in [0,t^*],\ V_{\mathbf{b}}(\mathbf{c})(T,x) \\ \\ =\ e^{\int_0^t M(x(s),u(s))ds} V_{\mathbf{b}}(\mathbf{c})(T-t,x(t)) + \displaystyle\int_0^t e^{\int_0^\tau M(x(s),u(s))ds} L(x(\tau),u(\tau))d\tau. \end{array}\right. \tag{34}$$

The first statement of Theorem 3.1 implies that the valuation function is a Frankowska contingent episolution to Hamilton–Jacobi–Bellman equations.

THEOREM 6.3 (Frankowska). *Let us assume that the control system* (P,f,L,M) *is Marchaud and that the functions* $\mathbf{b}$ *and* $\mathbf{c}$ *are nontrivial, nonnegative, and lower semicontinuous.*

Then the valuation function $V_{\mathbf{b}}(\mathbf{c})$ *is characterized as the smallest of the nonnegative lower semicontinuous functions* $\mathbf{v} : \mathbf{R}_+ \times X \mapsto \mathbf{R}_+ \cup \{+\infty\}$ *satisfying, for every* $(t,x) \in]0,\infty[\times X$*,*

$$\left\{\begin{array}{ll} \text{(i)} & \mathbf{b}(t,x) \ \le\ \mathbf{v}(t,x) \ \le\ \mathbf{c}(t,x), \\ \text{(ii)} & \textit{if } (t,x) \in \Omega(\mathbf{v}), \\ & \inf_{u\in P(x)} \left(D_\uparrow \mathbf{v}(t,x)(-1,f(x,u)) + L(x,u) + M(x,u)\mathbf{v}(t,x)\right) \ \le\ 0. \end{array}\right.$$

Let us set

$$\mathbf{R}(t,x) \ :=\ \{u \in P(x) \mid D_\uparrow V_{\mathbf{b}}(\mathbf{c})(t,x)(-1,f(x,u)) + L(x,u) + M(x,u)V_{\mathbf{b}}(\mathbf{c})(t,x) \le 0\}.$$

Knowing the valuation function, an optimal solution is obtained in the following way: Starting from x_0 *such that* $V_{\mathbf{b}}(\mathbf{c})(T,x_0) < \mathbf{c}(T,x_0)$*, any optimal solution* $(x(\cdot),u(\cdot))$ *is a solution to the control system*

$$\left\{\begin{array}{ll} \text{(i)} & x'(t) \ =\ f(x(t),u(t)), \\ \text{(ii)} & u(t) \ \in\ \mathbf{R}(T-t,x(t)) \end{array}\right. \tag{35}$$

until the first time $t^* \ge 0$ *when*

$$V_{\mathbf{b}}(\mathbf{c})(T-t^*,x(t^*)) \ =\ \mathbf{c}(T-t^*,x(t^*)).$$

The second part of Theorem 3.1 implies the following existence and uniqueness result:

THEOREM 6.4 (Frankowska). *Let us assume that the control system* (P,f,L,M) *is Marchaud and Lipschitz and that* $\mathbf{b}$ *and* $\mathbf{c}$ *are nontrivial, nonnegative, and lower semicontinuous.*

Then the valuation function $V_{\mathbf{b}}(\mathbf{c})$ *is the unique lower semicontinuous* episolution $\mathbf{v}$ *to the following system of differential inequalities: for every* $(t,x) \in \mathrm{Dom}(\mathbf{v})$*,*

$$\left\{\begin{array}{ll} \text{(i)} & \mathbf{b}(t,x) \ \le\ \mathbf{v}(t,x) \ \le\ \mathbf{c}(t,x), \\ \text{(ii)} & \textit{if } \mathbf{v}(t,x) < \mathbf{c}(t,x), \\ & \inf_{u\in P(x)} \left(D_\uparrow \mathbf{v}(t,x)(-1,f(x,u)) + L(x,u) + M(x,u)\mathbf{v}(t,x)\right) \ \le\ 0, \\ \text{(iii)} & \textit{if } \mathbf{v}(t,x) > \mathbf{b}(t,x), \\ & \sup_{u\in P(x)} (D_\uparrow \mathbf{v}(t,x)(1,-f(x,u)) - L(x,u) - M(x,u)\mathbf{v}(t,x)) \ \le\ 0, \\ \text{(iv)} & \textit{if } \mathbf{v}(t,x) = \mathbf{b}(t,x), \\ & \sup_{u\in P(x)} [\min(D_\uparrow \mathbf{v}(t,x)(1,-f(x,u)), D_\downarrow \mathbf{b}(t,x)(1,-f(x,u))) \\ & -L(x,u) - M(x,u)\mathbf{v}(t,x)] \ \le\ 0. \end{array}\right. \tag{36}$$

Remark. Condition (36) (iv) is automatically satisfied whenever

$$\sup_{u\in P(x)} (D_\downarrow \mathbf{b}(t,x)(1,-f(x,u)) - L(x,u) - M(x,u)\mathbf{v}(t,x)) \ \leq\ 0.$$

We refer to the papers [43, 41, 42, 44, 45, 46, 47] for other differential properties of the value function obtained using the tools of the epigraphical approach and, in particular, by duality, the links with viscosity solutions and lower semicontinuous bilateral solutions also introduced in [32, 33] by PDE methods.

7. Vector optimal control problems. When $Y := \mathbf{R}^n$ is supplied with the natural order relation $\leq$ associated with the positive orthant $\mathbf{R}^n_+$, we can associate with two maps $\vec{\mathbf{c}} : \mathbf{R}_+ \times X \rightsquigarrow \mathbf{R}^n_+$ and $\vec{\mathbf{b}} : \mathbf{R}_+ \times X \rightsquigarrow \mathbf{R}^n_+$ such that

$$\forall\ (t,x) \in \mathbf{R}_+ \times X,\ \ 0 \ \leq\ \vec{\mathbf{b}}(t,x) \ \leq\ \vec{\mathbf{c}}(t,x)$$

the set-valued maps Φ and Ψ defined by

$$\begin{cases} \text{(i)} & \Phi(t,x) \ := \ \vec{\mathbf{c}}(t,x) + \mathbf{R}^n_+, \\ \text{(ii)} & \Psi(t,x) \ := \ \vec{\mathbf{b}}(t,x) + \mathbf{R}^n_+. \end{cases}$$

In the following formulas, the supremums and the infimums are taken component by component. We set

$$\begin{cases} J_{\vec{\mathbf{c}}}(t;(x(\cdot),u(\cdot)))(T,x) \\ \\ := \ e^{\int_0^t M(x(s),u(s))ds}\vec{\mathbf{c}}(T-t,x(t)) + \displaystyle\int_0^t e^{\int_0^\tau M(x(s),u(s))ds} L(x(\tau),u(\tau))d\tau, \end{cases}$$

and then

$$K_{\vec{\mathbf{b}}}(t,x;(x(\cdot),u(\cdot))) \ := \ \sup_{s\in[0,t]} J_{\vec{\mathbf{b}}}(s,x;(x(\cdot),u(\cdot))).$$

We next integrate this cumulated cost together with the former cost $J_{\vec{\mathbf{c}}}(t,x;(x(s),u(s)))$ by introducing the new cost function

$$L^{\vec{\mathbf{c}}}_{\vec{\mathbf{b}}}(t;(x(\cdot),u(\cdot)))(T,x) \ := \ \max(K_{\vec{\mathbf{b}}}(t,x;(x(\cdot),u(\cdot))), J_{\vec{\mathbf{c}}}(t;(x(\cdot),u(\cdot)))(T,x)$$

and the subset

$$V_{\vec{\mathbf{b}}}(\vec{\mathbf{c}})(T,x) \ := \bigcup_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)} \ \bigcup_{t\in[0,T]} L^{\vec{\mathbf{c}}}_{\vec{\mathbf{b}}}(t;(x(\cdot),u(\cdot)))(T,x).$$

We shall deduce from Theorem 5.1 the following consequence.

THEOREM 7.1. *Let us assume that, for any $(x,u) \in X \times \mathcal{U}$ and for any $y \in \mathbf{R}^n_+$, whenever $y_i = 0$, then $(M(x,u)y)_i = 0$.*

Then

$$\forall\ T \geq 0, x \in X,\ \ V_\Psi(\Phi)(T,x) \ = \ V_{\vec{\mathbf{b}}}(\vec{\mathbf{c}})(T,x) + \mathbf{R}^n_+.$$

Proof. It is enough to check that

$$L^\Phi_\Psi(t;(x(\cdot),u(\cdot)))(T,x) \ = \ L^{\vec{\mathbf{c}}}_{\vec{\mathbf{b}}}(t;(x(\cdot),u(\cdot)))(T,x) + \mathbf{R}^n_+.$$

By assumption, the cone $\mathbf{R}^n_+$ is forward and backward invariant under the differential equation $y'(t) = M(x(t), u(t))y(t)$ so that

$$\forall\, y \in \mathbf{R}^n_+,\ \ e^{\pm \int_0^t M(x(\tau),u(\tau))d\tau} y \ \in\ \mathbf{R}^n_+.$$

Therefore, since $\vec{\mathbf{c}}(t,x) \in \mathbf{R}^n_+$, we infer that

$$J_\Phi(t;(x(\cdot),u(\cdot)))(T,x) \ =\ J_{\vec{\mathbf{c}}}(t;(x(\cdot),u(\cdot)))(T,x) + \mathbf{R}^n_+.$$

Next we observe that, if $a_i \in \mathbf{R}^n$,

$$\bigcap_{i\in I}(a_i + \mathbf{R}^n_+) \ =\ \sup_{i\in I} a_i + \mathbf{R}^n_+,$$

and thus

$$K_\Psi(t;(x(\cdot),u(\cdot)))(T,x) \ =\ L_{\vec{\mathbf{b}}}(t;(x(\cdot),u(\cdot)))(T,x) + \mathbf{R}^n_+$$

and

$$L^\Phi_\Psi(t;(x(\cdot),u(\cdot)))(T,x) \ =\ L^{\vec{\mathbf{c}}}_{\vec{\mathbf{b}}}(t;(x(\cdot),u(\cdot)))(T,x) + \mathbf{R}^n_+.$$

Finally, we note that

$$\left\{\begin{array}{l} V_\Psi(\Phi)(T,x) \\ \ =\ \bigcup_{(x(\cdot),u(\cdot))\in\mathcal{C}(x)} \bigcup_{t\in[0,T]} \left[L^{\vec{\mathbf{c}}}_{\vec{\mathbf{b}}}(t;(x(\cdot),u(\cdot)))(T,x) + \mathbf{R}^n_+\right] \\ \ =\ V_{\vec{\mathbf{b}}}(\vec{\mathbf{c}})(T,x) + \mathbf{R}^n_+. \qquad \square \end{array}\right.$$

Recall that, for a closed subset $A \subset \mathbf{R}^n$ satisfying $A = A + \mathbf{R}^n_+$, the interior of A is equal to

$$\mathrm{Int}(A) \ =\ A + \overset{\circ}{\mathbf{R}}{}^n_+,$$

and thus the boundary of A is equal to the set of (weak) Pareto optima of A: Indeed, $y \in \partial A$ if and only if, for any $z \in A$, there exists at least $i \in \{1,\dots,n\}$ such that $y_i \le z_i$. We say that $z >> y$ if, for any $i \in \{1,\dots,n\}$, $z_i > y_i$.

Hence we deduce the following consequence of Theorem 5.3.

THEOREM 7.2. *Let us consider $y_T >> \vec{\mathbf{b}}(T,x))$ to be a Pareto minimum of the set $J_{V_{\vec{\mathbf{b}}}(\vec{\mathbf{c}})}(T,x)$. Consider any solution $(x(\cdot),u(\cdot)) \in \mathcal{C}(x)$ starting from $x \in \mathrm{Dom}(V_{\mathbf{b}}(\mathbf{c}))$ satisfying*

$$y_T \ \ge\ J_{V_{\vec{\mathbf{b}}}(\vec{\mathbf{c}})}(t,x;(x(\cdot),u(\cdot))) \tag{37}$$

until the first time t^* when, for at least one component $i = 1,\dots,n$,

$$y_{T_i} \ \le\ K_{\vec{\mathbf{b}}_i}(t^*,x;(x(\cdot),u(\cdot))).$$

Then y_T actually remains a Pareto minimum of the sets $J_{V_{\vec{\mathbf{b}}}(\vec{\mathbf{c}})}(t;(x(\cdot),u(\cdot)))(T,x)$ whenever $t \in [0,t^*]$.

REFERENCES

[1] J.-P. Aubin, *Contingent derivatives of set-valued maps and existence of solutions to nonlinear inclusions and differential inclusions*, in Mathematical Analysis and Applications, Part A, Adv. in Math. Suppl. Stud. 7a, L. Nachbin, ed., Academic Press, New York, 1981, pp. 159–229.

[2] J.-P. Aubin, *Viability Theory*, Birkhäuser Boston, Boston, 1991.

[3] J.-P. Aubin, *Dynamic Economic Theory: A Viability Approach*, Springer-Verlag, Berlin, 1997.

[4] J.-P. Aubin, *Mutational and Morphological Analysis: Tools for Shape Regulation and Morphogenesis*, Birkhäuser Boston, Boston, 1999.

[5] J.-P. Aubin, *Impulse Differential Inclusions and Hybrid Systems: A Viability Approach*, Lecture notes, Université Paris-Dauphine, Paris, France, 2001.

[6] J.-P. Aubin, *Optimal impulse control problems and quasi-variational inequalities thirty years later: A viability approach*, in Contrôle optimal et EDP: Innovations et Applications, IOS Press, Amsterdam, 2000, pp. 311–324.

[7] J.-P. Aubin, *Boundary-value problems for systems of first-order partial differential inclusions*, NoDEA Nonlinear Differential Equations Appl., 7 (2000), pp. 67–90.

[8] J.-P. Aubin, *Lyapunov functions for impulse and hybrid control systems*, in Proceedings of the IEEE Conference on Decision and Control, Sydney, Australia, 2000.

[9] J.-P. Aubin, *The caliber of impulse and hybrid control systems*, in Proceedings of the CIB International Symposium on Evolution, Trento, 2000, Birkhäuser, Berlin, 2002.

[10] J.-P. Aubin, *The substratum of impulse and hybrid control systems*, in Hybrid Systems: Computation and Control, Proceedings of the HSCC 2001 Conference, Lecture Notes in Comput. Sci. 2034, Springer-Verlag, New York, 2001, pp. 105–118.

[11] J.-P. Aubin, *Dynamic core of fuzzy dynamical cooperative games*, in Annals of Dynamic Games, Ninth International Symposium on Dynamical Games and Applications, Adelaide, Australia, 2002.

[12] J.-P. Aubin, *Viability kernels and capture basins of sets under differential inclusions*, SIAM J. Control Optim., 40 (2001), pp. 853–881.

[13] J.-P. Aubin, A. Bicchi, and S. Pancanti, *Detectability of Evolutions by Tubes*, in preparation.

[14] J.-P. Aubin, N. Bonneuil, and F. Maurin, *Non-linear structured population dynamics with co-variables*, Math. Population Stud., 9 (2000), pp. 1–31.

[15] J.P. Aubin and F. Catté, *Fixed-point and algebraic properties of viability kernels and capture basins of sets*, Set-Valued Anal., to appear.

[16] J.-P. Aubin and G. Da Prato, *Solutions contingentes de l'équation de la variété centrale*, C. R. Acad. Sci. Paris Sér. I Math., 311 (1990), pp. 295–300.

[17] J.-P. Aubin and G. Da Prato, *Contingent solutions to the center manifold equation*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 9 (1992), pp. 13–28.

[18] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser Boston, Boston, 1990.

[19] J.-P. Aubin and H. Frankowska, *Inclusions aux dérivées partielles gouvernant des contrôles de rétroaction*, C. R. Acad. Sci. Paris Sér. I Math., 311 (1990), pp. 851–856.

[20] J.-P. Aubin and H. Frankowska, *Systèmes hyperboliques d'inclusions aux dérivées partielles*, C. R. Acad. Sci. Paris Sér. I Math., 312 (1991), pp. 271–276.

[21] J.-P. Aubin and H. Frankowska, *Hyperbolic systems of partial differential inclusions*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 18 (1992), pp. 541–562.

[22] J.-P. Aubin and H. Frankowska, *Partial differential inclusions governing feedback controls*, J. Convex Anal., 2 (1995), pp. 19–40.

[23] J.-P. Aubin and H. Frankowska, *The viability kernel algorithm for computing value functions of infinite horizon optimal control problems*, J. Math. Anal. Appl., 201 (1996), pp. 555–576.

[24] J.-P. Aubin and H. Frankowska, *Set-valued solutions to the Cauchy problem for hyperbolic systems of partial differential inclusions*, NoDEA Nonlinear Differential Equations Appl., 4 (1997), pp. 149–168.

[25] J.-P. Aubin and G. Haddad, *Cadenced runs of impulse and hybrid control systems*, Internat. J. Robust Nonlinear Control, 11 (2001), pp. 401–415.

[26] J.-P. Aubin and G. Haddad, *Path-dependent impulse and hybrid systems*, in Hybrid Systems: Computation and Control, Proceedings of the HSCC 2001 Conference, Lecture Notes in Comput. Sci. 2034, Springer-Verlag, New York, 2001, pp. 119–132.

[27] J.-P. Aubin and G. Haddad, *Detectability under impulse differential inclusions*, in Proceedings of the European Control Conference, Porto, Portugal, 2001.

[28] J.-P. Aubin and G. Haddad, *Detectability through measurements under impulse differential inclusions*, Math. Control Signals Systems, submitted.

[29] J.-P. Aubin, J. Lygeros, M. Quincampoix, S. Sastry, and N. Seube, *Impulse differential*

inclusions: A viability approach to hybrid systems, IEEE Trans. Automat. Control, (2001).

[30] J.-P. Aubin, D. Pujal, and P. Saint-Pierre, *Dynamic Management of Portfolios with Transaction Costs under Contingent Uncertainty*, preprint, Université Paris-Dauphine, Paris, France, 2001.

[31] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser Boston, Boston, 1997.

[32] E.N. Barron and R. Jensen, *Semicontinuous viscosity solutions for Hamilton-Jacobi equations with convex Hamiltonians*, Comm. Partial Differential Equations, 15 (1990), pp. 1713–1742.

[33] E.N. Barron and R. Jensen, *Optimal control and semicontiuous viscosity solutions*, Proc. Amer. Math. Soc., 113 (1991), pp. 393–402.

[34] A. Bensoussan and Menaldi, *Hybrid control and dynamic programming*, Dynam. Contin. Discrete Impuls. Systems, 3 (1997), pp. 395–442.

[35] P. Bernhard, *Expected values, feared values and partial information optimal control*, in New Trends in Dynamic Games and Applications, Ann. Internat. Soc. Dynam. Games 3, Birkhäuser Boston, Boston, 1995, pp. 3–24.

[36] P. Bernhard, *On the performance index of feared value control*, in Proceedings of the ISDG Symposium, Sils-Maria, Switzerland, 1997.

[37] P. Bernhard, *Max-plus algebra and mathematical fear in dynamic optimization. Set-valued analysis in control theory*, Set-Valued Anal., 8 (2000), pp. 71–84.

[38] P. Cardaliaguet, M. Quincampoix, and P. Saint-Pierre, *Set-valued numerical methods for optimal control and differential games*, in Stochastic and Differential Games. Theory and Numerical Methods, Ann. Internat. Soc. Dynam. Games 4, Birkhäuser Boston, Boston, 1999, pp. 177–247.

[39] O. Carja and C. Ursescu, *The characteristics method for a first-order partial differential equation*, An. Ştiinţ. Univ. Al. I. Cuza Iaşi Secţ. I a Mat., 39 (1993), pp. 367–396.

[40] O. Carja and C. Ursescu, *Viscosity solutions and partial differential inequalities*, in Evolution Equations, Control Theory and Biomathematics, Dekker, New York, 1994, pp. 39–44.

[41] H. Frankowska, *L'équation d'Hamilton-Jacobi contingente*, C. R. Acad. Sci. Paris Sér. I Math., 304 (1987), pp. 295–298.

[42] H. Frankowska, *Optimal trajectories associated to a solution of contingent Hamilton-Jacobi equations*, in Proceedings of the 26th IEEE Conference on Decision and Control, Los Angeles, CA, 1987.

[43] H. Frankowska, *Optimal trajectories associated to a solution of contingent Hamilton-Jacobi equations*, Appl. Math. Optim., 19 (1989), pp. 291–311.

[44] H. Frankowska, *Hamilton-Jacobi equation: Viscosity solutions and generalized gradients*, J. Math. Anal. Appl., 141 (1989), pp. 21–26.

[45] H. Frankowska, *Lower semicontinuous solutions to Hamilton-Jacobi-Bellman equations*, in Proceedings of the 30th IEEE Conference on Decision and Control, Brighton, UK, 1991.

[46] H. Frankowska, *Lower semicontinuous solutions of Hamilton–Jacobi–Bellman equations*, SIAM J. Control Optim., 31 (1993), pp. 257–272.

[47] H. Frankowska, *Control of Nonlinear Systems and Differential Inclusions*, Birkhäuser Boston, Boston, to appear.

[48] H. Frankowska, S. Plaskacz, and T. Rzezuchowski, *Measurable viability theorems and the Hamilton-Jacobi-Bellman equation*, J. Differential Equations, 116 (1995), pp. 265–305.

[49] H. Frankowska, S. Plaskacz, and T. Rzezuchowski, *Théorèmes de Viabilité Mesurable et l'équation d'Hamilton-Jacobi-Bellman*, C. R. Acad. Sci. Paris Sér. I Math., (1995), pp. 131–134.

[50] G. Haddad, *Monotone trajectories of differential inclusions with memory*, Israel J. Math., 39 (1981), pp. 83–100.

[51] G. Haddad, *Monotone viable trajectories for functional differential inclusions*, J. Differential Equations, 42 (1981), pp. 1–24.

[52] G. Haddad, *Topological properties of the set of solutions for functional-differential inclusions*, Nonlinear Anal., 5 (1981), pp. 1349–1366.

[53] A.S. Matveev and A.V. Savkin, *Qualitative Theory of Hybrid Dynamical Systems*, Birkhäuser Boston, Boston, 2000.

[54] A.S. Matveev and A.V. Savkin, *Hybrid Dynamical Systems: Controller and Sensor Switching Problems*, Birkhäuser Boston, Boston, 2001.

[55] D. Pujal, *Valuation et gestion dynamiques de portefeuilles*, Thèse de l'Université de Paris-Dauphine, Paris, France, 2000.

[56] D. Pujal and P. Saint-Pierre, *L'algorithme du bassin de capture appliqué pour évaluer des options européennes, américaines ou exotiques*, preprint, Université Paris-Dauphine, Paris,

France, 2001.
[57] M. Quincampoix, *Frontières de domaines d'invariance et de viabilité pour des inclusions différentielles avec contraintes*, C. R. Acad. Sci. Paris Sér. I Math., 311 (1990), pp. 411–416.
[58] M. Quincampoix, *Enveloppes d'invariance pour des inclusions différentielles Lipschitziennes: Applications aux problèmes de cibles*, C. R. Acad. Sci. Paris Sér. I Math., 314 (1992), pp. 343–347.
[59] M. Quincampoix and V. Veliov, *Viability with a target: Theory and applications*, in Applications of Mathematics in Engineering, Heron Press, Sofia, Bulgaria, 1998, pp. 47–54.
[60] R.T. Rockafellar and R. Wets, *Variational Analysis*, Springer-Verlag, New York, 1997.
[61] P. Saint-Pierre, *Approximation of the viability kernel*, Appl. Math. Optim., 29 (1994), pp. 187–209.
[62] P. Saint-Pierre, *Approximation of capture basins for hybrid systems*, in Hybrid Systems: Computation and Control, Proceedings of the HSCC 2002 Conference, Lecture Notes in Comput. Sci. 2034, Springer-Verlag, New York, 2002.
[63] A. van der Shaft and H. Schumacher, *An Introduction to Hybrid Dynamical Systems*, Lecture Notes in Control and Inform. Sci. 251, Springer-Verlag, New York, 1999.
[64] S. Shi, *Théorèmes de viabilité pour les inclusions aux dérivées partielles*, C. R. Acad. Sci. Paris Sér. I Math., 303 (1986), pp. 11–14.
[65] S. Shi, *Nagumo type condition for partial differential inclusions*, Nonlinear Anal., 12 (1988), pp. 951–967.
[66] S. Shi, *Optimal control of strongly monotone variational inequalities*, SIAM J. Control Optim., 26 (1988), pp. 274–290.
[67] S. Shi, *Viability theorems for a class of differential-operator inclusions*, J. Differential Equations, 79 (1989), pp. 232–257.
[68] L.E.O. Svensson and M. Woodford, *Indicator Variables for Optimal Policy*, NBRR WP 7953, 2000.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 2, pp. 457–486

NUMERICAL APPROXIMATIONS FOR STOCHASTIC DIFFERENTIAL GAMES*

HAROLD J. KUSHNER†

Abstract. The Markov chain approximation method is a widely used, robust, relatively easy to use, and efficient family of methods for the bulk of stochastic control problems in continuous time for reflected-jump-diffusion-type models. It has been shown to converge under broad conditions, and there are good algorithms for solving the numerical problems if the dimension is not too high. Versions of these methods have been used in applications to various two-player differential and stochastic dynamic games for a long time, and proofs of convergence are available for some cases, mainly using PDE-type techniques. In this paper, purely probabilistic proofs of convergence are given for a broad class of such problems, where the controls for the two players are separated in the dynamics and cost function, and which cover a substantial class not dealt with in previous works. Discounted and stopping time cost functions are considered. Finite horizon problems and problems where the process is stopped on first hitting an a priori given boundary can be dealt with by adapting the methods of [H. J. Kushner and P. Dupuis, *Numerical Methods for Stochastic Control Problems, in Continuous Time,* 2nd ed., Springer-Verlag, Berlin, New York, 2001] as done in this paper for the treated problems. The essential conditions are the weak-sense existence and uniqueness of solutions, an "almost everywhere" continuity condition, and that a weak local consistency condition holds "almost everywhere" for the numerical approximations, just as for the control problem. There are extensions to problems with controlled variance and jumps.

Key words. stochastic differential games, numerical methods, Markov chain approximations

AMS subject classifications. 60F17, 65C30, 65C40, 91A15, 91A23, 93E25

PII. S0363012901389457

1. Introduction.

The Markov chain approximation method of [25, 26, 32] is an effective and widely used method for the numerical solution of virtually all of the standard forms of stochastic control problems with reflected-jump-diffusion models. It is robust and can be shown to converge under very broad conditions. In this paper, the basic ideas will be extended to two-player stochastic dynamic games with the same systems model, but where the controls for the two players are separated in the dynamics and cost functions, and for certain classes of stopping time problems. Such "separated" models occur, for example, in pursuit-evasion games, where each player controls its own dynamics, risk-sensitive and robust control [2, 3, 8, 18, 36], Lagrangian formulation of optimization under side constraints, and controlled large deviation problems [12]. When the robust control is for controlled queues in heavy traffic, with or without finite buffers, or for its fluid limits, then the state is confined to some convex polyhedron by boundary reflection [28]. See section 8 for a few illustrations. The minimizing and maximizing players will be called, respectively, players 1 and 2.

Early results concerning algorithms and convergence for stochastic games for finite-state Markov chain models are in [30, 31], and a survey is in [37]. The performance of all of these algorithms can be improved with the use of multigrid, Gauss–Seidel, and various accelerated versions. See [32] for additional references and more

*Received by the editors May 17, 2001; accepted for publication (in revised form) November 30, 2001; published electronically June 18, 2002. This work was partially supported by contract DAAD19-99-1-0223 from the Army Research Office and National Science Foundation grant ECS 0097447.

http://www.siam.org/journals/sicon/41-2/38945.html

†Applied Mathematics Department, Lefschetz Center for Dynamical Systems, Brown University, Providence, RI 02912 (hjk@dam.brown.edu).

detail concerning such accelerated algorithms.

Partial results for the convergence problem for approximations of various forms of continuous state and time dynamic games have appeared, but there does not seem to be a complete development for the fully stochastic problem for reflected-jump-diffusion models. The upper value for a deterministic game (an ODE model) was treated by the Markov chain approximation method in [34, 35]. Results for various deterministic problems are in [4, 5, 6, 7, 40, 41]. The actual numerical methods which are used in the computations tend to be of the Markov chain approximation type, although the proofs are sometimes based on subsequent PDE techniques.

In this paper, we will use purely probabilistic methods of proof. Such methods have the advantage of providing intuition concerning numerical approximations, they cover many of the problem formulations to date, and they converge under quite general conditions. The essential conditions are *weak-sense* existence and uniqueness of the solution to the controlled equations, "almost everywhere" continuity of the dynamical and cost rate terms, and a natural "local consistency" condition: The local consistency and continuity need hold only almost everywhere with respect to the measure of the basic model; hence discontinuities and severe singularities in the dynamics and cost function can be treated under appropriate conditions (see, in particular, Theorems 4.7 and 7.1 and the treatment of discontinuities and complex variational problems with singularities in [32]). Furthermore, the numerical approximations are represented as processes which are close to the original, which gives additional intuitive and practical meaning to the method. Indeed, the Markov chain approximation method seems to provide the intuition for many of the actual numerical methods which are used, no matter what the method of proof of convergence.

We will treat only a selection of problems. The basic controlled process $x(\cdot)$ is defined by (2.2) or, equivalently, (2.4). We concentrate on discounted and stopping time cost functions. Others, such as finite horizon problems and problems where the process is stopped on first hitting an a priori given boundary, can be dealt with by adapting the methods of [32] as done in this paper for the treated problems.

In many applications, the state of the actual physical problem is confined to a bounded set, and the reflection term in (2.2) ensures the correct boundary behavior. One example is the heavy traffic limit of controlled queueing networks with finite buffers [1, 28] or robust control of such systems as in [2, 3], where the set is a hyperrectangle. Then robust control or the optimization under side constraints would lead to a game problem with a hyperrectangular state space. Another example would be the control of large deviations for such problems, along the lines of [12]. If the system state is not a priori confined to a bounded set, then, for numerical purposes, it is commonly necessary to bound the state space artificially and then experiment with the bounds. Such problems which are bounded for numerical purposes often involve reflecting boundaries. For this reason, our basic model is confined to a state space G that is a convex polyhedron, and it is confined by a "reflection" on the boundary. In [32], the boundary of the state space was determined by a set of smooth curved surfaces. We restrict our attention to the simpler polyhedral case, since that is the one most widely used, and it avoids details which distract from the general development. However, the approximations of the more general boundaries that were used in [32] can be carried over without change to the problem of this paper. Similarly, for simplicity, we drop the jump term (which is treated in [32]) since including it for the game involves no new issues. See also [27] for a setup where the jumps themselves are controlled. Again, for simplicity, we do not allow the variance to be controlled.

However, if (see (2.2)) $w(\cdot) = (w_1(\cdot), w_2(\cdot))$, where the $w_i(\cdot), i = 1, 2$, are mutually independent, and we have the separated form $\sigma(x, u)dw = \sigma_1(x, u_1)dw_1 + \sigma_2(x, u_2)dw_2$, then the methods in [32, Chapter 13] or [26] can be adapted.

The methods to be used are based on the theory of weak convergence [10, 14] as they are applied in [32]. For any process with values in a complete and separable metric space S, let $D(S; 0, \infty)$ denote the space of S-valued paths on the time interval $[0, \infty)$ which are right continuous and have left-hand limits, and with the Skorohod topology used. The path space for the state process $x(\cdot)$ is $D(G; 0, \infty)$, where $G \subset \mathbb{R}^r$, r-dimensional Euclidean space. The tightness criterion to be used implicitly is Theorem 2.7b of [24], which is restated as [32, Theorem 9.2.1].

The development involves various concepts from stochastic control and game theory, weak-sense solutions, the Skorohod problem, and numerical methods for stochastic control, not all of which will be familiar to many readers. Because of this, to make the material as accessible as possible as well as to minimize detail, the development has been structured to take advantage of the results in [32] whenever possible. The analysis for the game problem is more difficult than that for the pure control problem, since we must work with strategies and not simply controls, the strategies of the two players might be dependent, and they need to be approximated in various ways for purposes of the analysis.

Sections 2 and 3 give the basic systems model and describe the numerical method. They also contain necessary background material. The dynamical model is the reflected SDE (2.2) or (2.4), also called the *Skorohod problem* [11, 32, 28]. (See also the beginning of section 4.) The conditions on the boundary of the state space are A2.1–A2.2. Condition A2.1 covers the great majority of cases of current interest, including those that arise from queueing and communications networks, as noted in section 2. The condition is trivial to verify for the special case where the state space is a hyperrectangle, with reflection directions being the interior normals. As is common in control theory when limits of a sequence of controls are involved, much of the analysis uses the notion of relaxed control, and the necessary definitions are given. The definitions of the upper and lower value of the game requires a precise definition of the class of allowed strategies. These are given in section 2. Later, we will define various subclasses of these sets which are needed in the approximation and limit proofs. The bulk of the paper works with weak-sense solutions. This allows the greatest generality, including the possibility of using Girsanov transformation methods for constructing solutions, hence the possibility of discontinuous dynamics. However, it comes at a price since the notation is more complicated than what would be required if Lipschitz conditions (hence strong-sense solutions) were used.

The numerical method, which is the Markov chain approximation procedure, is discussed in section 3. The actual ways of approximating the original problem to get the approximating chain and associated cost function are the same as in [32] for the pure control problem since it is the process for arbitrary controls that is approximated. The basic and natural *local consistency conditions* are stated. The approximation to the original process $x(\cdot)$ is a continuous time interpolation of the chain, and this interpolation $\psi^h(\cdot)$ is defined, and a useful representation is given. The upper and lower values for the game for the chain are defined.

The actual proof of convergence of the numerical method in Theorem 7.1 is not long. However, it depends on many approximations and estimates as well as on the fact that the original game has a value. These issues are dealt with in sections 4–6. Under a Lipschitz condition, Theorem 4.3 shows that the costs are well approximated

(in a uniform sense) if the controls are, and Theorem 4.4 proves similar facts when there is only a weak-sense solution. Then it is shown that a fine discretization of any of the controls in space and time, and even slightly delaying the actions of any of the controls, changes the costs only slightly, again uniformly in the controls. Loosely speaking, the costs are continuous in the controls of either player and uniform in the control of the other. Theorem 4.7 shows, under appropriate conditions, how to get similar results when the dynamical and cost rate functions are discontinuous. These approximations are fundamental to the proof of the existence of the value of the game in section 5 since they imply that slight delays in any of the controls have little effect on the results, which in turn implies that "who goes first" is not too important.

Section 6 contains the final "auxiliary" result. In the proof of convergence of the numerical method, one needs to use ϵ-optimal strategies for the player who goes first. These strategies are for theoretical purposes only and *do not have any use in practice.* The construction of these strategies is much more complicated than what is required for the pure control problem, and it is done in Theorem 6.1.

The convergence of the numerical method is given in section 7. In the pure control problem of [32], the numerical approximations are controlled Markov chains, and one needs to show that the sequence of approximations to the optimal value function converges as the approximation parameter goes to zero. Here, the numerical approximations are games for Markov chains. They might or might not have a value, depending on the form of the approximation. However, one needs to show at least that the upper and lower values converge to the value of the original game as the approximation parameter goes to its limit. This is more difficult than the proof of convergence for the control problem, and one needs to keep careful track of the information available to the individual players.

Section 8 contains a brief discussion of some examples and extensions. The treatment of the ergodic cost case uses quite different methods and is in [29].

2. The system model.

Assumptions on the state space G. It is assumed that the system state $x(t)$ is confined to the set G by boundary reflections. Conditions A2.1 and A2.2 are common in treatments of SDEs with reflections and piecewise smooth boundaries [11, 32, 28].

A2.1. *G is a bounded convex polyhedron in r-dimensional Euclidean space $\mathbb{R}^r$ with an interior and a finite number of faces. Let d_i denote the direction of reflection to the interior of the ith face, assumed constant there. On any edge or corner, the reflection direction can be any nonnegative linear combination of the directions on the adjoining faces. Let $d(x)$ denote the set of reflection directions at $x \in \partial G$, the boundary of G. For an arbitrary corner or edge of ∂G, let $\bar{d}_i$ and $\bar{n}_i$ denote the direction of reflection and the interior normal, respectively, on the ith adjoining face. Then there are constants $a_i > 0$ (depending on the edge or corner) such that*

$$a_i \langle \bar{n}_i, \bar{d}_i \rangle > \sum_{j:j\neq i} a_j \left| \langle \bar{n}_i, \bar{d}_j \rangle \right| \quad \text{for all } i. \tag{2.1}$$

A2.2. *There is a neighborhood $N(\partial G)$ and an extension of $d(\cdot)$ to $\overline{N(\partial G)}$ such that the following holds: For each $\epsilon > 0$, there is $\mu > 0$ which goes to zero as $\epsilon \to 0$ and such that if $x \in \overline{N(\partial G)} - \partial G$ and distance$(x, \partial G) \leq \mu$, then $d(x)$ is in the convex hull of $\{d(v); v \in \partial G, \text{distance}(x, v) \leq \epsilon\}$.*

In A2.3, the real variables c_i are the coefficients in the cost rate term $c'dy$, $c = \{c_i\}$, in (2.5), and U_i is the space of values for the control $u_i(t)$ of player i in (2.2).

A2.3. *$U_i, i = 1, 2$, are compact subsets of some Euclidean space, and $c_i \geq 0$.*

A2.4. *The functions $k_i(\cdot)$ and $b_i(\cdot)$ are real-valued (resp., $\mathbb{R}^r$-valued) and continuous on $G \times U_i$. Let $\sigma(\cdot)$ be a Lipschitz continuous matrix-valued function on G, with r rows and with the number of columns being the dimension of the Wiener process in* (2.2). *The $b_i(\cdot, \alpha_i)$ are Lipschitz continuous, uniformly in α_i.*

Later, the continuity and Lipschitz conditions in A2.4 will be replaced by A2.5 and either A2.6 or A2.7, and then we will be concerned with weak-sense solutions.

Comments on A2.1 and A2.2. Condition A2.2 is unrestrictive since one can always construct the extension. That A2.1 is quite natural can be seen from the following comments. First, suppose that the state space is being bounded for purely numerical reasons. Then the reflections are introduced merely to give a compact set G, which should be large enough so that the effects on the solution in the region of main interest are small. Then one often uses a hyperrectangle with normal reflection directions, in which case the right side of (2.1) is zero. Next, consider a heavy traffic queueing network model [22, 28, 39] where the state space is the nonnegative orthant, and the probability that an output of the ith processor goes to the queue for the jth processor is q_{ij}. Define the routing matrix $Q = \{q_{ij}; i, j\}$. If the spectral radius of Q is less than unity, then all customers will eventually leave the system, with probability one. The model is a special case of (2.2), and we can write $z(t) = [I - Q']y(t)$, where $y_i(\cdot)$ is nondecreasing, continuous, and can increase only at t, where $x_i(t) = 0$. In this case, A2.1 implies (see [11, 28]) the so-called completely S condition [22, 28, 38], which is essential to ensure important properties of the representation (2.2); for example, that $z(\cdot)$ has bounded variation with probability one. Also, A2.1 implies the Lipschitz condition and bound in Theorem 4.2.

The system model. Let $w(\cdot)$ be a standard vector-valued Wiener process with respect to a filtration $\{\mathcal{F}_t, t < \infty\}$, which might depend on the control. Let $u_i(\cdot), i = 1, 2$, be U_i-valued, measurable, and $\mathcal{F}_t$-adapted processes. Such processes are to be called *admissible controls.*[1] Keep in mind that the mere fact that $u_i(\cdot), i = 1, 2$, are admissible does not imply that they are acceptable controls for the game since the two players will have different information available depending on who "goes first." Furthermore, controls for the game are defined in terms of "strategies," as discussed at the end of this section. Nevertheless, for any controls with the correct information dependencies, there will be a filtration with respect to which $w(\cdot)$ is a standard vector-valued Wiener process, and to which the controls are adapted. The concept of admissibility will be used in getting useful approximations and bounds.

The dynamical model for the game process is the reflected SDE

$$x(t) = x(0) + \sum_{i=1}^{2} \int_0^t b_i(x(s), u_i(s))ds + \int_0^t \sigma(x(s))dw(s) + z(t), \tag{2.2}$$

where $u_i(\cdot)$ is the control for player i, $i = 1, 2$. The process $z(\cdot)$ is due to the boundary reflections and ensures that $x(t) \in G$. It has the representation

$$z(t) = \sum_i d_i y_i(t), \tag{2.3}$$

where $y(0) = 0$ and the $y_i(\cdot)$ are continuous, nondecreasing, and can increase only at t, where $x(t)$ is on the ith face of ∂G. The condition (2.1) implies that the set of reflection directions on any set of intersecting boundary faces are linearly independent.

[1] They will sometimes be referred to as admissible ordinary controls to distinguish them from relaxed controls.

This implies that the representation (2.3) is unique. See [28, Chapter 3] or [11, 21, 32] for a discussion of equations such as (2.2).

Relaxed controls $r_i(\cdot)$. Suppose that, for some filtration $\{\mathcal{F}_t, t < \infty\}$ and some standard vector-valued $\mathcal{F}_t$-Wiener process $w(\cdot)$, each $r_i(\cdot), i = 1, 2$, is a measure on the Borel sets of $U_i \times [0, \infty)$ such that $r_i(U_i \times [0, t]) = t$ and $r_i(A \times [0, t])$ is $\mathcal{F}_t$-measurable for each Borel set $A \subset U_i$. Then $r_i(\cdot)$ is said to be an *admissible relaxed control* for player i, with respect to $w(\cdot)$. If the Wiener process and filtration have been given or are obvious or unimportant, we simply say that $r_i(\cdot)$ is an admissible relaxed control for player i [15, 32]. For Borel sets $A \subset U_i$, we will write $r_i(A \times [0, t]) = r_i(A, t)$.

For almost all (ω, t) and each Borel set $A \subset U_i$, one can define the derivative

$$r_{i,t}(A) = \lim_{\delta \to 0} \frac{r_i(t, A) - r_i(t - \delta, A)}{\delta}.$$

Without loss of generality, we can suppose that the limit exists for all (ω, t). Then, for all (ω, t), $r_{i,t}(\cdot)$ is a probability measure on the Borel sets of U_i, and, for any bounded Borel set B in $U_i \times [0, \infty)$,

$$r_i(B) = \int_0^\infty \int_{U_i} I_{\{(\alpha_i, t) \in B\}} r_{i,t}(d\alpha_i) dt.$$

An ordinary control $u_i(\cdot)$ can be represented in terms of the relaxed control $r_i(\cdot)$, defined by its derivative $r_{i,t}(A) = I_A(u_i(t))$, where $I_A(u_i)$ is unity if $u_i \in A$ and is zero otherwise. The weak topology [32] will be used on the space of admissible relaxed controls. Relaxed controls are commonly used in control theory to prove existence theorems since any sequence of relaxed controls has a convergent subsequence.

Define the relaxed control $r(\cdot) = (r_1(\cdot) \times r_2(\cdot))$, with derivative $r_t(\cdot) = r_{1,t}(\cdot) \times r_{2,t}(\cdot)$. The $r(\cdot)$ is a measure on the Borel sets of $(U_1 \times U_2) \times [0, \infty)$, with marginals $r_i(\cdot), i = 1, 2$. Sometimes we will just write $r(\cdot) = (r_1(\cdot), r_2(\cdot))$ without ambiguity. The pair $(w(\cdot), r(\cdot))$ is called an *admissible pair* if each of the $r_i(\cdot)$ is admissible with respect to $w(\cdot)$.

In relaxed control terminology, (2.2) is written as

$$x(t) = x(0) + \sum_{i=1}^{2} \int_0^t \int_{U_i} b_i(x(s), \alpha_i) r_{i,s}(d\alpha_i) ds + \int_0^t \sigma(x(s)) dw(s) + z(t). \tag{2.4}$$

The existence and uniqueness of solutions to (2.4) will be discussed in the next section. Until section 8, for $x(0) = x$ and $\beta > 0$, the cost function is

$$W(x, r_1, r_2) = E \int_0^\infty e^{-\beta t} \left[\sum_{i=1}^{2} \int_{U_i} k_i(x(s), \alpha_i) r_{i,t}(d\alpha_i) dt + c' dy(t) \right]. \tag{2.5}$$

Define $\alpha = (\alpha_1, \alpha_2)$, $u = (u_1, u_2)$, and $b(x, \alpha) = b_1(x, \alpha_1) + b_2(x, \alpha_2)$, and define $k(\cdot)$ analogously.

Weak-sense solution. Suppose that $(w(\cdot), r(\cdot))$ is admissible with respect to some filtration $\{\mathcal{F}_t, t < \infty\}$ on some probability space. If there is a probability space on which are defined a filtration $\{\tilde{\mathcal{F}}_t, t < \infty\}$ and an $\tilde{\mathcal{F}}_t$-adapted triple $(\tilde{x}(\cdot), \tilde{w}(\cdot), \tilde{r}(\cdot))$, where $(\tilde{w}(\cdot), \tilde{r}(\cdot))$ is admissible and has the same probability law as $(w(\cdot), r(\cdot))$, and the triple satisfies (2.4), then it is said that there is a *weak-sense* solution to (2.4) for $(w(\cdot), r(\cdot))$. (The associated reflection process $\tilde{z}(\cdot)$ is determined by $(\tilde{x}(\cdot), \tilde{w}(\cdot), \tilde{r}(\cdot))$.)

Unique weak-sense solution. Suppose that we are given two probability spaces (indexed by $i = 1, 2$) with filtrations $\{\mathcal{F}_t^i, t < \infty\}$ and on which are defined processes $(x^i(\cdot), w^i(\cdot), r^i(\cdot))$, where $w^i(\cdot)$ is a standard vector-valued $\mathcal{F}_t^i$-Wiener process, $(w^i(\cdot), r^i(\cdot))$ is an admissible pair, and $(x^i(\cdot), w^i(\cdot), r^i(\cdot))$ solves (2.4). If equality of the probability laws of $(w^i(\cdot), r^i(\cdot))$, $i = 1, 2$, implies equality of the probability laws of $(x^i(\cdot), w^i(\cdot), r^i(\cdot))$, $i = 1, 2$, then we say that there is a *unique weak-sense solution* to (2.4) for the admissible pair $(w^i(\cdot), r^i(\cdot))$.

When working with weak-sense solutions, condition A2.5 and either A2.6 or A2.7 will replace A2.4.

A2.5. *The functions $\sigma(\cdot), b_i(\cdot), k_i(\cdot), i = 1, 2$, are bounded and measurable. Equation* (2.4) *has a unique weak-sense solution for each admissible pair $(w(\cdot), r(\cdot))$ and each initial condition.*

A2.6. *The functions $\sigma(\cdot), b_i(\cdot)$, and $k_i(\cdot), i = 1, 2$, are continuous.*

In A2.7, let $(w(\cdot), r(\cdot))$ be an arbitrary admissible pair, and let $x(\cdot)$ be the corresponding solution. Condition A2.7 differs from A2.6 in that the dynamics can be discontinuous, provided that not much time is spent in a small neighborhood of the set of discontinuity. A "threshold" control example where A2.7 holds is where $\sigma(x)\sigma'(x)$ is uniformly positive definite in G, $b(x, \alpha) = \bar{b}(x, \alpha) + b_0(x)$, where $\bar{b}(\cdot) = \sum_i \bar{b}_i(\cdot)$ and $k_i(\cdot), \bar{b}_i(\cdot)$, and $\sigma(\cdot)$ are continuous, and $b_0(x)$ takes one of two values, depending on which side of a hyperplane x lies.

A2.7. *There is a Borel set $D_d \subset G$ such that $\sigma(\cdot), b_i(\cdot)$, and $k_i(\cdot), i = 1, 2$, are continuous when $x \notin D_d$, and, for each $\epsilon > 0$, there is $t_\epsilon > 0$, which goes to zero as $\epsilon \to 0$, and such that, for any real T,*

$$\limsup_{\epsilon \to 0} \sup_{x(0)} \sup_{admis.\ r(\cdot)} \sup_{t_\epsilon \le t \le T} P\{x(t) \in N_\epsilon(D_d)\} = 0,$$

where $N_\epsilon(D_d)$ is an ϵ-neighborhood of D_d.

Comment on the Girsanov transformation method for defining solutions. When there is not a uniform Lipschitz condition (i.e., A2.3 does not hold), a common and useful approach to modeling uses the Girsanov transformation method [9, 23, 28, 32]. Here one starts with either a unique strong- or weak-sense solution and then introduces the control by a change of measure. Under appropriate conditions, the transformation is used to "shift" the drift term so that it includes the desired control. This procedure does not change the filtration or the probability space, but it does change the Wiener process. The new solution will also be weak-sense unique. See the references for more detail.

Classes of controls and strategies. Definitions. Let $\{\mathcal{F}_t, t < \infty\}$ be a filtration, and let $w(\cdot)$ be a standard vector-valued $\mathcal{F}_t$-Wiener process. Let $\mathcal{U}_i$ denote the set of controls $u_i(\cdot)$ for player i that are admissible with respect to $w(\cdot)$. For $\Delta > 0$, let $\mathcal{U}_i(\Delta) \subset \mathcal{U}_i$ denote the subset of admissible controls $u_i(\cdot)$ that are constant on the intervals $[k\Delta, k\Delta + \Delta)$, $k = 0, 1, \dots$, and where $u_i(k\Delta)$ is $\mathcal{F}_{k\Delta}$-measurable. Let B be a Borel subset of U_1. Let $\mathcal{L}_1(\Delta)$ denote the set of such piecewise constant controls for player 1 that are represented by functions $Q_{1k}(B; \cdot), k = 0, 1, \dots$, of the conditional probability type

$$\begin{aligned} P\left\{u_1(k\Delta) \in B \middle| w(s), u_2(s), s < k\Delta; u_1(l\Delta), l < k\right\} \\ = Q_{1k}\left(B; w(s), u(s), s < k\Delta\right), \end{aligned} \tag{2.6}$$

where $Q_{1k}(B; \cdot)$ is a measurable function for each Borel set B. Controls determined by (2.6) can be called strategies, owing to their explicit dependence on the past actions of both players.

If a rule for player 1 is given by the form (2.6), then, in the arguments of the cost functions, it will sometimes be written as $u_1(u_2)$ to emphasize its dependence on $u_2(\cdot)$. Although there is also dependence on $w(\cdot)$, that dependence is suppressed in the notation. Define $\mathcal{L}_2(\Delta)$ and the associated rules $u_2(u_1)$ for player 2 analogously. The same terminology will be used for relaxed controls. Thus $r_i(\cdot) \in \mathcal{U}_i$ means that $r_i(\cdot)$ is admissible, $r_i(\cdot) \in \mathcal{U}_i(\Delta)$ means that $r_i(\cdot)$ is admissible, the derivative $r_{i,t}(\cdot)$ is constant on the intervals $[k\Delta, k\Delta + \Delta)$, and $r_{i,t}(\cdot)$ is $\mathcal{F}_{k\Delta}$-measurable. Thus the difference between $\mathcal{L}_i(\Delta)$ and $\mathcal{U}_i(\Delta)$ is that, in the former case, the control is determined by a conditional probability law such as (2.6). However, the uniqueness condition A2.5 implies that it is only the probability law of $(w(\cdot), u_1(\cdot), u_2(\cdot))$ (or, more generally, of $(w(\cdot), r_1(\cdot), r_2(\cdot))$) that determines the law of the solution and hence the value of the cost. Thus we can always suppose that if the control for, say, player 1 is determined by a form such as (2.6), then (in relaxed control terminology) the law for $(w(\cdot), r_2(\cdot))$ is determined recursively by a conditional probability law

$$P\left\{\{w(s), r_2(s), k\Delta \le s \le k\Delta + \Delta\} \in \cdot \middle| w(s), r_2(s), u_1(s), s < k\Delta\right\}.$$

Theorems 4.5–4.7 imply that the values defined by (2.7) and (2.8) would not change if admissible relaxed controls were used in lieu of admissible ordinary controls.

Upper and lower values. For initial condition $x(0) = x$, define the upper and lower values for the game as

$$V^+(x) = \lim_{\Delta \to 0} \inf_{u_1 \in \mathcal{L}_1(\Delta)} \sup_{u_2 \in \mathcal{U}_2} W(x, u_1(u_2), u_2), \tag{2.7}$$

$$V^-(x) = \lim_{\Delta \to 0} \sup_{u_2 \in \mathcal{L}_2(\Delta)} \inf_{u_1 \in \mathcal{U}_1} W(x, u_1, u_2(u_1)). \tag{2.8}$$

Discussion of (2.7), (2.8). Let us interpret (2.7). For fixed $\Delta > 0$, consider the right side of (2.7). For each k, at time $k\Delta$, player 1 uses a rule of the form (2.6) to decide on the constant action that it will take on $[k\Delta, k\Delta + \Delta)$. That is, it "goes first." Player 2 can decide on its action at $t \in [k\Delta, k\Delta + \Delta)$ at the actual time that it is to be applied. (Its choice for the discrete instants $k\Delta$ is irrelevant.) Thus player 2 "goes last." Player 2 selects its strategy simply to be admissible. This operation yields admissible $u(\cdot) = (u_1(\cdot), u_2(\cdot))$. Under the Lipschitz condition A2.4, there is clearly a unique solution to (2.4). Alternatively, under the weak-sense existence and uniqueness assumption A2.5, there is a probability space on which are defined $(\tilde{w}(\cdot), \tilde{u}(\cdot))$ (with the same distribution as $(w(\cdot), u(\cdot))$) and on which is defined a solution to (2.4). The distribution of the set (solution, Wiener process, control) does not depend on the probability space. Thus, either way, the $\sup_{u_2 \in \mathcal{U}_2}$ is well defined for each rule for player 1. As $\Delta \to 0$, the inf sup is monotonically decreasing since player 1 can make decisions more often. Similar monotonicity was discussed in [20]. The analogous comments hold for (2.8). In section 4, it will be seen that the infs and sups could be taken over the relaxed controls without changing the results. Under our conditions, Theorem 5.1 says that there is a saddle point in that

$$V^+(x) = V^-(x) = V(x) \quad \text{for all } x \in G. \tag{2.9}$$

The use of limits of discrete strategies to define the upper and lower values goes back to [16, 17, 20], where discrete time games were used to approximate continuous time games. The Elliott–Kalton definition [13] does not require discretization and admits the widest class of strategies. However, various approaches based on discretized strategies are shown to yield the same values as those given by the Elliott–Kalton definition (see, for example, [19]). The references [4, 5, 6] all use various discrete time approximations in defining value, similar to (2.7) and (2.8). The numerical approximations converge to the value given by the definition (2.7)–(2.9).

3. The numerical procedure: The Markov chain approximation method.

The Markov chain approximation. Since some facts concerning the Markov chain approximation method of [25, 26, 32] will be needed when dealing with the convergence of the numerical approximation, let us recall the basic numerical procedure for the control problem where there is only one player. Loosely speaking, the method consists of two steps. The first step is the determination of a finite-state controlled Markov chain that has a continuous time interpolation that is an "approximation" of the process $x(\cdot)$. The second step solves the optimization problem for the chain and a cost function that approximates the one used for $x(\cdot)$. Let h denote the approximation parameter. Under a natural "local consistency" condition, the minimal cost function $V^h(x)$ for the controlled approximating chain converges to the minimal cost function for the original problem. The optimal control for the original problem is also approximated. The method is a robust and effective method for solving optimal control problems for reflected-jump-diffusions under very general conditions. The approximating chain and local consistency conditions are the same for the game problems of this paper. There are many methods for getting suitable approximating chains, and the references contain a comprehensive discussion. An advantage of the approach is that the approximations "stay close" to the physical model and can be adjusted to exploit local features. Our main aim is the proof of convergence for the game problem, so only the essential details of the numerical approximations will be given, and the reader is referred to the references for more information.

To construct the approximation, start by defining S_h, a discretization of $\mathbb{R}^r$. This can be done in many ways. For example, S_h might be a regular grid with the distance between points in any coordinate direction being h. The precise requirements, as spelled out below, are quite general. It is only the points in G and their immediate neighbors that will be of interest. The next step is to define the approximating controlled Markov chain ξ_n^h and its state space, which will be a subset of S_h. The state space for the chain is usually divided into two parts. The first part is $G_h = G \cap S_h$, on which the chain approximates the diffusion part of (2.4). If the chain tries to leave G_h, then it is returned immediately, consistently with the local reflection direction. Thus define ∂G_h^+ to be the set of points not in G_h to which the chain might move in one step from some point in G_h. The set ∂G_h^+ is an approximation to the reflecting boundary. This two-step procedure on the boundary simplifies both coding and analysis. In particular, it allows us to introduce a reflection process $z^h(\cdot)$ that is analogous to $z(\cdot)$. This "approximating" reflection process is needed to get the correct form for the limits of the approximating chain and for the components of the cost function that are due to the boundary reflection.

Local consistency on G_h. First, we define local consistency at $x \in G_h$. Let $u_n^h = (u_{1,n}^h, u_{2,n}^h)$ denote the controls used at step n for the approximating chain ξ_n^h. Let $E_{x,n}^{h,\alpha}$ (resp., $\text{covar}_{x,n}^{h,\alpha}$) denote the expectation (resp., the covariance), given all of the data to step n, when $\xi_n^h = x, u_n^h = \alpha$. Then the chain satisfies the following

condition: There is a function $\Delta t^h(x,\alpha) > 0$ such that

$$\begin{gathered} E^{h,\alpha}_{x,n}\left[\xi^h_{n+1} - x\right] = b(x,\alpha)\Delta t^h(x,\alpha) + o(\Delta t^h(x,\alpha)), \\ \operatorname{covar}^{h,\alpha}_{x,n}\left[\xi^h_{n+1} - x\right] = a(x)\Delta t^h(x,\alpha) + o(\Delta t^h(x,\alpha)),\ a(x) = \sigma(x)\sigma'(x), \\ \lim_{h\to 0}\sup_{x,\alpha} \Delta t^h(x,\alpha) = 0, \\ \|\xi^h_{n+1} - \xi^n_h\| \le K_1 h \end{gathered} \tag{3.1}$$

for some real K_1. With the straightforward methods in [32], $\Delta t^h(\cdot)$ is obtained automatically as a byproduct of getting the transition probabilities, and it will be used as an interpolation interval. Thus, in G, the conditional mean first two moments of $\xi^h_{n+1} - \xi^h_n$ are very close to those of the "differences" of the $x(\cdot)$ of (2.4). The interpolation interval $\Delta t^h(x,\alpha)$ can always be selected so that it does not depend on the control α (or even on the state x), and this is often the choice since it simplifies both the coding and numerical computations.

Remark concerning discontinuous dynamical and cost terms. The consistency condition (3.1) need not hold at all points. For example, consider a case where A2.7 holds: Let $k(\cdot), \sigma(\cdot)$ be continuous, and let $b(\cdot)$ have the form $b(x,\alpha) = b_0(x) + \bar{b}(x,\alpha)$, where $\bar{b}(\cdot)$ is continuous but $b_0(\cdot)$ is discontinuous at $D_d \subset G$. If A2.7 holds for D_d, then we do not need local consistency there. A2.7 would hold if the "noise" $\sigma(x)dw$ "drives" the process away from the set D_d, no matter what the control. See [32, discussion in section 5.5 and Theorem 10.5.3, and also the discussion concerning discontinuous dynamics in section 10.2] for examples and more detail.

Local consistency on the reflecting boundary ∂G^+_h. From points in ∂G^+_h, the transitions of the chain are such that they move to G_h, with the conditional mean direction being a reflection direction at x. More precisely,

$$\lim_{h\to 0}\sup_{x\in\partial G^+_h} \text{distance}(x, G_h) = 0, \tag{3.2}$$

and there are $\theta_1 > 0$ and $\theta_2(h) \to 0$ as $h \to 0$ such that, for all $x \in \partial G^+_h$,

$$\begin{gathered} E^{h,\alpha}_{x,n}\left[\xi^h_{n+1} - x\right] \in \{a\gamma :\ \gamma \in d(x), \theta_2(h) \ge a \ge \theta_1 h\}, \\ \Delta t^h(x,\alpha) = 0 \ \text{ for } x \in \partial G^+_h. \end{gathered} \tag{3.3}$$

The last line of (3.3) says that the reflection from states on ∂G^+_h is instantaneous. Reference [32] has an extensive discussion of straightforward methods of obtaining useful approximations, which can also be used for the game problem.

A cost function. Define $\Delta t^h_n = \Delta t^h(\xi^h_n, u^h_n)$ and $t^h_n = \sum_{l=0}^{n-1} \Delta t^h_l$. When $\xi^h_n \in \partial G^+_h$, we can write (modulo an asymptotically negligible term) $\xi^h_{n+1} - \xi^h_n = \sum_i d_i \delta y^h_{i,n}$, where $\delta y^h_{i,n} \ge 0$ and represents the increments in the direction d_i. The $\delta y^h_{i,n} = 0$ for $\xi^h_n \notin \partial G^+_h$. See also the representation of $z^h(\cdot)$ above (3.9). One choice of discounted cost function for the approximating chain and initial condition $x = x(0)$ is

$$W^h(x,u^h) = E\sum_{n=0}^{\infty} e^{-\beta t^h_n}\left[k(\xi^h_n, u^h_n)\Delta t^h_n I_{\{\xi^h_n \in G_h\}} + c'\delta y^h_n\right]. \tag{3.4}$$

Admissible controls and the values. Let $p^h(x,y|u)$ denote the transition probability of the chain for $u = (u_1, u_2)$, $u_1 \in U_1, u_2 \in U_2$. We will define the strategies

for the game analogously to what was done in (2.6). If player i goes first, its strategy is defined by a conditional probability law of the type

$$P\left\{u_{i,n}^n \in \cdot \middle| \xi_l^h, l \le n;\, u_l^h, l < n\right\}.$$

The class of such rules is called $\mathcal{U}_i^h(1)$. If player i goes last, then its strategy is defined by a conditional probability law of the type

$$P\left\{u_{i,n}^n \in \cdot \middle| \xi_l^h, l \le n, u_l^h, l < n;\, u_{j,n}^h, j \ne i\right\}.$$

The class of such strategies is called $\mathcal{U}_i^h(2)$. Let $\{\delta\tilde{w}_n^h, n < \infty\}$ be mutually independent random variables and such that $\delta\tilde{w}_n^h$ is independent of the "past" $\{\xi_l^h, l \le n, u_l^h, l < n\}$. For technical reasons, in section 7, the conditioning data might be augmented by $\{\delta\tilde{w}_l^h, l \le n\}$, but the Markov property

$$P\left\{\xi_{n+1}^h = x \middle| \xi_l^h, u_l^h, l \le n\right\} = p^h\left(\xi_n^h, x \middle| u_n^h\right)$$

will always hold.

The same notation $\mathcal{U}_i^h(k)$ is used for the admissible relaxed controls. Define the upper values, respectively, as

$$V^{+,h}(x) = \inf_{u_1 \in \mathcal{U}_1^h(1)} \sup_{u_2 \in \mathcal{U}_2^h(2)} W^h(x, u_1, u_2), \tag{3.5}$$

$$V^{-,h}(x) = \sup_{u_2 \in \mathcal{U}_2^h(1)} \inf_{u_1 \in \mathcal{U}_1^h(2)} W^h(x, u_1, u_2). \tag{3.6}$$

In interpreting the cost function and the interpolations to be defined below, keep in mind that $\Delta t^h(x, \alpha) = 0$ for $x \in \partial G_h^+$. For $x \in G_h$, the dynamic programming equation for the upper value is ($\alpha = (\alpha_1, \alpha_2)$)

$$V^{+,h}(x) = \min_{\alpha_1 \in U_1}\{\max_{\alpha_2 \in U_2} E_x^\alpha[e^{-\beta\Delta t^h(x,\alpha)}V^{+,h}(\xi_1^h) + k(x,\alpha)\Delta t^h(x,\alpha)]\}, \tag{3.7}$$

and, for $x \in \partial G_h^+$, it is

$$V^{+,h}(x) = E_x\left[V^{+,h}(\xi_1^h) + c'\delta y_1^h\right]. \tag{3.8}$$

Here E_x^α denotes the expectation, given initial state x, with control pair α used, and E_x is the expectation, given initial state x (the reflection direction is not controlled). The equations are analogous for the lower value. Owing to the contraction implied by the discounting, there is a unique solution to (3.7) [35]. If desired, the transition probabilities could be constructed so that $\Delta t^h(\cdot)$ does not depend on α, and we have the separated form[2]

$$p^h(x, y|\alpha) = \bar{p}_1(x, y|\alpha_1) + \bar{p}_2(x, y|\alpha_2).$$

Such a form is useful for establishing the existence of a value for the game for the chain [31, 30], but it is not needed for the convergence of the numerical method.

Continuous time interpolation. The chain ξ_n^h is defined in discrete time, but $x(\cdot)$ is defined in continuous time. Only the chain is needed for the numerical

[2]For example, for the latter use, the splitting method of [32, subsection 5.3.2].

computations. However, for the proofs of convergence, the chain must be interpolated into a continuous time process which approximates $x(\cdot)$. The interpolation intervals are suggested by the $\Delta t^h(\cdot)$ in (3.1). We will use a Markovian interpolation, called $\psi^h(\cdot)$. Let $\{\Delta\tau_n^h, n < \infty\}$ be conditionally mutually independent and "exponential" random variables in that

$$P_{x,n}^{h,\alpha}\left\{\Delta\tau_n^h \geq t\right\} = e^{-t/\Delta t^h(x,\alpha)}.$$

Note that $\Delta\tau_n^h = 0$ if ξ_n^h is on the reflecting boundary ∂G_h^+. Define $\tau_0^h = 0$, and, for $n > 0$, set $\tau_n^h = \sum_{i=0}^{n-1} \Delta\tau_i^h$. The τ_n^h will be the jump times of $\psi^h(\cdot)$. Now define $\psi^h(\cdot)$ and the interpolated reflection processes by

$$\psi^h(t) = x(0) + \sum_{\tau_{i+1}^h \leq t} [\xi_{i+1}^h - \xi_i^h],$$

$$Z^h(t) = \sum_{\tau_{i+1}^h \leq t} [\xi_{i+1}^h - \xi_i^h] I_{\{\xi_i^h \in \partial G_h^+\}},$$

$$z^h(t) = \sum_{\tau_{i+1}^h \leq t} E_i^h[\xi_{i+1}^h - \xi_i^h] I_{\{\xi_i^h \in \partial G_h^+\}}.$$

Define the continuous time interpolations $u_i^h(\cdot)$ of the controls analogously. Let $r_i^h(\cdot)$ denote the relaxed control representation of $u_i^h(\cdot)$. The process $\psi^h(\cdot)$ is a continuous time Markov chain. When the state is x and control pair is α, the jump rate out of $x \in G_h$ is $1/\Delta t^h(x,\alpha)$. So the conditional mean interpolation interval is $\Delta t^h(x,\alpha)$; i.e., $E_{x,n}^{h,\alpha}[\tau_{n+1}^h - \tau_n^h] = \Delta t^h(x,\alpha)$.

Define $\tilde{z}^h(\cdot)$ by $Z^h(t) = z^h(t) + \tilde{z}^h(t)$. Note that this representation splits the effects of the reflection into two parts. The first is composed of the "conditional mean" parts $E_i^h[\xi_{i+1}^h - \xi_i^h] I_{\{\xi_i^h \in \partial G_h^+\}}$, and the second is composed of the perturbations about these conditional means [32, section 5.7.9]. The process $z^h(\cdot)$ is a reflection term of the classical type. Both components can change only at t, where $\psi^h(t)$ can leave G_h. Suppose that at some time t, $Z^h(t) - Z^h(t-) \neq 0$, with $\psi^h(t-) = x \in G_h$. Then by (3.3), $z^h(t) - z^h(t-)$ points in a direction in $d(N_h(x))$, where $N_h(x)$ is a neighborhood with radius that goes to zero as $h \to 0$. The process $\tilde{z}^h(\cdot)$ is the "error" due to the centering of the increments of the reflection term about their conditional means and has bounded (uniformly in x, h) second moments, and it converges to zero, as will be seen in Theorem 3.1. By A2.1, A2.2, and the local consistency condition (3.3), we can write (modulo an asymptotically negligible term)

$$z^h(t) = \sum_i d_i y_i^h(t),$$

where $y_i^h(0) = 0$ and $y_i^h(\cdot)$ is nondecreasing and can increase only when $\psi^h(t)$ is arbitrarily close (as $h \to 0$) to the ith face of ∂G.

The interpolated cost criterion. The cost criterion (3.4) can be written (modulo an asymptotically negligible error), where we use relaxed control terminology, $x(0) = x$, and $r_i^h(\cdot)$ is the relaxed control representation of $u_i^h(\cdot)$, as

$$W^h(x, r^h) = E\int_0^\infty e^{-\beta t}\left[\sum_{i=1}^2 \int_{U_i} k_i(\psi^h(s), \alpha_i) r_{i,t}^h(d\alpha_i)dt + c'dy^h(t)\right]. \tag{3.9}$$

In the numerical computations, the controls are ordinary and not relaxed, but it will be convenient to use the relaxed control terminology when taking limits. The proof of Theorem 7.1 implies that there is $\rho^h \to 0$ as $h \to 0$ such that

$$V^{+,h}(x) \leq V^{-,h}(x) + \rho^h. \tag{3.10}$$

This implies that either the upper or lower numerical game gives an approximation to the original game.

A representation for $\psi^h(\cdot)$. The process $\psi^h(\cdot)$ has a representation which makes it appear close to (2.4) and which is useful in the convergence proofs. Let $\xi_0^h = x$. If $a(\cdot)$ is not uniformly positive definite, then augment the probability space by adding a standard vector-valued Wiener process $\tilde{w}(\cdot)$, where, for each n, $\delta\tilde{w}_{n+1}^h = \tilde{w}(\tau_n^h + \cdot) - \tilde{w}(\tau_n^h)$ is independent of the "past" $\{\psi^h(s), u^h(s), \tilde{w}(s), s \leq \tau_n^h\}$. Then, by [32, sections 5.7.3 and 10.4.1], we can write

$$\begin{aligned} \psi^h(t) = x &+ \int_0^t b(\psi^h(s), u^h(s))ds \\ &+ \int_0^t \sigma(\psi^h(s))dw^h(s) + Z^h(s) + \epsilon^h(s), \end{aligned} \tag{3.11}$$

where $\psi^h(t) \in G$. The process $\epsilon^h(\cdot)$ is due to the $o(\cdot)$ terms in (3.1) and is asymptotically unimportant in that, for any T, $\lim_h \sup_{x,r^h} \sup_{s\leq T} E|\epsilon^h(s)|^2 = 0$. The process $w^h(\cdot)$ is a martingale with respect to the filtration induced by $(\psi^h(\cdot), u^h(\cdot), w^h(\cdot))$ and converges weakly to a standard (vector-valued) Wiener process. The $w^h(t)$ is obtained from $\{\psi^h(s), \tilde{w}(s), s \leq t\}$. All of the processes in (3.11) are constant on the intervals $[\tau_n^h, \tau_{n+1}^h)$.

Let $|z^h|(T)$ denote the variation of the process $z^h(\cdot)$ on the time interval $[0, T]$. Then we have the following theorem from [32].

THEOREM 3.1 (Theorem 11.1.3 and (5.7.5) [32]). *Assume* A2.1, A2.2, *and the local consistency conditions, and let $b(\cdot)$ and $\sigma(\cdot)$ be bounded and measurable. Then, for any $T < \infty$, there are $K_2 < \infty$ and δ_h, where $\delta_h \to 0$ as $h \to 0$, and which do not depend on the controls or initial condition, such that*

$$E\left|z^h\right|^2(T) \leq K_2, \tag{3.12}$$

$$E \sup_{s\leq T} \left|\tilde{z}^h(s)\right|^2 = \delta_h E\left|z^h\right|(T). \tag{3.13}$$

The inequalities hold for $y^h(\cdot)$ replacing $z^h(\cdot)$.

4. Auxiliary results: Bounds and approximations. This section is concerned with various estimates and approximations of the solution to (2.4) which are uniform in the control. The proofs of convergence of any numerical approximations involve approximations of the underlying process, especially when control is involved, and the results of this section will be used in section 6 to obtain nearly optimal strategies of a particular type that will play a fundamental role in the convergence proofs of the numerical algorithms. Furthermore, the approximations will be used in section 5 to show that the game has a value. This is critical in showing that the numerical approximations actually converge to the desired value. The approximations imply, among other things, that slight delays in the controls of any of the players affect the

costs only slightly. Delaying the control of the second player is equivalent to that player "going first" since its actual applied control at any time will depend on "old" information. This idea will be used in the next section to prove that the game has a value. The first part of the following theorem is [11, Theorem 2.2]. The inequality (4.3) is [32, Theorem 11.1.1].

DEFINITION 4.1 (the Skorohod problem). *Assume* A2.1 *and* A2.2*, and let the components of the $\mathbb{R}^r$-valued function $\psi(\cdot)$ be right continuous and have left-hand limits. Consider the equation $\bar{x}(t) = \psi(t) + \bar{z}(t)$, $x(t) \in G$. Then $\bar{x}(\cdot)$ is said to solve the Skorohod problem* [11, 32] *if the following holds. The components of $\bar{z}(\cdot)$ are right continuous with $\bar{z}(0) = 0$, and $\bar{z}(\cdot)$ is constant on the time intervals where $\bar{x}(t)$ is in the interior of G. The variation $|\bar{z}|(t)$ of $\bar{z}(\cdot)$ on each $[0,t]$ is finite. There is measurable $\gamma(\cdot)$ with values $\gamma(t) \in d(\bar{x}(t))$, the set of reflection directions at $\bar{x}(t)$, such that $\bar{z}(t) = \int_0^t \gamma(s) d|\bar{z}|(s)$. Thus $\bar{z}(\cdot)$ can change only when $\bar{x}(t)$ is on the boundary of G, and then its "increment" is in a reflection direction at $\bar{x}(t)$.*

THEOREM 4.2. *Assume* A2.1 *and* A2.2*. Let $\psi(\cdot) \in D(\mathbb{R}^r; 0, \infty)$, and consider the Skorohod problem $\bar{x}(t) = \psi(t) + \bar{z}(t)$, $x(t) \in G$. Then there is a unique solution $(\bar{x}(\cdot), \bar{z}(\cdot))$ in $D(\mathbb{R}^{2r}; 0, \infty)$. There is $K < \infty$ depending only on the $\{d_i\}$ such that*

$$|\bar{x}(t)| + |\bar{z}(t)| \le K \sup_{s \le t} |\psi(s)|, \tag{4.1}$$

and, for any $\psi^i(\cdot) \in D(\mathbb{R}^r; 0, \infty), i = 1, 2$, and corresponding solutions $(\bar{x}^i(\cdot), \bar{z}^i(\cdot))$,

$$|\bar{x}_1(t) - \bar{x}_2(t)| + |\bar{z}_1(t) - \bar{z}_2(t)| \le K \sup_{s \le t} |\psi_1(s) - \psi_2(s)|. \tag{4.2}$$

Consider (2.4)*, where $b(\cdot)$ and $\sigma(\cdot)$ are bounded and measurable, and use the representation* (2.3) *for the reflection process $z(\cdot)$. Then, for any $T < \infty$, there is a constant K_1 which does not depend on the initial condition or controls and such that*

$$\sup_{x \in G} E\,|y(1)|^2 \le K_1. \tag{4.3}$$

Approximations under the Lipschitz condition A2.4. Suppose that the Lipschitz and continuity condition A2.4 holds. Then the bound (4.1) and Lipschitz condition (4.2) ensure a unique strong-sense solution to the SDE (2.2) or (2.4) for any admissible controls. The proofs of the convergence of the numerical methods and of the existence of the value depend on our ability to approximate the controls. This is simplest under the Lipschitz condition A2.4, and we start with that case. Then the same approximations will be shown to hold if A2.5 and either A2.6 or A2.7 replace A2.4.

For each admissible relaxed control $r(\cdot)$, let $r^\epsilon(\cdot)$ be admissible relaxed controls with respect to the same filtration and Wiener process $w(\cdot)$ and that satisfy

$$\lim_{\epsilon \to 0} \sup_{r_i \in \mathcal{U}_i} E \sup_{t \le T} \left| \int_0^t \int_{U_i} \phi_i(\alpha_i) \left[r_{i,s}(d\alpha_i) - r^\epsilon_{i,s}(d\alpha_i) \right] ds \right| = 0, \quad i = 1, 2, \tag{4.4}$$

for each bounded and continuous real-valued nonrandom function $\phi_i(\cdot)$ and each $T < \infty$. For future use, note that if (4.4) holds, then it also holds for functions $\phi_i(\cdot)$ of (t, α_i) that are continuous except when t takes some value in a finite set $\{t_i\}$. Let $x(\cdot)$ and $x^\epsilon(\cdot)$ denote the solutions to (2.4) corresponding to $r(\cdot)$ and $r^\epsilon(\cdot)$, respectively, with the same Wiener process used. In particular,

$$x^\epsilon(t) = x(0) + \int_0^t \int_{U_1 \times U_2} b(x^\epsilon(s), \alpha) r^\epsilon_s(d\alpha) ds + \int_0^t \sigma(x^\epsilon(s)) dw(s) + z^\epsilon(t). \tag{4.5}$$

Define

$$\rho^\epsilon(t) = \int_0^t \int_{U_1\times U_2} b(x(s),\alpha)\left[r_s(d\alpha) - r_s^\epsilon(d\alpha)\right] ds.$$

The processes $x(\cdot), x^\epsilon(\cdot)$, and $\rho^\epsilon(\cdot)$ depend on $r(\cdot)$, but this dependence is suppressed in the notation. The next theorem shows that the set $\{x(\cdot)\}$ over all admissible controls is equicontinuous in probability in the sense that (4.6) holds, and that the costs corresponding to $r(\cdot)$ and $r^\epsilon(\cdot)$ are arbitrarily close for small ϵ, uniformly in $r(\cdot)$.

THEOREM 4.3. *Assume* A2.1 *and* A2.2*, and let $b(\cdot), \sigma(\cdot)$ be bounded and measurable. Then, for each real $\lambda > 0$,*

$$\lim_{\Delta\to 0} \sup_{x(0)} \sup_{t} \sup_{r_1\in\mathcal{U}_1} \sup_{r_2\in\mathcal{U}_2} P\left\{\sup_{s\le\Delta} |x(t+s) - x(t)| \ge \lambda\right\} = 0. \tag{4.6}$$

Now add the assumptions A2.3 *and* A2.4*, and let $(r(\cdot), r^\epsilon(\cdot))$ satisfy* (4.4) *for each bounded and continuous $\phi_i(\cdot), i = 1, 2$, and $T < \infty$. Define $\Delta^\epsilon(t) = \sup_{s\le t} |x(s) - x^\epsilon(s)|^2$. Then, for each t,*

$$\lim_{\epsilon\to 0} \sup_{x(0)} \sup_{r_1\in\mathcal{U}_1} \sup_{r_2\in\mathcal{U}_2} E\left|\sup_{s\le t} \rho^\epsilon(s)\right|^2 = 0, \tag{4.7}$$

$$\lim_{\epsilon\to 0} \sup_{x(0)} \sup_{r_1\in\mathcal{U}_1} \sup_{r_2\in\mathcal{U}_2} \left[E\Delta^\epsilon(t) + E\sup_{s\le t} |z(s) - z^\epsilon(s)|^2\right] = 0, \tag{4.8}$$

$$\lim_{\epsilon\to 0} \sup_{x} \sup_{r_1\in\mathcal{U}_1} \sup_{r_2\in\mathcal{U}_2} |W(x,r) - W(x,r^\epsilon)| = 0. \tag{4.9}$$

Proof. Assume the conditions in the first sentence of the theorem. Define $\psi(\cdot)$ by

$$\psi(t) = \int_0^t \int_{U_1\times U_2} b(x(s),\alpha) r_s(d\alpha) ds + \int_0^t \sigma(x(s)) dw(s).$$

Then

$$x(t+\delta) - x(t) = [\psi(t+\delta) - \psi(t)] + [z(t+\delta) - z(t)].$$

By Theorem 4.2, there is $K < \infty$ which does not depend on the control or initial condition and such that

$$\sup_{s\le\delta} |x(t+s) - x(t)| + \sup_{s\le\delta} |z(t+s) - z(t)| \le K \sup_{s\le\delta} [\psi(t+s) - \psi(t)].$$

Now using standard estimates for SDEs to evaluate the fourth moments of the right side of the last inequality yields, for some $K_1 < \infty$,

$$\sup_{x(0),t} \sup_{r_1\in\mathcal{U}_1} \sup_{r_2\in\mathcal{U}_2} E\sup_{s\le\delta} |x(t+s) - x(t)|^4 \le K_1\delta^2, \tag{4.10}$$

which implies Kolmogorov's criterion for equicontinuity in probability, which is (4.6) [33, Proposition III.5.3]. Write

$$\begin{aligned} x(t) - x^\epsilon(t) \;=\; & \int_0^t \int_{U_1\times U_2} [b(x(s),\alpha) - b(x^\epsilon(s),\alpha)]\, r_s^\epsilon(d\alpha) ds + \rho^\epsilon(t) \\ & + \int_0^t [\sigma(x(s)) - \sigma(x^\epsilon(s))]\, dw(s) + z(t) - z^\epsilon(t). \end{aligned}$$

Now assume the Lipschitz condition A2.4. Then the Lipshitz condition (4.2) together with standard estimates for SDEs, imply that there is a constant K not depending on $(r(\cdot), r^\epsilon(\cdot))$ or on the initial condition $x(0)$ and such that

(4.11)
$$E\Delta^\epsilon(t) \le K\left[E\sup_{s\le t}|\rho^\epsilon(s)|^2 + (t+1)\int_0^t E\Delta^\epsilon(s)ds + E\sup_{s\le t}|z(s)-z^\epsilon(s)|^2\right],$$
$$E\sup_{s\le t}|z(s)-z^\epsilon(s)|^2 \le K\left[E\sup_{s\le t}|\rho^\epsilon(s)|^2 + (t+1)\int_0^t E\Delta^\epsilon(s)ds\right].$$

Suppose that, in the definition of $\rho^\epsilon(\cdot)$, the function $b(x(t),\alpha)$ was replaced by a bounded nonrandom function $\phi(t,\alpha)$, which is continuous except when t takes values in some finite set $\{t_i\}$. Then (4.7) and (4.8) would hold by (4.4) and the use of Gronwall's inequality on the first line of (4.11), after the second line is substituted in to eliminate $z(\cdot)-z^\epsilon(\cdot)$. The equicontinuity in probability (4.6) and the boundedness and continuity of $b(\cdot)$ imply that $b(x(t),\alpha)$ can be approximated arbitrarily well by replacing $x(t)$ by $x(k\mu)$ for $t\in[k\mu,k\mu+\mu), k=0,1,\dots$, where μ can be chosen independently of $r(\cdot)$. Doing this approximation and using (4.4) imply (4.7) and (4.8).

Now we turn our attention to (4.9). By (4.7), (4.8), and the discounting, the parts of $W(x,r^\epsilon)$ that involve $k(\cdot)$ converge to the corresponding parts of $W(x,r)$. As noted below (2.3), the linear independence of the reflection directions on any set of intersecting boundary faces which is implied by (2.1) implies that $z(\cdot)$ uniquely determines $y(\cdot)$ with probability one. Thus $y^\epsilon(\cdot)$ converges to $y(\cdot)$ with probability one. This convergence, the uniform integrability of the set $\{|y^\epsilon(t+1)-y^\epsilon(t)|;\ t<\infty,\ \text{all } r(\cdot),\epsilon>0\}$ (which is implied by (4.3) and the compactness of G), and the discounting imply that the component of $W(x,r^\epsilon)$ involving $y^\epsilon(\cdot)$ converges to the component of $W(x(0),r)$ involving $y(\cdot)$. □

Weak-sense solutions. The next theorem uses only weak-sense solutions and does not require the Lipschitz condition A2.4. Except for the uniformity assertion, it is a slight variation of [32, Theorem 10.1.2] or, equivalently, of [26, Theorem 3.5.2]. The method of proof, using specially selected probability spaces, is very useful in general when dealing with sequences of solutions that are defined in the weak sense.

THEOREM 4.4. *Assume* A2.1–A2.3, A2.5, *and* A2.6. *Let* $r(\cdot)$ *and* $r^\epsilon(\cdot)$, $\epsilon>0$, *be admissible with respect to some Wiener process* $w^r(\cdot)$ *and satisfy* (4.4). *For each* $\epsilon>0$, *there is a probability space with an admissible pair* $(\tilde w^{r,\epsilon}(\cdot),\tilde r^\epsilon(\cdot))$ *which has the same probability law as* $(w^r(\cdot),r^\epsilon(\cdot))$ *and on which is defined a solution* $(\tilde x^{r,\epsilon}(\cdot),\tilde y^{r,\epsilon}(\cdot))$ *to* (2.4). *Let* $x^r(\cdot)$ *denote the solution to* (2.4), *corresponding to* $(w^r(\cdot),r(\cdot))$, *and let* $z^r(\cdot)=\sum_i d_i y_i^r(\cdot)$ *denote the associated reflection process. Let* $F(\cdot)$ *be a bounded and continuous real-valued function on the path space of the canonical set* $(x(\cdot),y(\cdot),r(\cdot))$. *Then the approximation of the solutions by using* $r^\epsilon(\cdot)$ *is uniform in that*

$$\lim_{\epsilon\to0}\sup_{x(0)}\sup_{r_1\in\mathcal{U}_1}\sup_{r_2\in\mathcal{U}_2}\left|EF(\tilde x^{r,\epsilon}(\cdot),\tilde y^{r,\epsilon}(\cdot),\tilde r^\epsilon(\cdot)) - EF(x^r(\cdot),y^r(\cdot),r(\cdot))\right| = 0. \tag{4.12}$$

Now let $F(\cdot)$ *be only continuous with probability one with respect to the measure of any solution set* $(x(\cdot),y(\cdot),r(\cdot))$. *Then, if* $(x^n(\cdot),y^n(\cdot),r^n(\cdot))$ *converges weakly to* $(x(\cdot),y(\cdot),r(\cdot))$, $F(x^n(\cdot),y^n(\cdot),r^n(\cdot))$ *converges weakly to* $F(x(\cdot),y(\cdot),r(\cdot))$. *Also,* (4.12) *continues to hold.*

Proof. Let $F(\cdot)$ be bounded and continuous. Let $(w^r(\cdot),r(\cdot))$ be an admissible pair on some probability space, with associated solution process $x^r(\cdot)$ and re-

flection process $z^r(\cdot) = \sum_i d_i y_i^r(\cdot)$. Since $(w^r(\cdot), r^\epsilon(\cdot))$ is an admissible pair, by the existence part of A2.5, there is a probability space on which are defined processes $(\tilde{x}^{r,\epsilon}(\cdot), \tilde{y}^{r,\epsilon}(\cdot), \tilde{w}^{r,\epsilon}(\cdot), \tilde{r}^\epsilon(\cdot))$, where the last two members are an admissible pair with the same law as $(w^r(\cdot), r^\epsilon(\cdot))$, and $\tilde{x}^{r,\epsilon}(\cdot)$ is the associated weak-sense solution to (2.4), with reflection process $\tilde{z}^{r,\epsilon}(\cdot) = \sum_i d_i \tilde{y}^{r,\epsilon}(\cdot)$. Any weak-sense limit (as $\epsilon \to 0$) $(\tilde{x}^r(\cdot), \tilde{y}^r(\cdot), \tilde{w}^r(\cdot), \tilde{r}(\cdot))$ of the quadruple must solve (2.4), and $\tilde{w}^r(\cdot)$ is a standard vector-valued Wiener process with respect to the filtration generated by $(\tilde{x}^r(\cdot), \tilde{w}^r(\cdot), \tilde{r}(\cdot))$. For a proof of such a characterization of a related limit, see [32, Theorem 11.1.2].

By (4.4) and the weak convergence, $(\tilde{w}^r(\cdot), \tilde{r}(\cdot))$ has the probability law of $(w^r(\cdot), r(\cdot))$. Thus by the uniqueness part of A2.5, $(\tilde{x}^r(\cdot), \tilde{y}^r(\cdot), \tilde{w}^r(\cdot), \tilde{r}(\cdot))$ has the probability law of $(x^r(\cdot), y^r(\cdot), w^r(\cdot), r(\cdot))$. This yields convergence (as $\epsilon \to 0$) in (4.12) for each pair $(w^r(\cdot), r(\cdot))$ and initial condition $x(0)$.

Suppose that the uniformity (in $r(\cdot)$ and $x(0)$) of the convergence in (4.12) does not hold. Then there are $x(0), x_n \to x(0)$, all in G, $\rho > 0$, $\epsilon_n \to 0$, bounded and continuous $F(\cdot)$, and for each n there is a probability space on which are defined an admissible pair $(w^n(\cdot), r^n(\cdot))$, an associated solution $(x^n(\cdot), y^n(\cdot))$, and approximations $r_i^{n,\epsilon_n}(\cdot)$ to $r_i^n(\cdot), i = 1, 2$, satisfying

$$\lim_{n\to\infty} E \sup_{t\le T} \left| \int_0^t \int_{U_i} \phi_i(\alpha_i) \left[r_{i,s}^n(d\alpha_i) - r_{i,s}^{n,\epsilon_n}(d\alpha_i) \right] ds \right| = 0, \quad i = 1, 2, \tag{4.13}$$

for each bounded and continuous real-valued nonrandom function $\phi_i(\cdot)$ and each $T < \infty$, and with the following additional properties. For each n, there is a probability space on which are defined an admissible pair $(\tilde{w}^{n,\epsilon}(\cdot), \tilde{r}^{n,\epsilon_n}(\cdot))$, which has the law of $(w^n(\cdot), r^{n,\epsilon_n}(\cdot))$, and associated solution $(\tilde{x}^{n,\epsilon_n}(\cdot), \tilde{y}^{n,\epsilon_n}(\cdot))$ with initial conditions $\tilde{x}^{n,\epsilon_n}(0) = x_n$ such that

$$\liminf_n \left| EF(x^n(\cdot), y^n(\cdot), r^n(\cdot)) - EF(\tilde{x}^{n,\epsilon_n}(\cdot), \tilde{y}^{n,\epsilon_n}(\cdot), \tilde{r}^{n,\epsilon_n}(\cdot)) \right| \ge \rho. \tag{4.14}$$

Equation (4.13) is implied by (4.4).

Now take a weakly convergent subsequence of $\{\tilde{x}^{n,\epsilon_n}(\cdot), \tilde{y}^{n,\epsilon_n}(\cdot), \tilde{w}^{n,\epsilon_n}(\cdot), \tilde{r}^{n,\epsilon_n}(\cdot)\}$, with limit $(\tilde{x}(\cdot), \tilde{y}(\cdot), \tilde{w}(\cdot), \tilde{r}(\cdot))$. Take a weakly convergent subsequence of $\{x^n(\cdot), y^n(\cdot), w^n(\cdot), r^n(\cdot), r^{n,\epsilon_n}(\cdot)\}$ with limit $(x(\cdot), y(\cdot), w(\cdot), r(\cdot), \hat{r}(\cdot))$. By (4.13), $r(\cdot) = \hat{r}(\cdot)$. Also, $(x(\cdot), y(\cdot), w(\cdot), r(\cdot))$ and $(\tilde{x}(\cdot), \tilde{y}(\cdot), \tilde{w}(\cdot), \tilde{r}(\cdot))$ both solve (2.4) with initial condition $x(0)$. Since $(w(\cdot), r(\cdot))$ has the same law as $(w(\cdot), \hat{r}(\cdot))$, hence as $(\tilde{w}(\cdot), \tilde{r}(\cdot))$, the quadruples in the last sentence must be identical in law, which (together with the weak convergence) contradicts (4.14). Thus (4.12) holds.

Now let $F(\cdot)$ be merely bounded and measurable. The first assertion of the last paragraph of the theorem follows from [14, Theorem 3.1(f), Chapter 3]. With this in hand, the uniformity of the convergence in (4.12) is treated as for the case of continuous $F(\cdot)$. □

Finite-valued and piecewise constant approximations $r^\epsilon(\cdot)$ in (4.4): Definitions. Now some approximations of subsequent interest will be defined. They are just piecewise constant and finite-valued ordinary controls. Consider the following discretization of the U_i. Given $\mu > 0$, partition U_i into a finite number of disjoint subsets $C_i^l, l \le p_i$, each with diameter no greater than $\mu/2$. Choose a point $\alpha_i^l \in C_i^l$. Henceforth let p_i be some given function of μ.

Now, given admissible $(r_1(\cdot), r_2(\cdot))$, define the approximating admissible relaxed control $r_i^\mu(\cdot)$ on the control value space $\{\alpha_i^l, l \le p_i\}$ by its derivative as $r_{i,t}^\mu(\alpha_i^l) =$

$r_{i,t}(C_i^l)$. Denote the set of such controls over all $\{C_i^l, \alpha_i^l, l \le p_i\}$ by $\mathcal{U}_i(\mu)$. Let $\mathcal{U}_i(\mu,\delta)$ denote the subset of $\mathcal{U}_i(\mu)$ that are ordinary controls and constant on the intervals $[l\delta, l\delta+\delta), l = 0, 1, \ldots$. Another subclass $\mathcal{U}_i(\mu,\delta,\Delta)$ will be defined above Theorem 4.6.

Finite-valued controls. The proof that (4.4) holds in the next two theorems is straightforward and the details are left to the reader. Under the Lipschitz conditions in A2.4, (4.9) follows from Theorem 4.3, and this implies (4.17). Under A2.5 and A2.6, use Theorem 4.4 to get (4.17).

THEOREM 4.5. *Assume* A2.1–A2.3, A2.5, A2.6*, and the above approximation of* $r_i(\cdot)$ *by* $r_i^\mu(\cdot) \in \mathcal{U}_i(\mu), i = 1, 2$. *Then* (4.4) *and* (4.9) *hold for* μ *replacing* ϵ*, no matter what the* $\{C_i^l, \alpha_i^l\}$. *The same result holds if we approximate only one of the* $r_i(\cdot)$.

Finite-valued, piecewise-constant, and "delayed" approximations. The proof that the game has a value in section 6 depends on showing that the cost changes little if the controls of any player are "delayed" since that implies that the order in which the players act is not important. Let $r_i^\mu(\cdot) \in \mathcal{U}_i(\mu)$, where the control-space values are $\{\alpha_i^l,\ l \le p_i\}$. Let $\Delta > 0$. Define the "backward" differences $\Delta_{i,k}^l = r_i^\mu(\alpha_i^l, k\Delta) - r_i^\mu(\alpha_i^l, k\Delta - \Delta), l \le p_i, k = 1, \ldots$. Define the piecewise constant ordinary controls $u_i^{\mu,\Delta}(\cdot) \in \mathcal{U}_i(\mu,\Delta)$ on the interval $[k\Delta, k\Delta+\Delta)$ by

$$u_i^{\mu,\Delta}(t) = \alpha_i^l \ \text{ for } \ t \in \left[k\Delta + \sum_{\nu=1}^{l-1} \Delta_{i,k}^\nu,\ k\Delta + \sum_{\nu=1}^{l} \Delta_{i,k}^\nu\right). \tag{4.15}$$

Note that, on $[k\Delta, k\Delta+\Delta)$, $u^{\mu,\Delta}(\cdot)$ takes the value α_i^l on a time interval of length $\Delta_{i,k}^l$. Note also that the $u_i^{\mu,\Delta}(\cdot)$ are "delayed" in that the values of $r_i(\cdot)$ on $[k\Delta - \Delta, k\Delta)$ determine the values of $u_i^{\mu,\Delta}(\cdot)$ on $[k\Delta, k\Delta+\Delta)$. Thus $u_i^{\mu,\Delta}(\cdot)$ is $\mathcal{F}_{k\Delta-}$-measurable. This delay will play an important role in the next two sections. Let $r_i^{\mu,\Delta}(\cdot)$ denote the relaxed control representation of $u_i^{\mu,\Delta}(\cdot)$.

The intervals $\Delta_{i,k}^l$ in (4.15) are just real numbers. For use in section 6, it is important to have them be some multiple of some small $\delta > 0$, where Δ/δ is an integer. Consider one method of doing this. Divide $[k\Delta, k\Delta+\Delta)$ into Δ/δ subintervals of length δ each. To each value α_i^l first assign (the integer part) $[\Delta_{i,k}^l/\delta]$ subintervals of length δ. Then assign the remaining unassigned subintervals to the values α_i^l at random with probabilities proportional to the residual (unassigned) lengths $\Delta_{i,k}^l - [\Delta_{i,k}^l/\delta]\delta$, $i \le p_i$. Call the resulting control $u_i^{\mu,\delta,\Delta}(\cdot)$, with relaxed control representation $r_i^{\mu,\delta,\Delta}(\cdot)$. Let $\mathcal{U}_i(\mu,\delta,\Delta)$ denote the set of such controls. If $u_i^{\mu,\delta,\Delta}(\cdot)$ is obtained from $r_i(\cdot)$ in this way, then we will henceforth write it as $u_i^{\mu,\delta,\Delta}(\cdot|r_i)$ to emphasize that fact. Similarly, if $u_i^{\mu,\Delta}(\cdot)$ is obtained from $r_i(\cdot)$, then it will be written as $u_i^{\mu,\Delta}(\cdot|r_i)$. Let $r_{i,t}^{\mu,\Delta}(\cdot|r_i)$ denote the time derivative of $r_i^{\mu,\Delta}(\cdot|r_i)$. As stated in the next theorem, for fixed μ and small δ, $u_i^{\mu,\delta,\Delta}(\cdot|r_i)$ well approximates $u_i^{\mu,\Delta}(\cdot|r_i)$ uniformly in $r_i(\cdot)$ and $\{\alpha_i^l\}$ in that (4.4) holds in the sense that, for each $\mu > 0$ $\Delta > 0$ and bounded and continuous $\phi_i(\cdot)$,

$$\lim_{\delta\to 0} \sup_{r_i \in \mathcal{U}_i} E \sup_{t \le T} \left| \int_0^t \int_{U_i} \phi_i(\alpha_i)[r_{i,s}^{\mu,\Delta}(d\alpha_i|r_i) - r_{i,s}^{\mu,\delta,\Delta}(d\alpha_i|r_i)]ds \right| = 0,\ i = 1, 2. \tag{4.16}$$

THEOREM 4.6. *Assume* A2.1–A2.3, A2.5*, and* A2.6. *For* $r_i(\cdot) \in \mathcal{U}_i$*, approximate as above the theorem to get* $r_i^{\mu,\Delta}(\cdot|r_i) \in \mathcal{U}_i(\mu,\Delta)$ *and* $r_i^{\mu,\delta,\Delta}(\cdot|r_i) \in \mathcal{U}_i(\mu,\delta,\Delta)$. *Then* (4.4) *holds for* $r_i^{\mu,\Delta}(\cdot|r_i)$ *and* (μ,Δ) *replacing* $r_i^\epsilon(\cdot)$ *and* ϵ*, respectively. Also,* (4.16)

holds and

$$\lim_{\Delta\to 0}\lim_{\delta\to 0}\sup_{x}\sup_{r_1\in\mathcal{U}_1}\sup_{r_2\in\mathcal{U}_2}|W(x,r_1,r_2)-W(x,r_1,u_2^{\mu,\delta,\Delta}(\cdot|r_2))|=0. \tag{4.17}$$

For each $\epsilon>0$, there are $\mu_\epsilon>0$ and $\delta_\epsilon>0$ such that, for $\mu\le\mu_\epsilon$ and $\delta\le\delta_\epsilon$ and $r_i(\cdot)\in\mathcal{U}_i, i=1,2$, there are $u_i^{\mu,\delta}(\cdot)\in\mathcal{U}_i(\mu,\delta)$ such that (4.4) holds for $u_i^{\mu,\delta}(\cdot)$ and (μ,δ) replacing $r_i^\epsilon(\cdot)$ and ϵ, respectively, and

$$\sup_{x}\sup_{r_1\in\mathcal{U}_1}\sup_{r_2\in\mathcal{U}_2}|W(x,r_1,r_2)-W(x,r_1,u_2^{\mu,\delta})|\le\epsilon. \tag{4.18}$$

The expressions (4.17) and (4.18) hold with the indices 1 and 2 interchanged.

THEOREM 4.7. *If* A2.7 *replaces* A2.6 *in Theorems* 4.4–4.6*, then their conclusions continue to hold.*

Proof. Only a few details will be given. The last part of Theorem 4.4 will be used, and we need only identify the $F(\cdot)$ for the present case. Theorem 4.2 and (4.6) required only measurability and boundedness of $b(\cdot)$ and $\sigma(\cdot)$. Also, the tightness of any solution sequence $\{x^\epsilon(\cdot),y^\epsilon(\cdot),w^\epsilon(\cdot),r^\epsilon(\cdot)\}$ requires only the measurability and boundedness of $b(\cdot)$ and $\sigma(\cdot)$.

For each $t>0$, define the bounded real-valued function $F(\phi(\cdot),m(\cdot))$ on the product path space of $x(\cdot)$ and $r(\cdot)$ by

$$F(\phi(\cdot),m(\cdot))=\int_0^t\int_{U_1\times U_2}b(\phi(s),\alpha)m_s(d\alpha)ds.$$

Under A2.7, $F(\phi(\cdot),m(\cdot))$ is continuous with probability one with respect to the measure induced by any pair $(x(\cdot),r(\cdot))$ solving (2.4). Let $(x^\epsilon(\cdot),y^\epsilon(\cdot),w^\epsilon(\cdot),r^\epsilon(\cdot))$ satisfy

$$x^\epsilon(t)=x(0)+\int_0^t\int_{U_1\times U_2}b(x^\epsilon(s),\alpha)r^\epsilon(d\alpha)ds+\int_0^t\sigma(x^\epsilon(s))dw^\epsilon(s)+z^\epsilon(t) \tag{4.19}$$

and converge weakly to $(x(\cdot),y(\cdot),w(\cdot),r(\cdot))$ as $\epsilon\to 0$.

For the sake of simplicity and without loss of generality, suppose that the Skorohod representation is used so that all processes are defined on the same probability space and the weak convergence is equivalent to convergence with probability one [14, Theorem 1.8, Chapter 3]. First, suppose that $\sigma(\cdot)$ is continuous but $b(\cdot)$ is not. By the asserted almost everywhere continuity of $F(\cdot)$ and the weak convergence and Skorohod representation, the integral in (4.19) involving $b(\cdot)$ converges to

$$\int_0^t\int_{U_1\times U_2}b(x(s),\alpha)r_s(d\alpha)ds$$

with probability one. Discontinuous $k(\cdot)$ is treated in the same way. Also, the stochastic integral and reflection term converge to those for the limit. The proof for the convergence of the stochastic integral uses a finite sum approximation $\sum_{l\gamma\le t}\sigma(x^\epsilon(l\gamma))\cdot[w^\epsilon(l\gamma+\gamma)-w^\epsilon(l\gamma)]$ and a standard estimate of the errors in this approximation. The proof for the reflection direction is similar to that in [32, Theorem 11.1.2]. The uniformity in $r(\cdot)$ of the approximations, as asserted in (4.9), (4.12), (4.17), and (4.18) in Theorems 4.4–4.6, is shown by a contradiction argument as in Theorem 4.4.

Now, suppose that $\sigma(\cdot)$ is discontinuous but that A2.7 holds. For $\rho>0$, define the real-valued function $f^\rho(\cdot)$ by

$$f^\rho(x)=\begin{cases}1 & \text{for } \operatorname{dist}(x,D_d)\ge\rho,\\ \operatorname{dist}(x,D_d)/\rho & \text{otherwise.}\end{cases}$$

The function $\sigma(\cdot)f^\rho(\cdot)$ is continuous, and, for each $T < \infty$, A2.7 implies that

$$\lim_{\rho\to 0} \sup_{x(0),r} \sup_{\epsilon} E \sup_{t\le T} \left| \int_0^t \sigma(x^\epsilon(s))\,[1 - f^\rho(x^\epsilon(s))]\,dw^\epsilon(s) \right|^2 = 0. \tag{4.20}$$

Now approximate $\sigma(\cdot)$ by $\sigma(\cdot)f^\rho(\cdot)$, and use (4.20) to get the convergence of the stochastic integrals

$$\int_0^t \sigma(x^\epsilon(s))dw^\epsilon(s) \to \int_0^t \sigma(x(s))dw(s). \qquad \square$$

5. Existence of the value of the game.

Motivating the proof that the value exists, i.e., that (2.9) holds. Let player 1 go first, and have a control that is constant on intervals of length Δ_1. Theorems 4.5–4.7 imply that, if the control for player 2 is delayed by more than Δ_1 and discretized in time, then the costs change little for small Δ_1. This delay will mean that player 1 will know player 2's actions before it selects its own. This, in turn, is equivalent to player 2 going first, which (together with the fact that the costs change little) essentially implies that the upper and lower values are as close as we wish. The proof formalizes this idea.

THEOREM 5.1. *Assume* A2.1–A2.3, A2.5, *and either* A2.6 *or* A2.7. *Then the game has a value in that* (2.9) *holds.*

Proof. Let $\Delta_1 > 0$. Let $\Delta, \delta, \mu, \epsilon$ be positive with Δ/δ being an integer. By Theorems 4.5–4.7, for small enough μ, δ, Δ, and large Δ/δ,

$$|W(x, u_1(r_2), r_2) - W(x, u_1(r_2), u_2^{\mu,\delta,\Delta}(\cdot|r_2))| \le \epsilon \tag{5.1}$$

for all $u_1(\cdot) \in \mathcal{L}_1(\Delta_1)$ and $r_2(\cdot) \in \mathcal{U}_2$. Also, for all $\Delta_1 > 0$,

$$\left| \inf_{u_1\in\mathcal{L}_1(\Delta_1)} \sup_{r_2\in\mathcal{U}_2} W(x, u_1(r_2), r_2) - \inf_{u_1\in\mathcal{L}_1(\Delta_1)} \sup_{r_2\in\mathcal{U}_2} W(x, u_1(r_2), u_2^{\mu,\delta,\Delta}(\cdot|r_2)) \right| \le \epsilon. \tag{5.2}$$

The results analogous to (5.1)–(5.2) hold if player 1 goes last. It follows that, in computing the upper or lower values, we can use either relaxed or ordinary controls for the player that goes last.

By the definition (2.7), for each $\Delta_1 > 0$,

$$V^+(x) \le \inf_{u_1\in\mathcal{L}_1(\Delta_1)} \sup_{r_2\in\mathcal{U}_2} W(x, u_1(r_2), r_2). \tag{5.3}$$

Let $\epsilon > 0$. By (5.2), there is $\Delta_\epsilon > 0$ such that, for $\Delta \le \Delta_\epsilon$, there are $\mu > 0$ and $\delta > 0$ such that, for all $\Delta_1 > 0$,

$$\begin{aligned} &\inf_{u_1\in\mathcal{L}_1(\Delta_1)} \sup_{r_2\in\mathcal{U}_2} W(x, u_1(r_2), r_2) \\ &\quad\le \inf_{u_1\in\mathcal{L}_1(\Delta_1)} \sup_{r_2\in\mathcal{U}_2} W(x, u_1(r_2), u_2^{\mu,\delta,\Delta}(\cdot|r_2)) + \epsilon \\ &\quad\le \inf_{u_1\in\mathcal{L}_1(\Delta_1)} \sup_{r_2\in\mathcal{U}_2} W(x, u_1(u_2^{\mu,\delta,\Delta}(\cdot|r_2)), u_2^{\mu,\delta,\Delta}(\cdot|r_2)) + \epsilon. \end{aligned} \tag{5.4}$$

Now let $\Delta_1 < \Delta$ with Δ/Δ_1 being an integer. Recall that the process $u_2^{\mu,\delta,\Delta}(\cdot|r_2)$ is constant on the intervals $[l\delta, l\delta+\delta), l = 0, 1, \dots$, and that $u_2^{\mu,\delta,\Delta}(t|r_2)$ is $\mathcal{F}_{q\Delta}$-measurable for $t \in [q\Delta, q\Delta+\Delta)$. Thus, for integers k and q such that $k\Delta_1 \in [q\Delta, q\Delta+\Delta)$, the inf sup and the use of $u_2^{\mu,\delta,\Delta}(\cdot|r_2)$ on the right of (5.4) can be interpreted to mean that, at each such time $k\Delta_1$, player 1 knows all of player 2's actions on the entire interval $[k\Delta_1, q\Delta+\Delta)$ as well as the data on the "past" up to $k\Delta_1$. Thus one computes the value of the main term on the right side of (5.4) as if player 2 goes first: For $k\Delta_1 \in [q\Delta, q\Delta+\Delta)$, player 1 uses a rule which can be represented in the form

$$P\{u_1(k\Delta_1) \in \cdot \big| u_1(l\Delta_1), l < k;\ u_2^{\mu,\delta,\Delta}(l\delta|r_2), l\delta < q\Delta+\Delta;\ w(s), s < k\Delta_1\}. \tag{5.5}$$

Since it is only the joint probability law that matters, it can be supposed that the value of $u_2(t) = u_2^{\mu,\delta,\Delta}(t|r_2)$ which is actually applied on $[q\Delta, q\Delta+\Delta)$ is determined by a conditional probability law which can be represented in the form

$$P\left\{(u_2(q\Delta+l\delta), l\delta < \Delta) \in \cdot \big| u_2(l\delta), l\delta < q\Delta;\ u_1(s),\ w(s), s < q\Delta\right\}, \tag{5.6}$$

where the $u_2(l\delta)$ take values in a μ-discretization of U_2. Let $\mathcal{L}_2(\mu,\delta,\Delta)$ denote the set of such rules for player 2. The main term on the right side of (5.4) involves arbitrary strategies for player 2 but which are discretized in space and time and delayed. By this fact and the use of the form (5.5), the $u_2^{\mu,\delta,\Delta}(\cdot|r_2)$ can be replaced by a control $u_2(\cdot)$ in $\mathcal{L}_2(\mu,\delta,\Delta)$, and we can write

$$\begin{aligned}
&\inf_{u_1\in\mathcal{L}_1(\Delta_1)} \sup_{r_2\in\mathcal{U}_2} W(x, u_1(u_2^{\mu,\delta,\Delta}(\cdot|r_2)), u_2^{\mu,\delta,\Delta}(\cdot|r_2)) \\
&\quad = \inf_{u_1\in\mathcal{L}_1(\Delta_1)} \sup_{u_2\in\mathcal{L}_2(\mu,\delta,\Delta)} W(x, u_1(u_2), u_2) \\
&\quad = \inf_{u_1\in\mathcal{U}_1(\Delta_1)} \sup_{u_2\in\mathcal{L}_2(\mu,\delta,\Delta)} W(x, u_1(u_2), u_2).
\end{aligned}$$

Now, since player 2 can be considered to "go first," we can write

$$\begin{aligned}
&\inf_{u_1\in\mathcal{U}_1(\Delta_1)} \sup_{u_2\in\mathcal{L}_2(\mu,\delta,\Delta)} W(x, u_1(u_2), u_2) \\
&\quad = \sup_{u_2\in\mathcal{L}_2(\mu,\delta,\Delta)} \inf_{u_1\in\mathcal{U}_1(\Delta_1)} W(x, u_1, u_2(u_1)).
\end{aligned} \tag{5.7}$$

By (5.3), (5.4), and (5.6),

$$V^+(x) \le \sup_{u_2\in\mathcal{L}_2(\mu,\delta,\Delta)} \inf_{u_1\in\mathcal{U}_1(\Delta_1)} W(x, u_1, u_2(u_1)) + \epsilon. \tag{5.8}$$

For small Δ_1 and large Δ/Δ_1,

$$\begin{aligned}
&\left| \sup_{u_2\in\mathcal{L}_2(\mu,\delta,\Delta)} \inf_{u_1\in\mathcal{U}_1(\Delta_1)} W(x, u_1, u_2(u_1)) - \sup_{u_2\in\mathcal{L}_2(\mu,\delta,\Delta)} \inf_{u_1\in\mathcal{U}_1} W(x, u_1, u_2(u_1)) \right| \\
&\qquad \le \epsilon.
\end{aligned}$$

It now follows from this, (5.8), and the definition of $V^-(x)$ that, for small Δ and μ and large Δ/δ and Δ/Δ_1,

$$\begin{aligned}
V^+(x) &\le \sup_{u_2\in\mathcal{L}_2(\mu,\delta,\Delta)} \inf_{u_1\in\mathcal{U}_1(\Delta_1)} W(x, u_1, u_2(u_1)) + \epsilon \\
&\le \sup_{u_2\in\mathcal{L}_2(\mu,\delta,\Delta)} \inf_{u_1\in\mathcal{U}_1} W(x, u_1, u_2(u_1)) + \epsilon \\
&\le \sup_{u_2\in\mathcal{U}_2(\delta)} \inf_{u_1\in\mathcal{U}_1} W(x, u_1, u_2(u_1)) + \epsilon \le V^-(x) + 2\epsilon.
\end{aligned} \tag{5.9}$$

Since ϵ is arbitrary and $V^+(x) \ge V^-(x)$, the theorem is proved. □

6. An auxiliary result: Nearly optimal policies. The proof of convergence of the numerical method in the next section will require the use of particular ϵ-optimal minimizing (resp., maximizing) strategies for player 1 when it goes first (resp., for player 2 when it goes first). Such strategies will be constructed in this section. They are for mathematical purposes only and do not have any practical value otherwise.

In order to motivate the construction, we will first recall the method of proof used for the pure control problem (where there is only a minimizing player) in [32, Chapters 10 and 11]. Let $r^h(\cdot)$ denote the continuous time interpolation of the relaxed control representation of the optimal control for the approximating chain ξ_n^h. Thus the optimal cost $V^h(x)$ equals $W^h(x, r^h)$. The corresponding set $\{\psi^h(\cdot), z^h(\cdot), w^h(\cdot), r^h(\cdot)\}$ was shown to be tight. The limit $(x(\cdot), z(\cdot), w(\cdot), r(\cdot))$ of any weakly convergent subsequence was shown to satisfy the (one-player form of) (2.4). Hence it cannot be better than an optimal solution for (2.4). This implies that $\liminf_h V^h(x) \geq V(x)$, the minimal value of the cost for (2.4).

To finish the convergence proof in [32], we had to show that $\limsup_h V^h(x) \leq V(x)$. This was done in the following way. Given arbitrary $\epsilon > 0$, a special ϵ-optimal control for (2.4) was constructed. This special control was such that it could be adapted for use on the approximating chain, and for small h the interpolated chain well approximated the limit process under that control. In more detail, let $r^\epsilon(\cdot)$ denote the relaxed control form of this special ϵ-optimal control for (2.4), with Wiener process $w^\epsilon(\cdot)$ and associated solution and reflection process $(x^\epsilon(\cdot), z^\epsilon(\cdot))$. Let $r^{\epsilon,h}(\cdot)$ denote the relaxed control form of the adaptation of this special control for use on the chain ξ_n^h, interpolated to continuous time, and let $(\psi^{\epsilon,h}(\cdot), z^{\epsilon,h}(\cdot), w^{\epsilon,h}(\cdot))$ denote the continuous time interpolation of the corresponding solution, reflection process and "pre-Wiener" process in the representation (3.11). Since $r^{\epsilon,h}(\cdot)$ is no better than the optimal control for the chain, $V^h(x) \leq W^h(x, r^{\epsilon,h})$. By the method of construction of $r^{\epsilon,h}(\cdot)$, the set $(\psi^{\epsilon,h}(\cdot), z^{\epsilon,h}(\cdot), w^{\epsilon,h}(\cdot), r^{\epsilon,h}(\cdot))$ converged weakly to the set $(x^\epsilon(\cdot), z^\epsilon(\cdot), w^\epsilon(\cdot), r^\epsilon(\cdot))$, with ϵ-optimal cost $W(x, r^\epsilon)$. Since ϵ is arbitrary, we have $\limsup_h V^h(x) \leq V(x)$, which completes the proof that $V^h(x) \to V(x)$. See the references for full details.

Such an ϵ-optimal control for (2.4) (whether minimizing or maximizing) for the player that goes first plays a similar role for the game problem of this paper. The construction follows the general lines of what was done in [32, Theorem 10.3.1], but there are some very important differences since we must work with strategies, where the two controls depend on each other, which is not the case for the pure (i.e., one-player) control problem. The construction is done as it is since we know little about nearly optimal policies in general. For example, we do know whether there are smooth ϵ-optimal feedback controls for either player, in general.

THEOREM 6.1. *Assume* A2.1–A2.3, A2.5, *and either* A2.6 *or* A2.7. *Let player* 1 *go first. Then, for each* $\epsilon > 0$, *there is an* ϵ*-optimal minimizing control law for player* 1 *with the following properties. For positive* Δ, δ, *and* ρ, *let* δ/ρ *and* Δ/δ *be integers. The control is constant on the intervals* $[k\Delta, k\Delta + \Delta)$, $k = 0, 1, \ldots$, *finite-valued, the value at* $k\Delta$ *is* $\mathcal{F}_{k\Delta}$*-measurable, and, for small* $\lambda > 0$, *it is defined by the conditional probability law* (*which defines the function* $q_{i,k}(\cdot)$)

$$
\begin{aligned}
& P\left\{u_1(k\Delta) = \gamma \middle| u_1(l\Delta), l < k;\ w(s), r_2(s), s < k\Delta\right\} \\
& = P\{u_1(k\Delta) = \gamma \big| w(l\lambda), l\lambda < k\Delta;\ u_1(l\Delta), l < k;\ u_2^{\mu,\rho,\delta}(l\rho|r_2), l\rho < k\Delta\} \\
& = q_{1,k}(\gamma;\ w(l\lambda), l\lambda < k\Delta;\ u_1(l\Delta), l < k;\ u_2^{\mu,\rho,\delta}(l\rho|r_2), l\rho < k\Delta).
\end{aligned} \tag{6.1}
$$

The function $q_{1k}(\cdot)$ *is continuous in the* w*-arguments for each value of the others. Since the rule* (6.1) *depends on* $r_2(\cdot)$ *only via* $u_2^{\mu,\rho,\delta}(\cdot|r_2)$ (*which is defined above*

Theorem 4.6), *we write the rule as* $\bar{u}_1^\epsilon(u_2^{\mu,\rho,\delta}(\cdot|r_2))$. *In particular, for small* λ, μ, Δ *and large* δ/ρ *and* Δ/δ, *it satisfies the inequality*

$$\sup_{r_2 \in \mathcal{U}_2} W(x, \bar{u}_1^\epsilon(u_2^{\mu,\rho,\delta}(\cdot|r_2)), r_2) \le V(x) + \epsilon. \tag{6.2}$$

Also, if $r_2^n(\cdot)$ *is a sequence which converges weakly to some* $r_2(\cdot)$, *then*

$$\limsup_n W(x, \bar{u}_1^\epsilon(u_2^{\mu,\rho,\delta}(\cdot|r_2^n)), r_2^n) \le V(x) + \epsilon. \tag{6.3}$$

For each $r_2(\cdot)$ *and* $l = 0, 1, \dots$, *let* $\tilde{u}_2^{\mu,\rho,\delta}(l\rho|r_2)$ *be a control that differs from* $u_2^{\mu,\rho,\delta}(l\rho|r_2)$ *by at most* μ *in absolute value. Then* (6.2) *and* (6.3) *hold for the perturbation* $\tilde{u}_2^{\mu,\rho,\delta}(\cdot|r_2)$ *replacing* $u_2^{\mu,\rho,\delta}(\cdot|r_2)$.

Similarly, if player 2 *goes first, then there is an* ϵ*-optimal control rule of the same type: In particular, and with the analogous terminology,*

$$\inf_{r_1 \in \mathcal{U}_1} W(x, r_1, \bar{u}_2^\epsilon(u_1^{\mu,\rho,\delta}(\cdot|r_1))) \ge V(x) - \epsilon, \tag{6.4}$$

and (6.4) *continues to hold with the perturbation* $\tilde{u}_1^{\mu,\rho,\delta}(\cdot|r_1)$ *replacing* $u_1^{\mu,\rho,\delta}(\cdot|r_1)$.

Proof. Recall the approximation of the U_i given above Theorem 4.5: Given $\mu_1 > 0$, U_i was partitioned into a finite number of disjoint subsets $C_i^l, l \le p_i$, each with diameter no greater than $\mu_1/2$. A point α_i^l in each C_i^l was chosen. These are the values of γ in (6.1). Let player 1 go first. Given $\epsilon > 0$, there are $\Delta > 0, \mu_1 > 0$, and an $\epsilon/8$-optimal rule for player 1 which can be represented in the "conditional probability" form

$$P\left\{u_1(k\Delta) = \gamma \middle| u_1(l\Delta), l < k;\, w(s), r_2(s), s \le k\Delta\right\}. \tag{6.5}$$

Call the rule $u_1^\epsilon(r_2)$. Then, by the $\epsilon/8$-optimality, for all $r_2(\cdot)$, we have

$$W(x, u_1^\epsilon(r_2), r_2) \le V^+(x) + \epsilon/8. \tag{6.6}$$

The rule (6.5) needs to be approximated so that it depends only on selected samples of the data.

Whatever $r_2(\cdot)$, $u_2^{\mu,\rho,\delta}(\cdot|r_2)$ is also an admissible control. Hence, by (6.6), for all $r_2(\cdot) \in \mathcal{U}_2$,

$$W(x, u_1^\epsilon(u_2^{\mu,\rho,\delta}(\cdot|r_2)), u_2^{\mu,\rho,\delta}(\cdot|r_2)) \le V^+(x) + \epsilon/8. \tag{6.7}$$

Indeed, (6.7) holds for all $u_2^{\mu,\rho,\delta}(\cdot|r_2)$, irrespective of $r_2(\cdot)$. Let μ, ρ, and δ be positive numbers with Δ/δ and δ/ρ being integers. For small μ and δ and large δ/ρ, Theorems 4.5–4.7 imply that

$$W(x, u_1^\epsilon(u_2^{\mu,\rho,\delta}(\cdot|r_2)), r_2) \le W(x, u_1^\epsilon(u_2^{\mu,\rho,\delta}(\cdot|r_2)), u_2^{\mu,\rho,\delta}(\cdot|r_2)) + \epsilon/8 \tag{6.8}$$

for all $r_2(\cdot) \in \mathcal{U}_2$. The control law $u_1^\epsilon(u_2^{\mu,\rho,\delta}(\cdot|r_2))$ can be represented in the form

$$\begin{aligned} &P\{u_1(k\Delta) = \gamma \big| u_1(l\Delta), l < k;\, r_2(s), w(s), s < k\Delta\} \\ &= P\{u_1(k\Delta) = \gamma \big| w(s), s < k\Delta\ \ u_1(l\Delta); l < k;\, u_2^{\mu,\rho,\delta}(l\rho|r_2), l\rho < k\Delta\}. \end{aligned} \tag{6.9}$$

Let $r_2^n(\cdot)$ converge to $r_2(\cdot)$ as $n \to \infty$. Then the discrete approximations $u_2^{\mu,\rho,\delta}(\cdot|r_2^n)$ do not necessarily converge to $u_2^{\mu,\rho,\delta}(\cdot|r_2)$ as $n \to \infty$. They will converge if the limit

"mass" on the boundaries of the sets $\{C_2^m, m \le p_2\}$ into which U_2 is subdivided is zero—more particularly, if

$$r_2(l\delta + \delta, \partial C_2^m) - r_2(l\delta, \partial C_2^m) = 0$$

for all m, l [14, Chapter 3, Theorem 3.1(f)]. Such convergence is hard to ensure for arbitrary $r_2(\cdot)$. The problem is due to the fact that the sets C_2^n are not all closed so that part of the boundary of some set will actually be in a neighboring set. Owing to our use of a μ-discretization of the U_i, the worst that can happen is that $u_2^{\mu,\rho,\delta}(l\rho|r_2^n)$ will differ from $u_2^{\mu,\rho,\delta}(l\rho|r_2)$ by at most μ for each l in the limit as $n \to \infty$. For each $r_2(\cdot)$, let $\tilde{u}_2^{\mu,\rho,\delta}(\cdot|r_2)$ be any admissible control satisfying

$$\sup_l |u_2^{\mu,\rho,\delta}(l\rho|r_2) - \tilde{u}_2^{\mu,\rho,\delta}(l\rho|r_2)| \le \mu.$$

Then, as $n \to \infty$ and $r_2^n(\cdot) \to r_2(\cdot)$, we will have $u_1^\epsilon(u_2^{\mu,\rho,\delta}(\cdot|r_2^n)) \to u_1^\epsilon(\tilde{u}_2^{\mu,\rho,\delta}(\cdot|r_2))$ for some perturbation $\tilde{u}_2^{\mu,\rho,\delta}(\cdot|r_2)$ that differs from $\tilde{u}_2^{\mu,\rho,\delta}(\cdot|r_2)$ by a most μ at each time point. For small μ and large δ/ρ, it will be seen that inequality (6.12) holds, and that is all that will be needed.

For small μ and δ and large δ/ρ, (6.8) yields

$$W(x, u_1^\epsilon(\tilde{u}_2^{\mu,\rho,\delta}(\cdot|r_2)), r_2) \le W(x, u_1^\epsilon(\tilde{u}_2^{\mu,\rho,\delta}(\cdot|r_2)), \tilde{u}_2^{\mu,\rho,\delta}(\cdot|r_2)) + \epsilon/8 \tag{6.10}$$

for all $r_2(\cdot) \in \mathcal{U}_2$. Inequalities (6.10) and (6.6) (with all $r_2(\cdot)$ replaced by $\tilde{u}_2^{\mu,\rho,\delta}(\cdot|r_2)$) imply that, for all $r_2(\cdot)$ and all such perturbations $\tilde{u}_2^{\mu,\rho,\delta}(\cdot|r_2)$,

$$W(x, u_1^\epsilon(\tilde{u}_2^{\mu,\rho,\delta}(\cdot|r_2)), r_2) \le V^+(x) + 2\epsilon/8. \tag{6.11}$$

Hence, for small μ and δ and large δ/ρ, the rule (6.9), but with any perturbation $\tilde{u}_2^{\mu,\rho,\delta}(\cdot|r_2)$ used in lieu of $u_2^{\mu,\rho,\delta}(\cdot|r_2)$, still yields a $2\epsilon/8$-optimal rule for player 1 if it goes first. Furthermore, if $r_2^n(\cdot)$ converges to $r_2(\cdot)$, then

$$\limsup_n W(x, u_1^\epsilon(u_2^{\mu,\rho,\delta}(\cdot|r_2^n)), r_2^n) = W(x, u_1^\epsilon(\tilde{u}_2^{\mu,\rho,\delta}(\cdot|r_2)), r_2) \le V^+(x) + 2\epsilon/8 \tag{6.12}$$

for some perturbation $\tilde{u}_2^{\mu,\rho,\delta}(\cdot|r_2)$.

The next step is to approximate the right side of (6.9) so that it depends only on samples of the $w(\cdot)$. Other than the $w(\cdot)$-variables, owing to the discretizations of time and control value, the conditioning data in (6.9) takes only a finite number of values. By the martingale convergence theorem, as $\lambda \to 0$, the function defined by

$$\begin{aligned} &q_{1,k}^\lambda(\gamma;\, w(l\lambda), l\lambda < k\Delta;\, u_1(l\Delta), l < k;\, u_2^{\mu,\rho,\delta}(l\rho|r_2), l\rho < k\Delta) \\ &\quad \equiv P\{u_1(k\Delta) = \gamma \big| w(l\lambda), l\lambda < k\Delta;\, u_1(l\Delta), l < k;\, u_2^{\mu,\rho,\delta}(l\rho|r_2), l\rho < k\Delta\} \end{aligned} \tag{6.13}$$

converges to

$$P\{u_1(k\Delta) = \gamma \big| w(s), s < k\Delta;\, u_1(l\Delta), l < k; u_2^{\mu,\rho,\delta}(l\rho|r_2), l\rho < k\Delta\}$$

for almost all $w(\cdot)$, for each value of the other conditioning variables. Thus, for small enough λ, the rule (6.9) can be approximated by (6.13). If the new rule is called $\hat{u}_1^{\epsilon,\lambda}(u_2^{\mu,\rho,\delta}(\cdot|r_2))$, then, for small λ,

$$W(x, \hat{u}_1^{\epsilon,\lambda}(u_2^{\mu,\rho,\delta}(\cdot|r_2)), r_2) \le V^+(x) + 6\epsilon/8,$$

and for small λ, μ, ρ, and δ and large Δ/δ and δ/ρ, the same inequality holds if the perturbation $\tilde{u}_2^{\mu,\rho,\delta}(\cdot|r_2)$ replaces $u_2^{\mu,\rho,\delta}(\cdot|r_2)$. We can suppose, without loss of generality, that Δ/λ is an integer.

There is one more approximation since we will require that the function $q_{1,k}(\cdot)$ in (6.1) be continuous in the $w(l\lambda)$-variables for each value of the others. Fix k, and let m denote the dimension of $w(1)$. Let $n = [k\Delta/\lambda] - 1$. Let $\bar{\alpha}$ denote the canonical value of the entire set $\{u_1(l\Delta), l < k\}$, and let $\bar{\beta}$ denote the canonical value of the entire set $\{u_2(l\rho), l\rho < k\Delta\}$. Let $w_\nu, v_\nu, \nu \le n$, be vectors in $\mathbb{R}^m$. For $\kappa > 0$, define the smoothed function

$$
\begin{aligned}
&q_{1,k}^{\lambda,\kappa}\left(\gamma\middle|\bar{\alpha},\bar{\beta};\, w_\nu, \nu \le n\right) \\
&= \frac{1}{[2\pi\kappa]^{nm/2}} \int \cdots \int e^{-|w_\nu - v_\nu|^2/[2\kappa]} q_{1,k}^{\lambda}\left(\gamma\middle|\bar{\alpha},\bar{\beta};\, v_\nu, \nu \le n\right) dv_1 \cdots dv_n.
\end{aligned} \tag{6.14}
$$

The smoothed function defined by (6.14) is continuous in $\{w_\nu, \nu \le n\}$ for each value of $\bar{\alpha}, \bar{\beta}, \gamma$. As $\kappa \to 0$, for each $\bar{\alpha}, \bar{\beta}, \gamma$, it converges to $q_{1,k}^{\lambda}\left(\gamma\middle|\bar{\alpha},\bar{\beta};\, w_\nu, \nu \le n\right)$ for almost all (Lebesgue measure) $\{w_\nu, \nu \le n\}$. Since the measure of $\{w(l\lambda), l\lambda < k\Delta\}$ is absolutely continuous with respect to Lebesgue measure, the convergence is for almost all $w(\cdot)$. Finally, defining the function $q_{1,k}(\cdot)$ in (6.1) by $q_{1,k}^{\lambda,\kappa}(\cdot)$ for small enough λ and κ and calling the resulting control law $\bar{u}_1^{\epsilon}(u_2^{\mu,\rho,\delta}(\cdot|r_2))$, we have (6.1) and (6.3), and $q_{1,k}(\cdot)$ is continuous in the w-variables for each value of the others.

If players 1 and 2 are interchanged in all of the above arguments, then we get an ϵ-optimal (maximizing) rule analogous to the form (6.1) for player 2, and (6.4) holds. □

7. Convergence of the numerical solutions. The next theorem establishes the convergence of the numerical procedure. It supposes the local consistency condition (3.1)–(3.3) everywhere, but recall the remarks concerning discontinuous dynamical and cost terms below (3.1). We do not show the convergence of the controls. In numerical examples, the sequence of optimal feedback controls for the chain does converge as well, and, in all examples of which we are aware, it is of a form that can be shown to be optimal. This would be the case if the optimal feedback controls $\bar{u}_i^h(\cdot)$ for the chain converged to feedback controls $\bar{u}_i(\cdot)$, where the convergence is uniform and the limits are continuous outside of an arbitrarily small neighborhood of a set D_d satisfying A2.7, and the process (2.4) under the $\bar{u}_i(\cdot)$ is unique in the weak sense. Then $W(x, \bar{u}_1, \bar{u}_2) = V(x)$.

THEOREM 7.1. *Assume the local consistency conditions* (3.1)–(3.3), A2.1–A2.3, A2.5, *and either* A2.6 *or* A2.7. *Then* $V^{\pm,h}(x) \to V(x)$ *as* $h \to 0$.

Proof. Let player 1 go first. Given $\epsilon > 0$, let us adapt the ϵ-optimal (minimizing) rule $\bar{u}_1^{\epsilon}(u_2^{\mu,\rho,\delta}(\cdot|r_2))$ for (2.4) that is defined by (6.1) for use on the chain. With player 1 using this rule for the Markov chain model, for each integer k player 1 uses a constant control value on the interpolated time interval $[k\Delta, k\Delta + \Delta)$. The continuous time interpolation of the relaxed control representation of the control processes which are used for the two players will be written as $r^h(\cdot) = (r_1^h(\cdot), r_2^h(\cdot))$. Thus, for some small positive μ, ρ, δ, and λ, the adaptation of the rule (6.1) for player 1 for the chain can be represented by the form, where $w^h(\cdot)$ is the "pre-Wiener" process in (3.11),

$$
\begin{aligned}
&P\left\{u_1^h(k\Delta) = \gamma\middle|u_1^h(l\Delta), l < k;, r_2^h(s), w^h(s), s \le t\right\} \\
&= P\{u_1^h(k\Delta) = \gamma\middle|u_1^h(l\Delta), l < k;\, u_2^{\mu,\rho,\delta}(l\rho|r_2^h), l\rho < k\Delta;\, w^h(l\lambda), l\lambda < k\Delta\} \\
&= q_{1,k}(\gamma\middle|u_1^h(l\Delta), l < k;\, u_2^{\mu,\rho,\delta}(l\rho|r_2^h), l\rho < k\Delta;\, w^h(l\lambda), l\lambda < k\Delta).
\end{aligned} \tag{7.1}
$$

Given the rule (7.1) for player 1, player 2 selects a maximizing control at each state transition. Let $u_2^h(\cdot)$ denote player 2's optimal choice. Then $r_2^h(\cdot)$ is its relaxed control representation, and $r_1^h(\cdot)$ is the relaxed control representation of the realization of player 1's actions which are determined by (7.1).

Choose a weakly convergent subsequence of $\{\psi^h(\cdot), r^h(\cdot), w^h(\cdot), y^h(\cdot)\}$ (abusing terminology, for simplicity this subsequence is also indexed by h). This converges weakly to a solution $(x(\cdot), r(\cdot), w(\cdot), y(\cdot))$ of (2.4) and $W^h(x, r^h) \to W(x, r)$. The proofs of these facts are the same as for the pure control problem in [32, Theorems 11.1.2 and 11.1.5]. Let us use the Skorohod representation [14, Theorem 1.8, Chapter 3] so that all processes are defined on the same probability space, and weak convergence becomes convergence with probability one. If

$$r_2(l\delta + \delta, \partial C_2^m) - r_2(l\delta, \partial C_2^m) = 0$$

for all m, l, then, using the continuity of $q_{1k}(\cdot)$ in the w-variables,

$$\begin{aligned} &q_{1,k}(\gamma\big|u_1^h(l\Delta), l < k;\, u_2^{\mu,\rho,\delta}(l\rho|r_2^h), l\rho < k\Delta;\, w^h(l\lambda), l\lambda < k\Delta) \\ &\quad \to q_{1,k}(\gamma\big|u_1(l\Delta), l < k;\, u_2^{\mu,\rho,\delta}(l\rho|r_2), l\rho < k\Delta;\, w(l\lambda), l\lambda < k\Delta) \end{aligned} \tag{7.2}$$

with probability one,

$$W^h(x, \bar{u}_1^\epsilon(u_2^{\mu,\rho,\delta}(\cdot|r_2^h)), r_2^h) \to W(x, \bar{u}_1^\epsilon(u_2^{\mu,\rho,\delta}(\cdot|r_2)), r_2), \tag{7.3}$$

and (6.2) holds. In any case, as noted in the paragraph below (6.9),

$$W^h(x, \bar{u}_1^\epsilon(u_2^{\mu,\rho,\delta}(\cdot|r_2^h)), r_2^h) \to W(x, \bar{u}_1^\epsilon(\tilde{u}^{\mu,\rho,\delta}(\cdot|r_2)), r_2), \tag{7.4}$$

where $\tilde{u}^{\mu,\rho,\delta}(\cdot|r_2)$ is a perturbation of the type defined in Theorem 6.1. We have, where $\tilde{r}_i(\cdot)$ is the relaxed control representations of the canonical $\tilde{u}_i(\cdot)$,

$$\begin{aligned} V^{+,h}(x) &= \inf_{\tilde{u}_1 \in \mathcal{U}_1^h(1)} \sup_{\tilde{u}_2 \in \mathcal{U}_2^h(2)} W^h(x, \tilde{u}_1, \tilde{u}_2) \le \sup_{\tilde{u}_2 \in \mathcal{U}^h(2)} W^h(x, \bar{u}_1^\epsilon(u_2^{\mu,\rho,\delta}(\cdot|\tilde{r}_2)), \tilde{r}_2) \\ &= W^h(x, \bar{u}_1^\epsilon(u_2^{\mu,\rho,\delta}(\cdot|r_2^h)), r_2^h) \to W(x, \bar{u}_1^\epsilon(\tilde{u}_2^{\mu,\rho,\delta}(\cdot|r_2)), r_2). \end{aligned} \tag{7.5}$$

This and the inequality (6.3) imply that, for any $\epsilon > 0$,

$$\limsup_h V^{+,h}(x) \le V(x) + \epsilon. \tag{7.6}$$

Now repeat the procedure, but with player 2 going first. Use the analogue of the ϵ-optimal rule (6.1) for player 2. Then, given that rule for player 2, let player 1 optimize (minimize). Writing $r^h(\cdot)$ for the actual control process, we have the analogue of (7.5), namely,

$$\begin{aligned} V^{-,h}(x) &= \sup_{\tilde{u}_2 \in \mathcal{U}_2^h(1)} \inf_{\tilde{u}_1 \in \mathcal{U}_1^h(2)} W^h(x, \tilde{u}_1, \tilde{u}_2) \ge \inf_{\tilde{u}_1 \in \mathcal{U}_1^h(2)} W^h(x, \tilde{r}_1, \bar{u}_2^\epsilon(u_1^{\mu,\rho,\delta}(\cdot|\tilde{r}_1))) \\ &= W^h(x, r_1^h, \bar{u}_2^\epsilon(u_1^{\mu,\rho,\delta}(\cdot|r_1^h))) \to W(x, r_1, r_2) = W(x, r_1, \bar{u}_2^\epsilon(\tilde{u}_1^{\mu,\rho,\delta}(\cdot|r_1))), \end{aligned}$$

where $\tilde{u}_1^{\mu,\rho,\delta}(\cdot|r_1)$ is a perturbation of $u_1^{\mu,\rho,\delta}(\cdot|r_1)$. Using this and (6.4) yields

$$\liminf_h V^{-,h}(x) \ge V(x) - \epsilon. \tag{7.7}$$

Finally, (7.5) and (7.7) yield

$$\limsup_h V^{+,h}(x) - \liminf_h V^{-,h}(x) \le 2\epsilon,$$

and the proof is concluded since $\epsilon > 0$ is arbitrary. □

8. Comments and extensions.

8.1. Examples with separated dynamics. Only a few comments will be made since the interest is in reminding the reader of the connection between risk-sensitive, robust, constrained, and large deviations control and differential games.

Risk-sensitive control. Let $\epsilon > 0$, and consider the problem of minimizing

$$\Lambda^{\epsilon}(u_1) = \lim_{T\to\infty} \frac{1}{T} \log E_x \exp\left[\frac{1}{\epsilon}\int_0^T L(x(s), u_1(s))ds\right]$$

for bounded and continuous $L(\cdot)$ with dynamics given by

$$dx = b(x, u_1)dt + \left[\frac{\epsilon}{2\gamma}\right]^{1/2} \sigma(x)dw + dz,$$

where $u_1(t) \in U_1$, a compact set. This is part of the subject of risk-sensitive control [18]. Under appropriate conditions, the solution reduces to that of a differential game with separated dynamics. Let $\bar{\Lambda}^{\epsilon} = \inf_{u_1} \Lambda^{\epsilon}(u_1)$, and define $a(x) = \sigma(x)\sigma'(x)$, assumed positive definite for each x. For x in the interior of G, the Isaacs equation is [18]

$$\bar{\Lambda}^{\epsilon} = \frac{\epsilon}{4\gamma^2}\sum_{i,j} V_{x_i x_j}(x)a_{ij}(x) + \max_{u_2}[V_x'(x)u_2 - \gamma^2 |u_2|^2] + \min_{u_1}\left[V_x'(x)b(x, u_1) + L(x, u_1)\right].$$

This corresponds to a two-person game with cost rate $k(x, u) = L(x, u_1) - \gamma^2|u_2|^2$. Only u_1 appears in the dynamical equation.

In applications, the set U_1 is often unbounded. Effective approaches to dealing with unbounded sets for the control problem are in the chapters concerning the variational problems in [32], and they can be adapted to the game problem under appropriate conditions.

Constrained optimization via the Lagrangian method. Consider the model (2.4), but with only one control $u_1(\cdot)$. Let $q_i(\cdot), i \le p$, be bounded, continuous, and continuously differentiable real-valued functions on G, and consider the minimization of

$$E\int_0^{\infty} e^{-\beta t} k_1(x(t), u_1(t))dt$$

for a bounded and continuous function subject to the constraints $Eq_i(x(t)) \le 0$ for almost all t, $i = 1, \dots, \mu$. The problem can be formulated as a game, via the introduction of Lagrange multipliers $u_{2,i}(t) \ge 0, i = 1, \dots, p$. Define

$$W(x, u_1, u_2) = E\int_0^{\infty} e^{-\beta t}\left[k_1(x(t), u_1(t)) + u_2'(t)q(x(t))\right]dt.$$

Then the solution is obtained from the game with upper (and lower as well, since the game has a value) value

$$\lim_{\Delta\to 0} \inf_{u\in\mathcal{U}_1(\Delta)} \sup_{u_2\in\mathcal{U}_2} W(x, u_1(u_2), u_2).$$

Here the set U_2 is $[0,\infty)$. However, for numerical purposes, one bounds the interval and then experiments with the bound until the desired solution is obtained.

Controlled large deviations problems. Consider the problem in controlled large deviations where one wishes to minimize (over choice of a control) the large deviations estimate of the probability of an event, say the probability that a set will be exited over some time interval. The mathematical formulation of such problems for diffusion-type models often reduces to that of a game, where the dynamics and cost function are separated, analogously to the forms of $b(\cdot)$ and $k(\cdot)$ in (2.4) and (2.5). See for example, the development in [12].

8.2. Stopping time problems and pursuit-evasion games.

Stopping cost not depending on who stops first. Suppose that player i now has a choice of an $\mathcal{F}_t$-stopping time τ_i as well as of the controls. Define $\tau = \min\{\tau_1, \tau_2\}$. For a continuous function $g(\cdot)$, replace (2.5) by

$$W(x,r,\tau) = E\int_0^\tau e^{-\beta t}\left[\int_{U_i}\sum_{i=1}^{2} k_i(x(t),\alpha_i)r_{i,t}(d\alpha_i)dt + c'dy(t)\right] + Ee^{-\beta\tau}g(x(\tau)). \tag{8.1}$$

Thus, in this model, the stopping cost $g(x(\tau))$ does not depend on who selects the stopping time.

The control spaces such as $\mathcal{U}_i$, $\mathcal{U}_i(\Delta), \mathcal{L}_i(\Delta)$, and $\mathcal{U}_i(\mu,\delta,\Delta)$, etc. need to be extended so that they include the stopping times. Let $\overline{\mathcal{U}}_i$ be the set of pairs $(u_i(\cdot),\tau)$, where $u_i(\cdot)\in\mathcal{U}_i$ and τ is an $\mathcal{F}_t$-stopping time. Let $\overline{\mathcal{U}}_i(\Delta)$ denote the subset where $u_i(\cdot)\in\mathcal{U}_i(\Delta)$ and τ takes values $k\Delta, k=0,1\ldots$, where the set $\{\omega:\tau=k\Delta\}$ is $\mathcal{F}_{k\Delta}$-measurable. Similarly, $\overline{\mathcal{U}}_i(\mu,\delta,\Delta)$ denotes the subset of $\overline{\mathcal{U}}_i(\Delta)$, where $u_i(\cdot)\in\mathcal{U}_i(\mu,\delta,\Delta)$. Let $\overline{\mathcal{L}}_1(\Delta)$ denote the set of controls in $\overline{\mathcal{U}}_1(\Delta)$ for player 1 which can be represented in the form

$$\begin{aligned} &P\left\{\tau_1 > k\Delta \middle| w(s), u_2(s), s<t; u_1(l\Delta), l<k, \tau_1\ge k\Delta\right\},\\ &P\left\{u_1(k\Delta)\in\cdot\middle| w(s), u_2(s), s<t; u_1(l\Delta), l<k;\ \tau_1 > k\Delta\right\}. \end{aligned} \tag{8.2}$$

Define $\mathcal{L}_2(\Delta)$ analogously for player 2.

The definitions of the upper and lower values in (2.6) are replaced by, respectively,

$$\begin{aligned} V^+(x) &= \lim_{\Delta\to 0}\inf_{u_1,\tau_1\in\overline{\mathcal{L}}_1(\Delta)}\sup_{(u_2,\tau_2)\in\overline{\mathcal{U}}_2} W(x,u_1,u_2,\tau),\\ V^-(x) &= \lim_{\Delta\to 0}\sup_{(u_2,\tau_2)\in\overline{\mathcal{L}}_2(\Delta)}\inf_{(u_1,\tau_1)\in\overline{\mathcal{U}}_1} W(x,u_1,u_2,\tau). \end{aligned} \tag{8.3}$$

The first line of (8.3) is to be understood as follows. Suppose that the game has not stopped by time $k\Delta$. Then, at $k\Delta$, player 1 goes first and decides whether to stop based on data to time $k\Delta-$. If it stops, the game is over. If not, it selects the control value $u_1(k\Delta)$ (which it will use until $(k\Delta+\Delta)-$ or until player 2 stops, whichever comes first) based on data to time $k\Delta-$. If the game is not stopped at $k\Delta$ by player 1, then player 2 has the opportunity to stop at any time on $[k\Delta, k\Delta+\Delta)$, with the decision to stop at any time being based on all data to that time. Until it stops (if it does), it chooses admissible control values $u_2(\cdot)$. The procedure is then repeated at time $k\Delta+\Delta$, and so forth. With these changes and minor (mostly notational) modifications, the previous theorems continue to hold. In particular, Theorem 7.1 holds.

Stopping cost depends on who stops first. Now let the cost be

$$W(x,r,\tau_1,\tau_2) = E\int_0^{\tau} e^{-\beta t}\left[\int_{U_i}\sum_{i=1}^{2} k_i(x(t),\alpha_i)r_{i,t}(d\alpha_i)dt + c'dy(t)\right] + Ee^{-\beta\tau_1}g_1(x(\tau_1))I_{\{\tau_1<\tau_2\}} + Ee^{-\beta\tau_2}g_2(x(\tau_2))I_{\{\tau_2\le\tau_1\}}, \tag{8.4}$$

where $\tau = \min\{\tau_1,\tau_2\}$ and the $g_i(\cdot)$ are bounded and continuous. The proof in Theorem 5.1 that the game has a value does not carry over to the present case, since the stopping cost depends on who stops first. However, if the game has a value, then Theorem 7.1 holds.

Consider the approximating Markov chain. Let player 1 go first, and let I_1 denote the indicator of the event that player 1 stops at the current step. Then the Bellman equation for the (for example) upper value is

$$\begin{aligned} V^{+,h}(x) = \min_{I_1,\alpha_1}\{&g_1(x)I_1, \\ &(1-I_1)\max[\max_{\alpha_2}(E_x^{\alpha}e^{-\beta\Delta t^h(x,\alpha)}V^{+,h}(\xi_1^h) + k(x,\alpha)\Delta t^h(x,\alpha)), g_2(x)]\}. \end{aligned} \tag{8.5}$$

Acknowledgments. The author greatly appreciates the many helpful suggestions of Paul Dupuis and Wendell Fleming.

REFERENCES

[1] E. Altman and H. J. Kushner, *Admission control for combined guaranteed performance and best effort communications systems under heavy traffic*, SIAM J. Control Optim., 37 (1999), pp. 1780–1807.

[2] J. A. Ball, M. Day, and P. Kachroo, *Robust feedback control of a single server queueing system*, Math. Control Signals Systems, 12 (1999), pp. 307–345.

[3] J. A. Ball, M. Day, P. Kachroo, and T. Yu, *Robust L_2-gain for nonlinear systems with projection dynamics and input constraints: An example from traffic control*, Automatica J. IFAC, 35 (1999), pp. 429–444.

[4] M. Bardi, S. Bottacin, and M. Falcone, *Convergence of discrete schemes for discontinuous value functions of pursuit-evasion games*, in New Trends in Dynamic Games and Applications, G.I. Oldser, ed., Birkhäuser Boston, Boston, 1995, pp. 273–304.

[5] M. Bardi, M. Falcone, and P. Soravia, *Fully discrete schemes for the value function of pursuit-evasion games*, in Advances in Dynamic Games and Applications, T. Basar and A. Haurie, eds., Birkhäuser Boston, Boston, 1994, pp. 89–105.

[6] M. Bardi, M. Falcone, and P. Soravia, *Numerical methods for pursuit-evasion games via viscosity solutions*, in Stochastic and Differential Games: Theory and Numerical Methods, M. Bardi, T. E. S. Raghavan, and T. Parthasarathy, eds., Birkhäuser Boston, Boston, 1999, pp. 105–175.

[7] M. Bardi and P. Soravia, *Approximation of differential games of pursuit-evasion by discrete time games*, in Differential Games: Developments in Modelling and Computation, R. P. Hämäläinen and H. K. Ehtamo, eds., Springer-Verlag, Berlin, New York, 1991, pp. 131–143.

[8] T. Basar and P. Bernhard, *H_∞-Optimal Control and Related Minimax Problems*, Birkhäuser Boston, Boston, 1991.

[9] V. E. Benes, *Existence of optimal stochastic control laws*, SIAM J. Control, 9 (1971), pp. 446–472.

[10] P. Billingsley, *Convergence of Probability Measures*, 2nd ed., Wiley, New York, 1999.

[11] P. Dupuis and H. Ishii, *On Lipschitz continuity of the solution mapping to the Skorokhod problem, with applications*, Stochastics Stochastic Rep., 35 (1991), pp. 31–62.

[12] P. Dupuis and W. M. McEneaney, *Risk-sensitive and robust escape criteria*, SIAM J. Control Optim., 35 (1997), pp. 2021–2049.

[13] R. J. Elliott and N. J. Kalton, *Existence of Value in Differential Games*, Mem. Amer. Math. Soc., 126, AMS, Providence, RI, 1972.

[14] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*, Wiley, New York, 1986.

[15] W. F. Fleming, *Generalized solutions in optimal stochastic control*, in Differential Games and Control Theory III, E. Roxin, P. T. Liu, and R. Sternberg, eds., Marcel Dekker, New York, 1977, pp. 147–165.
[16] W. H. Fleming, *The convergence problem for differential games*, J. Math. Anal. Appl., 3 (1961), pp. 102–116.
[17] W. H. Fleming, *The convergence problem for differential games* II, in Advances in Game Theory, Ann. of Math. Stud. 52, Princeton University Press, Princeton, NJ, 1964, pp. 195–210.
[18] W. H. Fleming and W. M. McEneaney, *Risk-sensitive control on an infinite time horizon*, SIAM J. Control Optim., 33 (1995), pp. 1881–1915.
[19] W. H. Fleming and P. E. Souganidis, *On the existence of value functions for two-player zero-sum differential games*, Indiana Univ. Math. J., 38 (1989), pp. 293–314.
[20] A. Friedman, *Differential Games*, Wiley, New York, 1971.
[21] J. M. Harrison and M. I. Reiman, *Reflected Brownian motion on an orthant*, Ann. Probab., 9 (1981), pp. 302–308.
[22] J. M. Harrison and R. J. Williams, *Brownian models of open queueing networks with homogeneous customer populations*, Stochastics Stochastics Rep., 22 (1987), pp. 77–115.
[23] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1988.
[24] T. G. Kurtz, *Approximation of Population Processes*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 36, SIAM, Philadelphia, 1981.
[25] H. J. Kushner, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Academic Press, New York, 1977.
[26] H. J. Kushner, *Numerical methods for stochastic control problems in continuous time*, SIAM J. Control Optim., 28 (1990), pp. 999–1048.
[27] H. J. Kushner, *Jump-diffusions with controlled jumps: Existence and numerical methods*, J. Math. Anal. Appl., 249 (2000), pp. 179–198.
[28] H. J. Kushner, *Heavy Traffic Analysis of Controlled Queueing and Communication Networks*, Springer-Verlag, Berlin, New York, 2001.
[29] H. J. Kushner, *Numerical Approximations for Stochastic Differential Games: The Ergodic Case*, Report, Applied Math., Brown University, Providence, RI, 2001.
[30] H. J. Kushner and S. G. Chamberlain, *Finite state stochastic games: Existence theorems and computational procedures*, IEEE Trans. Automat. Control, 14 (1969), pp. 248–255.
[31] H. J. Kushner and S. G. Chamberlain, *On stochastic differential games: Sufficient conditions that a given strategy be a saddle point and numerical procedures for the solution of the game*, J. Math. Anal. Appl., 26 (1969), pp. 560–575.
[32] H. J. Kushner and P. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, 2nd ed., Springer-Verlag, Berlin, New York, 2001.
[33] J. Neveu, *Mathematical Foundations of the Calculus of Probability*, Holden-Day, San Francisco, 1965.
[34] O. Pourtallier and M. Tidball, *Approximation of the Value Function for a Class of Differential Games with Target*, Research report 2942, INRIA, Le Chesnay, France, 1996.
[35] O. Pourtallier and B. Tolwinski, *Discretization of Isaac's Equation*, Report, INRIA, Le Chesney, France, 1992.
[36] P. Soravia, H_∞ *control of nonlinear systems: Differential games and viscosity solutions*, SIAM J. Control Optim., 34 (1996), pp. 1071–1097.
[37] T. E. S. Raghavan and J. A. Filar, *Algorithms for stochastic games: A survey*, Z. Oper. Res., 35 (1991), pp. 437–472.
[38] M. I. Reiman and R. J. Williams, *A boundary property of semimartingale reflecting Brownian motions*, Probab. Theory Related Fields, 77 (1988), pp. 87–97.
[39] M. R. Reiman, *Open queueing networks in heavy traffic*, Math. Oper. Res., 9 (1984), pp. 441–458.
[40] M. Tidball, *Undiscounted zero-sum differential games with stopping times*, in New Trends in Dynamic Games and Applications, G. J. Oldser, ed., Birkhäuser Boston, Boston, 1995, pp. 305–322.
[41] M. Tidball and R. L. V. González, *Zero-sum differential games with stopping times: Some results and about its numerical resolution*, in Advances in Dynamic Games and Applications, T. Basar and A. Haurie, eds., Birkhäuser Boston, Boston, 1994, pp. 106–124.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 2, pp. 487–510

CONSISTENT APPROXIMATIONS AND APPROXIMATE FUNCTIONS AND GRADIENTS IN OPTIMAL CONTROL*

OLIVIER PIRONNEAU† AND ELIJAH POLAK‡

Abstract. Because of the unavoidable use of numerical integration methods, such as Runge–Kutta or finite elements, the numerical solution of optimal control problems, with either ODE or PDE dynamics, is governed by a discretization parameter such as the integration mesh-size. Usually, when explicit integration techniques are used, function and derivative values can be computed exactly for the discretized problems. Recently, we have come across some examples where function and derivative values of the explicitly discretized problems had to be approximated by the outcome of N iterations of a solver. Consequently, the discretization of these problems is controlled by *two* parameters: the mesh-size and the number of iterations of the solver.

Referring to [E. Polak, *Optimization: Algorithms and Consistent Approximations*, Springer-Verlag, 1997], we find a theory for solving optimization problems that require discretization. It deals with two situations. In the first, which is referred to as that of *consistent approximations*, it is assumed that an infinite dimensional optimization problem can be suitably approximated by a family of progressively higher dimensional optimization problems. In this case, strategies, in the form of algorithm models, are presented for "diagonalizing" the solution process. In the second situation, it is assumed that numerical solution of the dynamic equations does not result in a family of finite dimensional consistent approximations (e.g., when implicit integration methods are used). For this case, the theory provides models for the implementation of *conceptual* algorithms. Unfortunately, neither of these situations envisions the possibility of two discretization parameters.

In this paper, we present new algorithm models that can be used with two discretization parameters. The first one controls the mesh-size of an explicit integration scheme, and the second one controls the precision with which functions and gradients associated with a fixed mesh-size are computed. The result can be seen as a framework of *quasi-consistent approximations*.

We implemented these new algorithm models using an approximate steepest descent method for the solution of two problems: a two-point boundary value problem in which the discretized linear ODE dynamics are solved approximately using the Gauss–Seidel method and a distributed control problem in which the discretized dynamics are solved using a domain decomposition algorithm which can be implemented on parallelized computers. Our numerical results show that these new algorithms perform quite well and are fairly insensitive to the selection of user-set parameters. Also, they appear to be superior to some alternative, ad hoc schemes.

Key words. optimization, PDEs, acceleration methods

AMS subject classifications. 49M25, 49M37, 65K05, 90C30

PII. S0363012900369599

1. Introduction.

Many classes of optimization problems, such as semi-infinite optimization problems, continuous optimal control problems with ODE dynamics, and optimal control problems with PDE dynamics, cannot be solved numerically without resorting to a discretization strategy. The simplest but least efficient approach is to discretize the problem with desired precision and to solve the discretized problem using finite dimensional optimization algorithms. It is much more efficient to start out with low discretization precision and to increase the precision progressively as the computation proceeds. Referring to [21], we see that there are essentially two distinct approaches to "dynamic" discretization. The first and oldest is that of *algorithm*

*Received by the editors March 21, 2000; accepted for publication (in revised form) January 14, 2002; published electronically June 26, 2002. This work was supported in part by the National Science Foundation under grant ECS-9900985 and by the Institut Universitaire de France.

http://www.siam.org/journals/sicon/41-2/36959.html

†LAN & IUF, University of Paris VI, Paris, France (pironneau@math.jussieu.fr).

‡EECS, University of California at Berkeley, Berkeley, CA 94720 (polak@eecs.berkeley.edu).

implementation; see, e.g., [1, 3, 4, 5, 7, 8, 9, 10, 11, 12, 19, 18, 23]. In this approach, first one develops a *conceptual* algorithm for the original problem and then a *numerical implementation* of this algorithm. In each iteration, the numerical implementation adjusts the precision with which the function and derivative values used by the conceptual algorithm are approximated so as to ensure convergence to a stationary point of the original problem. When far from a solution, the approximate algorithms perform well at low precision, but, as a solution is approached, the demand for increased precision progressively increases.

The second and more recent approach to dynamic discretization uses sequences of finite dimensional approximating problems. It was formalized in [20, 21] in the form of a theory of consistent approximations. Applications to optimal control are described in [24, 26], and a software package for optimal control, based on consistent approximations, can be obtained from [25]. Within this approach, an infinite dimensional problem, $\mathbf{P}$, such as an optimal control problem with either ODE or PDE dynamics, is replaced by an infinite sequence of "nested," epi-converging finite dimensional problems $\{\mathbf{P}_k\}$. Epi-convergence ensures that the *global* optimal solutions of the approximating problems $\{\mathbf{P}_k\}$ converge to *global* optimal solutions of the infinite dimensional problem $\mathbf{P}$. Problem $\mathbf{P}$ is then solved by a recursive scheme which consists of applying a nonlinear programming algorithm to problem $\mathbf{P}_k$ until a test is satisfied, at which point one proceeds to solve problem $\mathbf{P}_{k+1}$, using the last point obtained for $\mathbf{P}_k$ as the initial point for the new calculation. In [21], we find a number of Algorithm Models for organizing such a calculation. These range from simple schemes that ensure the convergence of the subsequence of points, at which discretization has been increased, to a stationary point, to quite complex schemes that ensure the convergence of the entire sequence, constructed by the master algorithm, to a stationary point, with rate of convergence determined by the nonlinear programming algorithms being used. The advantages of the consistent approximations approach over the algorithm implementation approach are that (i) there is a much richer set of possibilities for constructing precision refinement tests, and hence for devising one that enhances computational efficiency, and (ii) one can use unmodified nonlinear programming code libraries as subroutines; see [24, 25].

In this paper, we deal with a situation that has not been considered before: the case where it is either impossible or uneconomical to compute with high precision the values and gradients of functions appearing in the finite dimensional consistent approximating problems $\mathbf{P}_k$, introduced above. We develop a two-tier algorithm, which, in the first tier, constructs an infinite sequence of epi-converging finite dimensional approximating problems $\{\mathbf{P}_k\}$ and, in the second tier, uses an algorithm implementation strategy in solving each $\mathbf{P}_k$. The main task that we had to address was that of constructing efficient tests for dynamically adjusting *two* precision parameters, k and N, where N determines the precision used in the implementation strategy. The end result can be viewed as a *quasi-consistent approximations* approach. As we will see in section 3, our new algorithm performs considerably better than an ad hoc algorithm implementation scheme on the problems tested.

2. Basic definitions and a motivational example. To make this paper reasonably self-contained, we begin with a definition.

DEFINITION 1. *Let S be a normed space, let $\{S_h\}_{h=\alpha}^{0}$ be a sequence of finite dimensional subspaces of S such that $\cup S_h$ is dense in S, and consider the problems*

$$(\mathbf{P}) \qquad \min_{v \in V} f(v), \tag{1}$$

where V is a subset of S and $f : S \to \mathbb{R}$ is continuous, together with the approximating problems

$$(\mathbf{P}_h) \qquad \min_{v_h \in V_h} f_h(v_h), \tag{2}$$

where V_h is a subset of S_h and $f_h : S_h \to \mathbb{R}$ is continuous.

(a) *We say that the problems $\mathbf{P}_h$* epi-converge[1] *to $\mathbf{P}$ if* (i) *for every $v \in V$, there exists a sequence $\{v_h\}$, with $v_h \in V_h$, such that $v_h \to v$, and $\limsup f_h(v_h) \le f(v)$; and* (ii) *for every infinite sequence $\{v_h\}$, such that $v_h \in V_h$ and $v_h \to v$, $v \in V$, and $\liminf f_h(v_h) \ge f(v)$.*

(b) *We say that upper-semicontinuous, nonpositive-valued functions $\theta_h : V_h \to \mathbb{R}$ ($\theta : V \to \mathbb{R}$) are optimality functions for $\mathbf{P}_h$ ($\mathbf{P}$) if they vanish at local minimizers of $\mathbf{P}_h$ ($\mathbf{P}$).*[2]

(c) *We say that the problem-optimality function pairs $\{\mathbf{P}_h, \theta_h\}$ are* consistent approximations *to the problem-optimality function pair $\{\mathbf{P}, \theta\}$ if the $\mathbf{P}_h$ epi-converge to $\mathbf{P}$, and, for every infinite sequence $\{v_h\}$, such that $v_h \in V_h$ and $v_h \to v \in V$, $\limsup \theta_h(v_h) \le \theta(v)$.*[3]

The reason for introducing optimality functions into the definition of consistency of approximation is that it enables us to ensure that not only *global* optimal solutions of the problems $\mathbf{P}_h$ converge to *global* optimal solutions of $\mathbf{P}$, but also *local* optimal solutions converge to either local solutions or stationary points.

The motivation for this work stems from the fact that, while attempting to solve some optimal control problems with distributed dynamics (see Lions [13]), using the consistent approximations framework, we came across a new difficulty, caused by the fact that even the discretized state equation cannot be solved with adequate precision in reasonable time. In such problems, there are *two* precision parameters to control: the mesh-size h, which defines the approximating problem, and the number of iterations N used by a "solver" in solving the discretized state equations. Since the parameter N seriously impacts the behavior of optimization algorithms as well as the total work needed to solve a problem, it is desirable to control the two precision parameters individually.

For example, consider an optimization problem of the form

$$(\mathbf{P}) \qquad \min_{v \in V} f(v), \tag{3}$$

where $V = L^2(0,1)$,

$$f(v) = J(u(v), v) = \int_0^2 |u(v) - u_d|^2 dx, \tag{4}$$

and $u(v)$ is the solution of an equation of the form

$$Cu = Bv \tag{5}$$

[1]The epigraphs of f_h, restricted to V_h, converge to the epigraph of f, restricted to V, in the Painlevé–Kuratowski sense (see [22]).

[2]When optimality functions are properly constructed, their zeros are standard stationary points; for examples, see [21].

[3]Note that this property ensures that the limit point of a converging sequence of approximate stationary points for the $\mathbf{P}_h$ must be a stationary point for $\mathbf{P}$.

such as

$$-u''(x) = v(x)I_{(0,1)} \quad \forall x \in (0,2), \quad u(0) = u(2) = 0, \tag{6}$$

where u_d is given and I_D is the characteristic function of a set D.

This problem can be approximated by a finite dimensional problem of the form

$$(\mathbf{P}_h) \qquad \min_{v_h \in V_h} f_h(v_h), \tag{7}$$

where V_h is the space of piecewise constant functions defined on a mesh for $(0,1)$, $h > 0$ is the mesh-size,

$$f_h(v_h) = J(u_h(v_h), v_h), \tag{8}$$

and $u_h(v_h)$ is the solution of a discretized equation of the form

$$C_h u_h = B_h v_h, \tag{9}$$

arising from a centered finite difference approximation to (6).

It is not difficult to show that the problems $\mathbf{P}_h$ epi-converge to the problem $\mathbf{P}$ as $h \to 0$, and, if $\{v_h\}$ is a sequence of points such that $v_h \in V_h$, $h \to 0$, and $v_h \to v$ as $h \to 0$, then $\mathrm{grad} f_h(v_h) \to \mathrm{grad} f(v)$ as $h \to 0$. These facts show that the pairs $(\mathbf{P}_h, -\|\mathrm{grad} f_h(\cdot)\|)$ form a family of *consistent approximations* for the pair $(\mathbf{P}, -\|\mathrm{grad} f(\cdot)\|)$. Hence any accumulation point of global optimizers of the problems $\mathbf{P}_h$ is a global optimizer of the problem $\mathbf{P}$, and, if $\{v_h\}$ is a sequence of points such that $v_h \in V_h$, $h \to 0$, $v_h \to v$, and $\mathrm{grad} f_h(v_h) \to 0$, then $\mathrm{grad} f(v) = 0$ also.

The fact that the approximating pairs $(\mathbf{P}_h, -\|\mathrm{grad} f_h(\cdot)\|)$ are a family of consistent approximations for the pair $(\mathbf{P}, -\|\mathrm{grad} f(\cdot)\|)$ lays a basis for the solution of $\mathbf{P}$ by the type of algorithm outlined in section 3.3 of [21]. Unfortunately, for small h, C_h is a large sparse matrix, and it is quite possible that all efficient methods for solving the linear system for $u_h(v_h)$ are iterative, and, realistically, only a reasonable number of iterations of an iterative "solver" can be contemplated. Similar facts apply to the computation of $f_h(v_h)$ and $\mathrm{grad} f_h(v_h)$. Thus let $u_{h,N}(v_h)$ denote the result of N iterations of an iterative "solver" applied to the linear system (9), and let $f_{h,N}(v_h)$ denote the associated approximation to $f_h(v_h)$. Similarly, let $\mathrm{grad}_N f_h(v_h)$ denote the result of N iterations of an iterative "solver" applied to the defining equations for $\mathrm{grad} f_h(v_h)$. For instance, if the Gauss–Seidel relaxation algorithm is used to solve (9), then $u_{h,N}(v_h)$ is the Nth iterate of the recursion

$$L_h u^p = B_h v_h - U_h u^{p-1}, \quad p = 1, \ldots, N, \quad u^0 \text{ given}, \tag{10}$$

where L_h is the lower diagonal part of C_h and U_h its upper part.

Thus we see that, in this case, the "discretized" functions $f_h(v_h)$ are not computable exactly, and, for obvious reasons, neither are their gradients. We will see later that this is also the case when domain decomposition is used to solve discretized PDEs. A quick reference to section 3.3 of [21] shows that the Master Algorithm Models outlined there are not applicable to these cases, because there are no standard nonlinear programming algorithms that use approximate function and gradient values, necessitating the development of a new computational scheme, which we will present in the next section. At this point, we drop the subscript h on the "controls" v_h since it will be clear from the context as to which subspace V_h a control v is in.

3. An algorithm model. We will construct a new algorithm model for solving problems of the form **P** that uses only approximations $f_{h,N}(v)$ and $\mathrm{grad}_N f_h(v)$ to the cost function $f_h(v)$ and its gradient $\mathrm{grad} f_h(v)$ by making use of some existing results in [21]. The relevant results are as follows: First, on page 406 in [21], we find the following Algorithm Model 3.3.17 for solving problems of the form **P**, in (4) above, which uses an iteration function $A_h : V_h \to V_h$, $h \in (0, h^{-1}]$, which can be of the form

$$A_h(v) = v - \lambda(v)\mathrm{grad} f_h(v), \tag{11}$$

where $\lambda(v) > 0$ is a step-size.

Algorithm Model 1: Solves problem **P**.

Parameters. $\omega \in (0,1)$, $\sigma > 0$.

Data. $h^{-1} \in \mathbb{R}_+$, and $v^0 \in V_{h^{-1}}$.

Step 0. Set $i = 0$.

Step 1. Compute the largest h^i, of the form $h^{i-1}/2^k$, $k \in \mathcal{N} := \{0, 1, 2, 3, \ldots\}$, and v^{i+1}, such that $h^i \leq h^{i-1}$ and

$$v^{i+1} = A_{h^i}(v^i), \tag{12}$$

and

$$f_{h^i}(v^{i+1}) \; - \; f_{h^i}(v^i) \leq -\sigma(h^i)^\omega. \tag{13}$$

Step 2. Replace i by $i \; + \; 1$, and go to **Step 1**.

Unfortunately, as we have explained in the preceding section, we may not have explicit formulas for computing $f_h(v)$ and $\mathrm{grad} f_h(v)$, and hence we may be forced to use the limited precision results of N iterations of an iterative solver for computing these quantities. Defining, as before, $u_{h,N}(v)$ to be the result of N iterations of a solver applied to the defining equation (9), we define

$$f_{h,N}(v) := J(u_{h,N}(v), v). \tag{14}$$

As we will see later, $\mathrm{grad} f_h(v)$ is usually determined as a solution of an adjoint equation. Hence $\mathrm{grad}_N f_h(v)$ is defined as the result of N iterations of a solver applied to the adjoint equation. This leads to an approximation $A_{h,N}(v)$ to the ideal iteration map $A_h(v)$. For example, the ideal iteration map $A_h(v)$ defined in (11) has to be replaced by

$$A_{h,N}(v) = v - \lambda \mathrm{grad}_N f_h(v), \tag{15}$$

where the step-size λ is determined either by a modified Armijo rule or by one dimensional minimization.

There are obviously any number of ways of making the parameter N a function of h, or even a function of h and v, which results in a new approximation to the cost function

$$\hat{f}_h(v) := f_{h,N(h,v)}(v) \tag{16}$$

and iteration map

$$\hat{A}_h(v) := A_{h,N(h,v)}(v), \tag{17}$$

which, hopefully, can be used within the structure of Algorithm Model 1. One can classify the rules for making N a function of h (or h and v) as *open-loop* or *closed-loop*.

An example of an open-loop rule is to set $N = \text{int}(1/h)$, the integer part of $1/h$. A closed-loop rule can be made more subtle and can be designed to produce as small a parameter N as is compatible with the convergence of the overall solution scheme in the form of an algorithm model. An example of a closed-loop (feedback) rule can be found in Algorithm Model 1.2.36 in [21].

The integer $N^0 > 0$ and an increment integer $K > 0$ are fixed parameters: given h and v, $N(h, v) := N^0 + kK$, where $k \geq 0$ is the smallest integer such that, for $N = N^0 + kK$,

$$f_{h,N}(A_{h,N}(v)) \; - \; f_{h,N}(v) \leq -\frac{\sigma}{N^\omega}, \tag{18}$$

where σ and ω are as in Algorithm Model 1, say.

Proceeding formally from this point on, we assume that, for every $h > 0$, we can construct an iteration map $A_{h,N} : V_h \to V_h$. In our analysis, we will depend on the following assumption.

ASSUMPTION 1. *We will assume the following:*

(i) *The function $f(\cdot)$ is continuous and bounded from below, and, for all $h \in (0, h_{max}]$, the functions $f_h(\cdot)$ are continuous and bounded from below.*

(ii) *For every bounded set $B \subset V$, there exist $\kappa < \infty$, a function $N^* : \mathbb{R}_+ \to \mathcal{N}$ (the set of positive integers), and functions $\varphi : \mathbb{R}_+ \times \mathcal{N} \to \mathbb{R}_+$, $\Delta : \mathbb{R}_+ \to \mathbb{R}_+$ with the properties*

$$\lim_{h \to 0} N^*(h) = \infty, \tag{19}$$

$$\lim_{N \to \infty} \varphi(h, N) = 0 \quad \forall h > 0, \tag{20}$$

$$\lim_{h \to 0} \varphi(h, N_h) = 0 \quad \forall N_h \geq N^*(h), \tag{21}$$

$$\lim_{h \to 0} \Delta(h) = 0, \tag{22}$$

such that, for all $h \in (0, h_{max}]$, $v \in V_h \cap B$,

$$|f_h(v) - f(v)| \leq \kappa \Delta(h), \tag{23}$$

and, for all $h \in (0, h_{max}]$, $N \in \mathcal{N}$, $v \in V_h \cap B$,

$$|f_{h,N}(v) - f_h(v)| \leq \kappa \varphi(h, N). \tag{24}$$

(iii) *For every $v^* \in V$ such that $\text{grad} f(v^*) \neq 0$, there exist $\rho^* > 0$, $\delta^* > 0$, $h^* > 0$, and $N^{**} < \infty$, such that*

$$f_{h,N}(A_{h,N}(v)) - f_{h,N}(v) \leq -\delta^* \;\; \forall v \in V_h \cap B(v^*, \rho^*), \;\; \forall h \leq h^*, \;\; \forall N \geq N^{**}. \tag{25}$$

Algorithm Model 2: Solves problem **P**.
Parameters. $\omega \in (0, 1)$, $\gamma > 0$, $n, K \in \mathcal{N}$, $N^*(\cdot)$, $\Delta(\cdot)$, $\varphi(\cdot, \cdot)$ verifying (19), (20), (21), (22).
Data. $h^0 \in (0, h_{max}]$, $v^0 \in V_{h^0}$.
Begin Outer Loop
Step 0. Set $i = 0$.
Begin Inner Loop (Computes $\hat{f}_h(v^i)$, $\hat{A}_{h^i}(v^i)$), and $\hat{\Delta}(h^i, v^i)$).
Step 1. Set $N^i = N^*(h^i)$.

Step 2. Compute a point $v^* = A_{h^i,N^i}(v^i)$.
Step 3. If $N^i < nN^*(h^i)$ and

$$f_{h^i,N^i}(v^*) - f_{h^i,N^i}(v^i) > -\varphi(h^i, N^i)^\omega, \tag{26}$$

replace N^i by $N^i + K$ and go to **Step 2.**
Else, set

$$N(h^i, v^i) := N^i, \tag{27}$$

$$\hat{f}_{h^i}(v^i) := f_{h^i,N(h^i,v^i)}(v^i), \tag{28}$$

$$\hat{A}_{h^i}(v^i) := A_{h^i,N(h^i,v^i)}(v^i), \tag{29}$$

and

$$\hat{\Delta}(h^i, v^i) = \Delta(h^i) + \varphi(h^i, N(h^i, v^i)). \tag{30}$$

End Inner Loop
Step 4. If

$$\hat{f}_{h^i}(v^*) - \hat{f}_{h^i}(v^i) > -\gamma\hat{\Delta}(h^i, v^i)^\omega, \tag{31}$$

replace the mesh-size h^i by $h^i/2$ and go to **Step 1**.
Else, set

$$v^{i+1} = \hat{A}_{h^i}(v^i), \tag{32}$$

replace i by $i + 1$, and go to **Step 2**.
End Outer Loop

Remark 1.

1. The main function of the test (26) is to increase N over the initial value of $N = N^*(h^i)$ if that is necessary. It gets reset to $N = N^*(h^i)$ whenever h^i is halved.
2. Note that the faster $\varphi(h, N) \to 0$ as $N \to \infty$, the easier it is to satisfy the test (26) at a particular value of N. Thus, when the solver is fast, the precision parameter N will be increased more slowly than when it is slow. A similar argument applies to the reduction of the mesh-size, h^i, on the basis of the test in (31). In the context of dynamics defined by differential equations, the integration mesh-size will be refined much faster when the Euler method is used for integration than when a Runge–Kutta method is used for integration.
3. Note that the test (31) effectively requires the computation of v^{i+2}. In an efficient implementation, this fact must be taken into account so as to avoid unnecessary duplication of computations.
4. It would be mathematically less elegant, but computationally more efficient, to replace the test (31) by

$$f_{h^i,N^i}(v^*) - \hat{f}_{h^i}(v^i) > -\hat{\Delta}(h^i, v^i)^\omega. \tag{33}$$

The proof of convergence of the resulting algorithm would, if anything, be slightly simpler.

We define the problems

$$(\hat{\mathbf{P}}_h) \qquad \inf_{v \in V_h} \hat{f}_h(v), \tag{34}$$

where $\hat{f}_h(v)$ is defined as in (28) for every $h \in (0, h_{max}]$ and $v \in V_h$.

For every $h > 0$, $N \in \mathcal{N}$, and $v \in V_h \cap B$, let $\mathrm{grad}_N f_h(v)$ denote the approximation to $\mathrm{grad} f_h(v)$ obtained by means of N iterations of a solver, and let $N(h, v)$ be defined by (27), in the Inner Loop of Algorithm Model 2. Finally, let

$$\mathrm{grd}\hat{f}_h(v) := \mathrm{grad}_{N(h,v)} f_h(v). \tag{35}$$

Although the functions $-\|\mathrm{grd}\hat{f}_h(v)\|$ are not the negatives of norms of gradients of the functions $\hat{f}_h(v)$ and hence not optimality functions for the problems $\hat{\mathbf{P}}_h$, it will become clear, under the following assumption, that the pairs $\{\hat{\mathbf{P}}_h, -\|\mathrm{grd}\hat{f}_h)\|\}$ are *pseudo*consistent approximations to $\{\mathbf{P}, -\|\mathrm{grad} f\|\}$, in the sense that the problems $\hat{\mathbf{P}}_h$ epi-converge to $\mathbf{P}$, and, for any sequences $\{h^i\}$, $h^i \to 0$, $\{v^i\}$, $v^i \in V_{h^i}$, $v^i \to v \in V$, $\|\mathrm{grd}\hat{f}_{h^i}(v^i)\| \to \|\mathrm{grad} f(v)\|$. We found this observation helpful in constructing the proof of convergence of Algorithm Model 2.

ASSUMPTION 2. *For every $h > 0$, $N \in \mathcal{N}$, and $v \in V_h \cap B$, let $\mathrm{grad}_N f_h(v)$ denote the approximation to $\mathrm{grad} f_h(v)$ obtained by means of N iterations of a solver. We will assume that, for all $h \in (0, h_{max}]$, $v \in V_h \cap B$,*

$$\|\mathrm{grad} f_h(v) - \mathrm{grad} f(v)\| \leq \kappa \Delta(h), \tag{36}$$

and, for all $h \in (0, h_{max}]$, $N \in \mathcal{N}$, $v \in V_h \cap B$,

$$\|\mathrm{grad}_N f_h(v) - \mathrm{grad} f_h(v)\| \leq \kappa \varphi(h, N), \tag{37}$$

where κ, $\Delta(\cdot)$, and $\varphi(\cdot, \cdot)$ are as in Assumption 1.

In order to establish the fact that the pairs $\{\hat{\mathbf{P}}_h, -\|\mathrm{grd}\hat{f}_h)\|\}$ are pseudoconsistent approximations to $\{\mathbf{P}, -\|\mathrm{grad} f\|\}$ and to establish the convergence of Algorithm Model 2, we need the following result.

LEMMA 2. *Let $h^0 > 0$ be as in the Data of Algorithm Model 2, and, for any $h \in (0, h^0]$ and $v \in V_h$, let $N(h, v)$, $\hat{f}_h(v)$, $\hat{A}_h(v)$, and $\hat{\Delta}(h, v)$ be defined as in the Inner Loop of Algorithm Model 2, i.e., by (27), (28), (29), (30), respectively. Then the following hold.*

(a) *For every bounded set $B \subset V$, $\hat{\Delta}(h, v) \to 0$ as $h \to 0$, uniformly in $v \in B$, and there exists a $\kappa < \infty$, such that, for all $h \in (0, h^0]$, $v \in V_h \cap B$,*

$$|\hat{f}_h(v) - f(v)| \leq \kappa \hat{\Delta}(h, v), \tag{38}$$

and

$$\|\mathrm{grd}\hat{f}_h(v) - \mathrm{grad} f(v)\| \leq \kappa \hat{\Delta}(h, v). \tag{39}$$

(b) *For every $\hat{v} \in V$ such that $\mathrm{grad} f(\hat{v}) \neq 0$, there exist $\hat{\rho} > 0$, $\hat{\delta} > 0$, $\hat{h} \in (0, h^0]$ such that*

$$\hat{f}_h(\hat{A}_h(v)) - \hat{f}_h(v) \leq -\hat{\delta} \quad \forall v \in V_h \cap B(\hat{v}, \hat{\rho}), \quad \forall h \leq \hat{h}, \tag{40}$$

where $\hat{A}_h(v)$ is defined by (29).

Proof. (a) It follows from (23) and (24) that, for all $h \in (0, h_{max}]$, $v \in V_h$, and $N \in \mathcal{N}$,

$$\begin{aligned} |f_{h,N}(v) - f(v)| &\leq |f_{h,N}(v) - f_h(v)| + |f_h(v) - f(v)| \\ &\leq \kappa\varphi(h,N) + \kappa\Delta(h). \end{aligned} \tag{41}$$

Hence we have that

$$|\hat{f}_h(v) - f(v)| = |f_{h,N_h(v)}(v) - f(v)| \leq \kappa(\varphi(h, N(h,v)) + \Delta(h)) \equiv \kappa\hat{\Delta}(h,v). \tag{42}$$

Since

$$\hat{\Delta}(h,v) = \varphi(h, N(h,v)) + \Delta(h) \tag{43}$$

and $N(h,v) \geq N^*(h)$, it follows that $\hat{\Delta}(h,v) \to 0$ as $h \to 0$, uniformly in $v \in V \cap B$.

The fact that (39) holds follows by similar arguments, and hence we omit its proof.

(b) Suppose that $v^* \in V$ is such that $\mathrm{grad} f(v^*) \neq 0$. Then, by Assumption 2 (iii), there exist $\rho^* > 0$, $\delta^* > 0$, $h^* \in (0, h^0]$, and $N^{**} < \infty$ such that (25) holds. Let $\hat{h} \in (0, h^*]$ be such that $N^*(h) \geq N^{**}$ for all $h \in (0, \hat{h}]$. Let $h \in (0, \hat{h}]$ and $v \in V_h \cap B(v^*, \rho^*)$ be arbitrary, and let $v' = \hat{A}_h(v)$. Then, because, for any $v \in V_h$, $N(h,v) \geq N^*(h)$ by construction, it follows from (23), (24), and (25) that

$$\begin{aligned} \hat{f}_h(v') - \hat{f}_h(v) &= f_{h,N(h,v')}(v') - f_{h,N(h,v)}(v) \\ &= [f_{h,N(h,v')}(v') - f_{h,N(h,v)}(v')] + [f_{h,N(h,v)}(v') - f_{h,N(h,v)}(v)] \\ &\leq [f_{h,N(h,v')}(v') - f_{h,N(h,v)}(v')] - \delta^* \\ &= [f_{h,N(h,v')}(v') - f(v')] + [f(v') - f_{h,N(h,v)}(v')] - \delta^* \\ &\leq \kappa[\varphi(h, N(h,v')) + \varphi(h, N(h,v)) + 2\Delta(h)] - \delta^* \\ &\leq -\frac{1}{2}\delta^*, \end{aligned} \tag{44}$$

provided that $h^* > 0$ is taken sufficiently small. This completes our proof. ∎

The following corollary follows directly from Lemma 2(a).

Corollary 3. *Suppose that Assumptions* 1 *and* 2 *are satisfied. Then the problems* $\{\hat{\mathbf{P}}_h\}$ *epi-converge to the problem* $\mathbf{P}$*, and, for any sequences* $\{h^i\}$*,* $h_i \to 0$*,* $\{v^i\}$*,* $v^i \in V_{h^i}$*,* $v^i \to v \in V$*,* $\|\mathrm{grd}\hat{f}_{h^i}(v^i)\| \to \|\mathrm{grad} f(v)\|$*; i.e., the pairs* $\{\hat{\mathbf{P}}_h, -\|\mathrm{grd}\hat{f}_h\|\}$ *are pseudoconsistent approximations to* $\{\mathbf{P}, -\|\mathrm{grad} f\|\}$.

Lemma 4. *Suppose that Assumption* 1 *is satisfied.*

(a) *If* $v^i \in V$ *is such that* $\mathrm{grad} f(v^i) \neq 0$*, then there exists an* $h^i > 0$ *such that* (31) *fails.*

(b) *If Algorithm Model* 2 *constructs an infinite sequence* $\{v^i\}_{i=0}^{\infty}$ *that has at least one accumulation point, then* $h^i \to 0$ *as* $i \to \infty$.

Proof. (a) Suppose that $v^i \in V$ is such that $\mathrm{grad} f(v^i) \neq 0$. Then, by Lemma 2(b), there exists an $\hat{h} > 0$ such that, for all $h \leq \hat{h}$,

$$\hat{f}_h(\hat{A}_h(v^i)) - \hat{f}_h(v^i) \leq -\hat{\delta} \leq -\hat{\Delta}(h, v^i)^{\omega}, \tag{45}$$

which shows that there exists an $h^i > 0$ such that (31) fails.

(b) For the sake of contradiction, suppose that the monotone decreasing sequence $\{h^i\}_{i=0}^{\infty}$ is bounded from below by $b > 0$. Then there exists an i_0 such that $h^i = h^{i_0} = h* > 0$ for all $i \geq i_0$. Hence it follows from the fact that (31) fails, at each $i \geq i_0$, that $\hat{f}_{h*}(v^i) \to -\infty$ as $i \to \infty$. It now follows from (38) that $f(v^i) \to -\infty$ as $i \to \infty$. However, by assumption, there exist an infinite subsequence $\{v^{i_j}\}$ and a $v* \in V_{h*}$, such that $v^{i_j} \to v*$ as $j \to \infty$. Since $f(\cdot)$ is continuous, by assumption, we conclude that $f(v^{i_j}) \to f(v*)$ as $j \to \infty$, which is a contradiction and completes our proof. ∎

THEOREM 5. *Suppose that Assumption* 1 *is satisfied.*

(a) *If* $\{v^i\}_{i=0}^{\infty}$ *is a sequence constructed by Algorithm Model* 2 *in solving the problem* $\mathbf{P}$*, then every accumulation point* $v*$ *of* $\{v^i\}_{i=0}^{\infty}$ *satisfies* $\mathrm{grad} f(v*) = 0$.

(b) *If* $f(\cdot)$ *is strictly convex, with bounded level sets, and* $\{v^i\}_{i=0}^{\infty}$ *is a sequence constructed by Algorithm Model* 2 *in solving the problem* $\mathbf{P}$*, then* $\{v^i\}_{i=0}^{\infty}$ *converges to the unique solution of* $\mathbf{P}$.

Proof. (a) Suppose that $\{v^i\}_{i=0}^{\infty}$ is a sequence constructed by Algorithm Model 2 and that $\{v^{i_j}\}_{j=0}^{\infty}$ is a subsequence converging to a point $\hat{v}$ and that $\mathrm{grad} f(\hat{v}) \neq 0$. ∎

Now, by Lemma 4, $h^i \to 0$ as $i \to \infty$, and, by Lemma 2(b), there exist $\hat{\rho} > 0$, $\hat{\delta} > 0$, $\hat{h} > 0$, such that

$$\hat{f}_{h^i}(\hat{A}_h(v^i)) - \hat{f}_{h^i}(v^i) \leq -\hat{\delta} \quad \forall v^i \in B(\hat{v}, \hat{\rho}), \quad \forall h^i \leq \hat{h}. \tag{46}$$

Let i_0 be such that, for all $i_j \geq i_0$, $v^{i_j} \in B(\hat{v}, \hat{\rho})$, and

$$2\kappa\hat{\Delta}(h^{i_j}, v^{i_j}) \leq \frac{1}{2}\hat{\delta}, \tag{47}$$

$$2\kappa\hat{\Delta}(h^{i_j}, v^{i_j})^{1-\omega} \leq \gamma. \tag{48}$$

Finally, let $i_1 \geq i_0$ be such that $h^i \leq \hat{h}$ for all $i \geq i_1$. Then, for the subsequence $\{v^{i_j}\}_{j=0}^{\infty}$, with $i_j \geq i_1$,

$$f(v^{i_j+1}) - f(v^{i_j}) \leq -\hat{\delta} + 2\kappa\hat{\Delta}(h^{i_j}, v^{i_j}) \leq -\frac{1}{2}\hat{\delta}, \tag{49}$$

and, in addition, in view of Lemma 2 and the test (31), for all $i \geq i_1$,

$$\begin{aligned} f(v^{i+1}) - f(v^i) &\leq 2\kappa\hat{\Delta}(h^i, v^i) - \gamma\hat{\Delta}(h^i, v^i)^{\omega} \\ &= -\hat{\Delta}(h^i, v^i)^{\omega}[\gamma - 2\kappa\hat{\Delta}(h^i, v^i)^{1-\omega}] \leq 0. \end{aligned} \tag{50}$$

Hence we see that the sequence $\{f(v^i)\}_{i=i_1}^{\infty}$ is monotone decreasing, and, therefore, because $f(\cdot)$ is continuous, it must converge to $f(\hat{v})$. Since this is contradicted by (48), our proof is complete.

(b) Since a strictly convex function, with bounded level sets, has exactly one stationary point, the desired result follows from (a) and the fact that $\{f(v^i)\}_{i=i_1}^{\infty}$ is monotone decreasing.

Remark 2. The following Algorithm Model differs from Algorithm Model 2 in two respects: first, the integer N is never reset and hence increases monotonically, and second, the test for reducing h is based on the magnitude of the norm of the approximate cost-gradient. As a result, the proof of its convergence is substantially simpler than that for Algorithm Model 2. However, convergence can be established only for the diagonal subsequence $\{v^{i_j}\}_j$ at which h is halved.

Algorithm Model 3: Solves problem **P**.
Parameters. $\omega \in (0,1)$, $\epsilon^0 > 0$, $K \in \mathcal{N}$, $N^*(\cdot)$, $\varphi(\cdot,\cdot)$ verifying (19), (20), (21).
Data. $h^0 > 0$, $v^0 \in V_{h^0}$.
Begin Outer Loop
Step 0. Set $i = 0$, $j = 0$, $N^0 = N^*(h^0)$.
Begin Inner Loop
Step 1. Compute a point $v^* = A_{h^i,N^i}(v^i)$.
Step 2. If

$$f_{h^i,N^i}(v^*) - f_{h^i,N^i}(v^i) > -\varphi(h^i,N^i)^\omega, \tag{51}$$

replace N^i by $N^i + K$, and go to **Step 1**.
Else, set $v^{i+1} = v^*$, and go to **Step 3**.
End Inner Loop
Step 3. If

$$\|\mathrm{grad}_{N^i} f_{h^i}(v^{i+1})\| \le \epsilon^i \text{ and } N^i \ge N^*(h^i), \tag{52}$$

set $v_*^{j+1} = v^{i+1}$, $N_*^{j+1} = N^i$, $h_*^{j+1} = h^i$, replace j by $j+1$, h^i by $h^i/2$, ϵ^i by $\epsilon^i/2$, i by $i+1$, and go to **Step 1**.
Else, replace i by $i+1$, and go to **Step 1**.
End Outer Loop

THEOREM 6. *Suppose that Assumptions* 1 *and* 2 *are satisfied and that* $\{v_*^j\}$ *is a sequence constructed by Algorithm Model* 3 *in solving the problem* **P**.

(a) *If* $\{v_*^j\}$ *is finite, then the sequence* $\{v^i\}_{i=0}^\infty$ *has no accumulation points.*

(b) *If* $\{v_*^j\}$ *is infinite, then every accumulation point* $v*$ *of* $\{v_*^j\}_{j=0}^\infty$ *satisfies* $\mathrm{grad} f(v*) = 0$.

(c) *If* $f(\cdot)$ *is strictly convex, with bounded level sets, and* $\{v_*^j\}_{j=0}^\infty$ *is a bounded sequence constructed by Algorithm Model* 3 *in solving the problem* **P**, *then it converges to the unique solution of* **P**.

Proof. (a) Suppose that the sequence $\{v_*^j\}$ is finite and that the sequence $\{v^i\}_{i=0}^\infty$ has an accumulation point $v*$. Then there exist an i_0, an $h* > 0$, and an $\epsilon* > 0$, such that, for all $i \ge i_0$, $h^i = h*$, $\epsilon^i = \epsilon*$, and $\|\mathrm{grad}_{N^i} f_{h^i}(v^i)\| > \epsilon*$. But, in this case, for $i \ge i_0$, the Inner Loop of Algorithm Model 2 is recognized as being of the form of Master Algorithm Model 1.2.36 in [21]. It now follows from Theorem 1.2.37 in [21] that $N^i \to \infty$ as $i \to \infty$ and that $\mathrm{grad} f_{h*}(v*) = 0$. It now follows from (20) in Assumption 1 and from Assumption 2 that, for some infinite subsequence $\{v^{i_j}\}$, $\mathrm{grad}_{N^{i_j}} f_{h*}(v^{i_j}) \to \mathrm{grad} f_{h*}(v*) = 0$, which shows that (52) cannot be violated an infinite number of times, which is a contradiction.

(b) When the sequence $\{v_*^j\}$ is infinite, it follows directly from Assumptions 1 and 2 and the test (52) that, if $v*$ is an accumulation point of $\{v_*^j\}$, then $\mathrm{grad} f(v*) = 0$.

(c) When the function $f(\cdot)$ is strictly convex, with bounded level sets, it has a unique minimizer $v*$, which is the only point in V satisfying $\mathrm{grad} f(v*) = 0$. Hence the desired result follows from (b). ∎

4. A two-point boundary value control problem. Consider again the two-point boundary value control problem first stated in section 2:

$$(\mathbf{P_1}) \qquad \left| \begin{array}{l} \min\limits_{v \in L^2(0,1)} f(v) := J(u(v)) := \displaystyle\int_0^2 |u - u_d|^2 dx \quad \text{subject to} \\ \qquad\qquad -u''(x) = v(x) I_{(0,1)} \quad \forall x \in (0,2), \quad u(0) = u(2) = 0. \end{array} \right. \tag{53}$$

The gradient of $f(\cdot)$ with respect to v can be expressed in terms of p, the solution of the adjoint equation

$$-p'' = 2(u - u_d), \quad p(0) = p(2) = 0. \tag{54}$$

Thus

$$\delta f = 2\int_0^2 (u - u_d)\delta u = -\int_0^2 p''\delta u = -\int_0^2 p\delta u'' = \int_0^1 p\delta v, \tag{55}$$

which shows that $\mathrm{grad} f(v) = p$ on (0,1).

To approximate the problem $\mathbf{P}_1$, we use a finite difference method with uniform mesh of size $h = 1/M$ to solve the differential equation (54). This results in the approximating problems

$$(\mathbf{P}_{1h}) \quad \left| \begin{array}{c} \min_{v\in V_h} f_h(v) := \sum_1^{2M-1} |u_j - u_d(jh)|^2 \quad \text{subject to} \\ -\frac{1}{h^2}(u_{j+1} - 2u_j + u_{j-1}) = v_j I_{j\le M}, \quad j = 1,\dots,2M-1, \\ u_0 = u_{2M} = 0, \end{array} \right. \tag{56}$$

where V_h is the set of piecewise constant functions on the intervals $(jh,(j+1)h]$, $j = 1,\dots,M$. Note that the coefficients u_j define a piecewise constant function $u(\cdot)$ on $[0,2]$.

As in the continuous case

$$\delta f = \sum_1^{2M-1} 2(u_j - u_d(jh))\delta u_j \tag{57}$$

and if

$$-\frac{p_{j+1} - 2p_j + p_{j-1}}{h^2} = 2(u_j - u_d(jh)), \quad j = 1,\dots,2M-1, \quad p_0 = p_{2M} = 0, \tag{58}$$

then

$$\begin{aligned} \sum_1^{2M-1} 2(u_j - u_d(jh))\delta u_j &= -\sum_1^{2M-1} \frac{p_{j+1} - 2p_j + p_{j-1}}{h^2}\delta u_j \\ &= -\sum_1^{2M-1} \frac{\delta u_{j+1} - 2\delta u_j + \delta u_{j-1}}{h^2} p_j. \end{aligned} \tag{59}$$

Therefore,

$$\delta f = \sum_1^M p_j \delta v_j, \tag{60}$$

and hence the gradient $\mathrm{grad} f_h(v)$ is the piecewise constant function $p_h(\cdot)$, on $(h, 1+h]$, defined by the coefficients $p_0, p_1, \dots, p_M$.

To illustrate the theory, the difference equations in (56), (58) will be solved by the Gauss–Seidel method, and the optimization problem will be solved by the method of steepest descent.

4.1. Verification of the hypotheses. Algorithm Model 2 depends on Assumption 1 to be satisfied and, in particular, on the existence of three appropriate functions $\varphi(h, N)$, $N^*(h)$, and $\Delta(h)$ and of an appropriate iteration map $A_{h,N}(\cdot)$.

We begin by showing that parts (i) and (ii) of Assumption 1 are satisfied. When the ODE for u is multiplied by u and integrated in x, we obtain, after integrating by parts, that

$$-\int_0^2 (uu'') = \int_0^1 uv = \int_0^2 (u'^2). \tag{61}$$

Applying the Schwarz inequality to the middle integral leads to $\|u'\|_0 \leq \|v\|_0$. It follows from the Poincaré inequality that $\|u\|_0 \leq C\|u'\|_0$ for some $C < \infty$. Hence we conclude that u is continuous with respect to v in L^2:

$$\|u\|_0 \leq C\|v\|_0. \tag{62}$$

Now the function $u \to J(u)$ is obviously continuous in u, and hence $f(\cdot)$ is continuous in v.

Using similar arguments, we find that p is continuous in v, and hence $\operatorname{grad} f(\cdot)$ exists and is continuous.

For the discrete problem, we note that $(u_1, u_2, \ldots, u_{2M-1})^T$ is the solution of a linear system with right-hand side $(v_1, \ldots, v_M, 0.., 0)^T$, and the matrix of the linear system is tridiagonal with $2/h^2$ on the main diagonal and $-1/h^2$ on the diagonals below and above the main one. This is a positive definite matrix, and hence u is continuous with respect to v. Similarly, p is continuous with respect to u and with respect to v by transitivity.

Next, it follows from the error analysis for the finite difference scheme that, for some $C < \infty$,

$$\|u_h - u\|_0 < Ch^2, \quad |J_h(u, v) - J(u, v)| < Ch^2, \tag{63}$$

which implies that

$$|f_h(v_h) - f(v)| < Ch^2. \tag{64}$$

Now the Gauss–Seidel algorithm is linearly convergent, but the constant of convergence is proportional to the condition number of the linear system. In particular, for some $C, c < \infty$,

$$\|u_{h,N} - u_h\| \leq C(1 - ch^2)^N \quad \forall N \in \mathcal{N}. \tag{65}$$

By inspection, a bound function φ is $\varphi(h, N) = C'(1 - ch^2)^N$ with any $C' \geq C$. However, it contains an unknown constant. We have the choice of either guessing this constant or replacing the function φ with a conservative estimate, such as $\varphi(h, N) = (1 - h^{2+\epsilon})^N$, with $\epsilon < 1$, small; i.e., we replace c with h^ϵ. In either event, and to satisfy the hypothesis, we may take

$$N^*(h) = \frac{C}{h^{2+2\epsilon}}, \tag{66}$$

with C a generic constant. Indeed,

$$(1 - h^{2+\epsilon})^{\frac{C}{h^{2+2\epsilon}}} = \exp \frac{C \log(1 - h^{2+\epsilon})}{h^{2+2\epsilon}} \approx e^{-\frac{C}{h^\epsilon}} \to 0 \text{ as } h \to 0. \tag{67}$$

We have thus shown that parts (i) and (ii) of Assumption 1 are satisfied.

To conclude, we must show that part (iii) of Assumption 1 is satisfied. We will derive the iteration map $A_{h,N}(\cdot)$ from the standard steepest descent algorithm with exact step-size. We recall that, for the problems $\mathbf{P}_{1h}$, this algorithm is defined by the following iteration function:

$$A_h(v) := v - \lambda(v)\text{grad}f_h(v), \tag{68}$$

where

$$\lambda(v) := \arg\min_{\lambda} f_h(v - \lambda\text{grad}f_h(v)). \tag{69}$$

Note that, for our problem, $\lambda(v)$ can be computed exactly because $f_h(v-\lambda\text{grad}f_h(v))$ is a quadratic function of λ.

Next, we define $A_{h,N}$ as follows:

$$A_{h,N}(v) := v - \lambda_N(v)\text{grad}_N f_h(v), \tag{70}$$

with

$$\lambda_N(v) := \arg\min_{\lambda} f_{h,N}(v - \lambda\text{grad}_N f_h(v)), \tag{71}$$

where $f_{h,N}(v)$ and $\text{grad}_N f_h(v)$ are computed using N iterations of the Gauss–Seidel algorithm on the difference equation in (56) and the adjoint equation (58), respectively.

Now, it follows from the properties of the method of steepest descent that, given any $v^* \in V = L^2(0,1)$ such that $\text{grad}f(v^*) \neq 0$, there exist $\rho^* > 0$, $\delta^* > 0$, λ^*, and $h^* > 0$, such that, for all $v \in V \cap B(v^*, \rho)$, (i) $\text{grad}f_h(v) \neq 0$ and (ii)

$$f(v - \lambda(v)\text{grad}f(v)) - f(v) \leq f(v - \lambda^*\text{grad}f(v)) - f(v) \leq -\delta^*, \tag{72}$$

where $\lambda(v)$ is the exact step-size computed by the Steepest Descent Algorithm. It now follows from (20), (21), (23), (24) that there exist an $h^* > 0$ and an $N^{**} < \infty$, such that, for all $h \leq h^*$, $N \geq N^{**}$, and $v \in V_h \cap B(v^*, \rho)$,

$$\begin{aligned} f_{h,N}(v - \lambda_N(v)\text{grad}_N f_h(v)) - f_{h,N}(v) &\leq f_{h,N}(v - \lambda(v)\text{grad}_N f_h(v)) \\ &\quad - f_{h,N}(v) \leq -\delta^*/2, \end{aligned} \tag{73}$$

which shows that part (iii) of Assumption 1 is satisfied.

4.2. Implementation of Algorithm Model 2. For any positive real number α, we define $\texttt{ceil}[\alpha]$ to be the smallest integer larger than α. Then, making use of the maps defined in the preceding subsection, we now obtain from Algorithm Model 2 the following.

Algorithm 1.

Data. $C_1 > 0$, $C_2 > 0$, $C_3 > 0$, $\epsilon > 0$, $h > 0$, $K \in \mathcal{N}$, $v_0 \in V_h$.

Step 0. Set $i = 0$.

Step 1. Set $M = 1/h$, $N = \texttt{ceil}(\frac{C_1}{h^{2+2\epsilon}})$.

Step 2. Compute $\{u_j^i\}$ using N Gauss–Seidel iterations.

Step 3. Compute $\{p_j^i\}$ using N Gauss–Seidel iterations.

Step 4. Compute $\lambda^i = \text{argmin}_\lambda f_{h,N}(v^i - \lambda p^i)$ using N Gauss–Seidel iterations.

Step 5. Set $v_j^{i+1} = v_j^i - \lambda^i p_j^i, \quad j = 1, \ldots, M.$

Step 6. **If** $f_{h,N}(v^{i+1}) - f_{h,N}(v^i) > -C_2(1 - C_4h^{2+\epsilon})^N$, replace N by $N+K$, and go to **Step 2**.
Else, go to **Step 7**.

Step 7. **If** $f_{h,N}(v^{i+1}) - f_{h,N}(v^i) > -C_3[h^2 + (1 - C_4h^{2+\epsilon})^N]$, replace h by $h/2$, and go to **Step 1**.
Else, replace i by $i+1$, and go to **Step 2**.

PROPOSITION 7. *There exists C^* such that, if $C_2 \geq C^*, C_3 \leq C^*$, and $\{v^i\}$ is a sequence of piecewise constant functions constructed by Algorithm* 1*, then $\{v^i\}$ converges to the solution of $\mathbf{P}_1$ as $i \to \infty$.*

Proof. This is Algorithm Model 2 with N^*,φ, and Δ multiplied by constants, and so, if the last two are smaller than the theoretical constants in (64), (65), then the method converges in the sense that any accumulation point satisfies the optimality conditions of the problem. Since the control problem is linear-quadratic, any solution of the optimality conditions is the solution of the problem. ∎

4.3. Numerical results. Problem (53) was solved with $u_d = \sin(\pi x)$ starting from $v = 0$, first using the standard steepest descent method, with a fixed mesh of 256 points, and solving the linear system using 500 Gauss–Seidel iterations. Then it was solved using Algorithm 1 (see Figures 1 and 2).

In the second case, the initial mesh had 8 points, and the final mesh had 512. The algorithm constants were

$$C_1 = 1, \quad C_2 = 0.1, \quad C_3 = 2\ 10^{-4}, \quad C_4 = 5, \quad \epsilon = 0.1, \quad K = 20.$$

These constants were chosen using trial and error to obtain good efficiency, but, as shown in Figure 3, the results are not very sensitive to these choices as long as the order of magnitude is right.

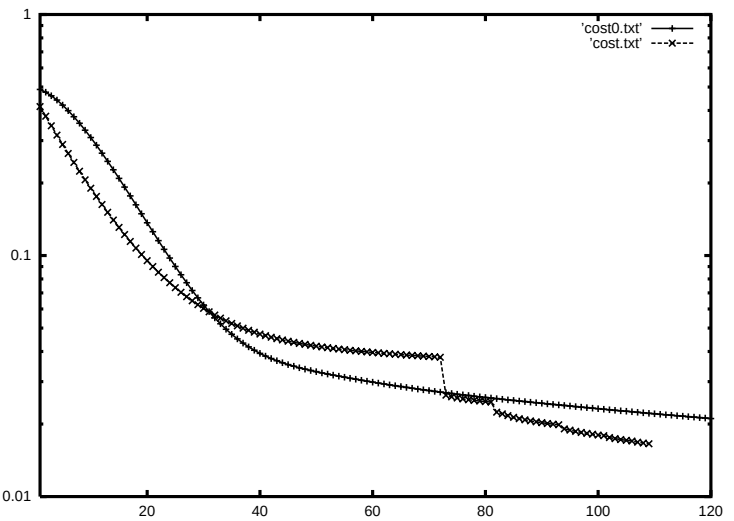


FIG. 1. *Cost function versus iteration number with and without mesh adaptation for problem P_1. The smooth curve (–+–) corresponds to standard steepest descent on the finest mesh with* 500 *Gauss–Seidel iterations for the linear systems. The broken curve (–×–) shows cost function decrease with Algorithm* 1*. Although the two curves are similar, there is an order of magnitude decrease in computing time using Algorithm* 1*.*

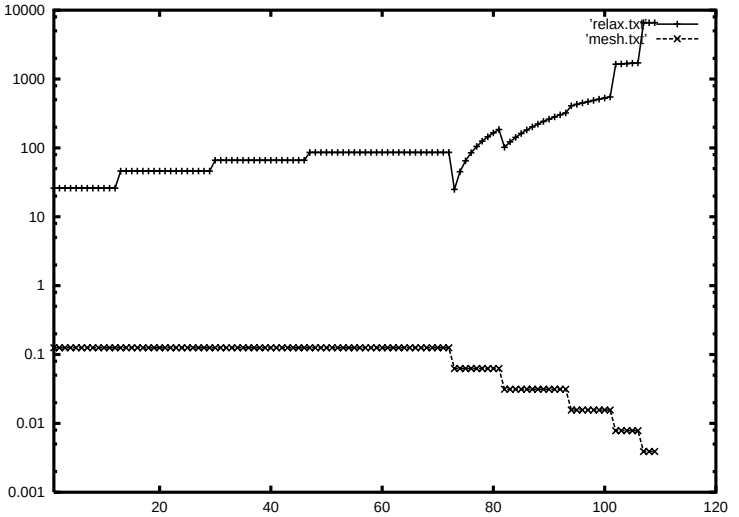


FIG. 2. *Mesh-size and number of Gauss–Seidel iterations versus iteration number for problem* P_1 *solved using Algorithm* 1. *The top curve is the history of Gauss–Seidel iterations count, and the lower one is the history of the mesh-size.*

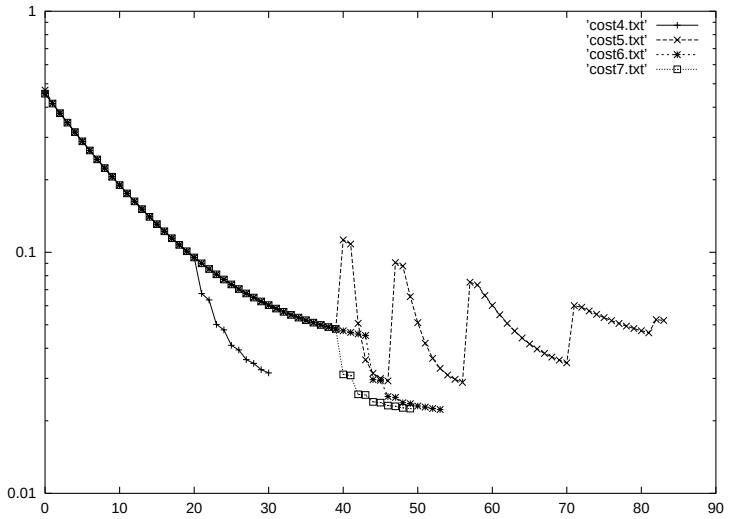


FIG. 3. *Cost function for other values of* C_i. *The behavior of Algorithm* 1 *on problem* P_1 *is shown for different values of the implementation constants:* C_1 *divided by* 10 ($-+-$), C_2 *divided by* 10 ($-\times-$), C_3 *divided by* 10 ($-*-$), *and* C_4 *divided by* 10 ($-\diamond-$).

Figure 1 shows the convergence history of the cost function for both tests. Figure 2 shows the history of the number of Gauss–Seidel iterations for the second case.

Figure 3 shows the behavior resulting from replacing each constant C_i by $C_i/10$, $i = 1, 2, 3, 4$ (one at a time).

In Figure 4, Algorithm 1 is compared with three other methods:

- *Implementation of steepest descent.* Use any approximation of the continuous gradient combined with a mesh refinement/Gauss–Seidel iteration increase based on when the approximate gradient is no longer an approximate feasible descent direction; for instance,

  (74)
  $$\text{if } f_{h,N}(v^{i+1}) - f_{h,N}(v^i) > -\varepsilon, \text{ then replace } (N, h, \varepsilon) \text{ by } (2N, h/2, \varepsilon/2).$$

  However, in Figure 4, we see that such a strategy is too crude, especially because the relationship between the change of h and N is linear.
- *Open-loop mesh refinement*: In Algorithm 1, we replace Step 7 with the rule that the mesh should be refined every 20 steps.
- *Implementation of Algorithm Model* 3: Finally, we compare Algorithm 1 with Algorithm Model 3, where mesh refinement is based on the norm of the gradient rather than on cost function decrease.

We have also tested a number of other parameter values and other functions $N^*(h, v)$, $\varphi(h, N)$ in Algorithm 1. Most of the time, similar computational behavior to that described here was obtained. However, sometimes the mesh was refined too quickly, and sometimes the number of Gauss–Seidel iterations became too large too soon, etc.

It is clear that the common strategy of simply discretizing a problem is easily modified to conform to Algorithm Model 2 or 3. The computing time will always be smaller when dynamic precision adjustment is used, and, for reasonable values of

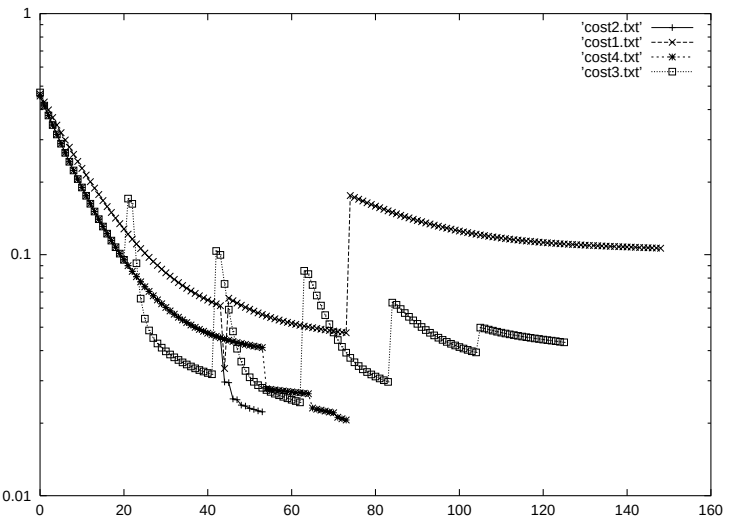


FIG. 4. *Comparison with other methods. Algorithm* 1 (−+−) *is compared to three other methods:* (i) *Implementation of Algorithm Model* 3 (−∗−), (ii) *Heuristic precision refinement* (75) (−×−), (iii) *Algorithm* 1 *with the test on mesh refinement replaced by a division by two every* 20 *iterations* (−◇−).

algorithm parameters, the computing time will be of an order of magnitude smaller than in the case where dynamic precision adjustment is not used.

5. A distributed control problem. Let S be a given subset of the boundary Γ of an open bounded subset Ω of $\mathbb{R}^d$; let ξ be a given function on S (added to this academic problem only to make it nontrivial (see (81), (91))), and consider the boundary control problem

$$(\mathbf{P}_2) \qquad \left| \begin{array}{l} \min\limits_{v\in L^2(S)} f(v) = \dfrac{1}{2}\displaystyle\int_\Omega [(u-u_d)^2 + |\nabla(u-u_d)|^2] \quad \text{subject to} \\ \qquad u - \Delta u = 0 \text{ in } \Omega, \quad \dfrac{\partial u}{\partial n}|_S = \xi v, \quad u_{\Gamma - S} = u_d. \end{array} \right. \tag{75}$$

The gradient of $f(\cdot)$ can be obtained by making use of the fact that

$$\delta f = \int_\Omega ((u-u_d)\delta u + \nabla(u-u_d)\cdot\nabla\delta u) + o(|v|) = \int_S \xi(u-u_d)\delta v, \tag{76}$$

which follows from the fact that the PDE in variational form is: find $u \in H^1(\Omega)$, the Sobolev space of order 1, such that $u - u_d \in H^1_{0_{\Gamma-S}}(\Omega) := \{u \in H^1(\Omega) \ : \ v|_{\Gamma-S} = 0\}$ and

$$\int_\Omega (uw + \nabla u\cdot\nabla w) = \int_S \xi v w \quad \forall w \in H^1_{0_{\Gamma-S}}(\Omega). \tag{77}$$

So, by inspection of (76), we see that the gradient of $f(\cdot)$ with respect to the $L^2(S)$ norm is

$$\operatorname{grad}_v f(v) = \xi(u-u_d)|_S. \tag{78}$$

To approximate the problem (75), we used a finite element method with $u \in V_h$, continuous, and piecewise linear on the triangles of a triangulation of Ω. This results in the discretized, finite dimensional optimization problem

$$(\mathbf{P}_{2h}) \qquad \left| \begin{array}{l} \min\limits_{v\in V_h} f_h(v) = \dfrac{1}{2}\displaystyle\int_\Omega [(u-u_{dh})^2 + |\nabla(u-u_{dh})|^2] \quad \text{subject to} \\ \qquad \displaystyle\int_\Omega (uw + \nabla u\cdot\nabla w) = \int_S \xi v w \quad \forall w \in V_h^0, \end{array} \right. \tag{79}$$

where V_h^0 is the approximation of $H^1_{0_{\Gamma-S}}(\Omega)$ consisting of continuous piecewise linear functions on the triangulation, which are zero on $\Gamma - S$.

The gradient of the discrete cost function $f_h(\cdot)$ can be obtained exactly as in the continuous case

$$\delta f_h = \int_S \xi(u-u_{dh})\delta v. \tag{80}$$

Therefore,

$$\operatorname{grad}_v f_h(v) = \mathcal{P}_h(\xi(u-u_{dh}))|_S, \tag{81}$$

where $\mathcal{P}_h$ is the projection operator from $L^2(S)$ into $V_h \cap L^2(S)$.

Strictly speaking, (81) holds only if Ω is a polygonal domain, but this is a standard technical problem with the finite element method which can be dealt with easily.

5.1. The Schwarz algorithm. Now, for some reason (parallel computing, for instance), suppose that we want to solve the discrete PDE (i.e., its equivalent sparse linear system) by a domain decomposition method (see [14, 15, 16, 17]).

Let $\Omega = \Omega_1 \cup \Omega_2$ with $\Omega_1 \cap \Omega_2 \neq \emptyset$; let $\Gamma = \partial\Omega$, and let $\Gamma_{ij} = \partial\Omega_i \cap \Omega_j$. To compute u, the solution of

$$u - \Delta u = f \text{ in } \Omega, \quad u|_{\Gamma - S} = u_\Gamma, \quad \frac{\partial u}{\partial n}|_S = \xi v, \tag{82}$$

the multiplicative Schwarz algorithm starts from a guess u_1^0, u_2^0 and computes $u_j = u|_{\Omega_j}$ as the limit when $n \to \infty$, of the sequence $u_j^n, j = 1, 2$, defined by

$$\left.\begin{aligned} & u_1^n - \Delta u_1^n = f \text{ in } \Omega_1, \\ & u_1^n|_{\Gamma \cap \bar{\Omega}_1 - S} = u_\Gamma, \quad u_1^n|_{\Gamma_{12}} = u_2^{n-1}, \quad \frac{\partial u_1^n}{\partial n}|_{S \cap \bar{\Omega}_1} = \xi v, \\ & u_2^n - \Delta u_2^n = f \text{ in } \Omega_2, \\ & u_2^n|_{\Gamma \cap \bar{\Omega}_2 - S} = u_\Gamma, \quad u_2^n|_{\Gamma_{21}} = u_1^{n-1}, \quad \frac{\partial u_2^n}{\partial n}|_{S \cap \bar{\Omega}_2} = \xi v, \end{aligned}\right\} \tag{83}$$

and

$$u^n = u_1^n \text{ on } \Omega_1 \cap \Omega_2^c,\ u^n = u_2^n \text{ on } \Omega_2 \cap \Omega_1^c,\ u^n = \frac{1}{2}(u_1^n + u_2^n) \text{ on } \Omega_1 \cap \Omega_2. \tag{84}$$

5.2. The doubly discretized problem. The introduction of the Schwarz algorithm leads to a doubly discretized problem, as follows. Let $\mathcal{T}_h$ be a triangulation of Ω of average edge size h such that, by removing triangles, we also obtain proper triangulations $\{\mathcal{T}_{jh}\}_{j=1,2}$ of Ω_1 and Ω_2.

Let V_{1h} and V_{2h} be the finite element spaces of continuous piecewise affine functions on $\{\mathcal{T}_{jh}\}_{j=1,2}$. Let V_{jh}^0 be the subspaces of continuous piecewise linear functions which are zero on the Dirichlet boundaries Γ_{ij}, $j = 1, 2$.

Then the doubly discretized problem is

(85)

$$(\mathbf{P}_{2h,N}) \left| \begin{array}{c} \min_{v \in V_h} f_{h,N}(v) = \|u^N - u_d\|_\Omega^2 : \quad u_j^0 = 0, \quad n = 1, \ldots, N, \\ u_j^n \in V_{jh} \ : \ u_j^n|_{\Gamma_{ij}} = u_i^{n-1}, \quad \displaystyle\int_{\Omega_j} [u_j^n w + \nabla u_j^n \nabla w] = \int_S \xi v w \quad \forall w \in V_{jh}^0, \\ j = 1, 2, \quad i = (j+1) \bmod 2, \\ u^N = u_1^N \text{ on } \Omega_1 \cap \Omega_2^c,\ u^N = u_2^N \text{ on } \Omega_2 \cap \Omega_1^c,\ u^N = \dfrac{1}{2}(u_1^N + u_2^N) \text{ on } \Omega_1 \cap \Omega_2. \end{array} \right.$$

So N is the number of Schwarz iterations applied to the discretized PDE in (79).

Consider the mapping from $V_{1h} \times V_{2h}$ onto itself which defines u^n from u^{n-1} by (84) and

$$\int_{\Omega_j} [u_j^n w + \nabla u_j^n \nabla w] = \int_S \xi v w \ \ \forall w \in V_{jh}^0, \quad u_j^n \in V_{jh}, \ \ u_j^n|_{\partial\Omega_{ij}} = u_i^{n-1}, \tag{86}$$

for $j = 1, 2$, $i = j \bmod 2$, $n = 1, \ldots, N$. Let $\{A, B, C\}$ be the finite element matrices associated with this operation. In matrix form, (86) is

$$AU^n = BU^{n-1} + CV, \tag{87}$$

where U denotes the vector of values of u_1 at the vertices of $\mathcal{T}_{1h}$ and of u_2 at the vertices of $\mathcal{T}_{2h}$, and V is the vector of values of v at the vertices of S.

For simplicity, we choose $U^0 = 0$. The doubly discretized problem (85) can now be rewritten as

(88)
$$\min_v \left\{ (U^N - U_d)^T G (U^N - U_d) : \begin{pmatrix} A & 0 & 0 & ... & 0 & 0 \\ -B & A & 0 & ... & 0 & 0 \\ 0 & -B & A & .. & 0 & 0. \\ ... & & & & & ... \\ ... & & & & -B & A \end{pmatrix} \begin{pmatrix} U^1 \\ U^2 \\ U^3 \\ ... \\ U^N \end{pmatrix} = \begin{pmatrix} CV \\ CV \\ CV \\ CV \\ CV \end{pmatrix} \right\},$$

where G is the finite element mass matrix (see Ciarlet [6] for more details).

We can express the exact gradient $\operatorname{grad} f_{h,N}(v)$ of $f_{h,N}(v)$ in terms of the solution of the adjoint equation

(89)
$$\begin{pmatrix} A & -B^T & 0 & ... & 0 & 0 \\ 0 & A & -B^T & ... & 0 & 0 \\ 0 & 0 & A & .. & 0 & 0. \\ ... & & & & ... & -B^T \\ ... & & & & 0 & A \end{pmatrix} \begin{pmatrix} P^1 \\ P^2 \\ P^3 \\ ... \\ P^N \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 2G(U^N - U_d) \end{pmatrix}$$

by making use of the fact that $\delta f_{h,N} = \left(\sum_1^N P^n\right)^T C \delta V$. Thus we see that $\operatorname{grad} f_{h,N}(v) = C^T \sum_1^N P^n$.

The interpretation is that P, like U, is the set of values at vertices of the Schwarz system

$$p^N - \Delta p^N = 2(u^N - u_d), \quad p^{N-1} - \Delta p^{N-1} = 0, \quad p^{N-1}_{\Gamma_{ij}} = p^N. \tag{90}$$

These equations are difficult to implement because, in principle, we must store all intermediate functions generated by the Schwarz algorithm and integrate the system for p^n in reverse order, although here it is not necessary because the problem is linear. Hence we will use approximations to the gradients $\operatorname{grad} f_{h,N}(v)$ defined by

$$\operatorname{grad}_N f_h(v) := \Pi_h(\xi(u_{h,N}(v) - u_d))|_S, \tag{91}$$

where Π_h is the interpolation operator ($\Pi_h g$ is the piecewise linear function which coincides with g at the vertices of S) and $u_{h,N}$ is computed by N iterations of the Schwarz algorithm with the convention that on $\Omega_1 \cap \Omega_2$, $u_{h,N} = \frac{1}{2}(u_{1h,N} + u_{2h,N})$.

5.3. Verification of the hypotheses. We proceed exactly as in the one dimensional case to show that Assumption 1 is satisfied.

(i) Continuity of $f(\cdot)$ with respect to the control is established in Lions [13]. Continuity of $f_h(\cdot)$ with respect to the control is obvious from (89).

(ii) It follows from the finite element error estimates given in [6] that the error analysis (63), (64) holds for this case as well. Hence we can set $\Delta(h) = h^2$.

(iii) The Schwarz algorithm converges linearly with rate $(1 - d/D)$, where d is the diameter of $\Omega_1 \cap \Omega_2$ and D is the diameter of $\Omega_1 \cup \Omega_2$; so, instead of (65), we have the bound

$$\|u_{h,N} - u_h\| \leq C \left(1 - \frac{d}{D}\right)^N \quad \forall N \in \mathcal{N}, \tag{92}$$

for some $C \in (0, \infty)$, which implies that we can set $\varphi(h, N) = (1 - \frac{d}{D})^N$. Note that, in this case, $\varphi(h, N)$ is actually independent of h. In view of this, we can take $N^*(h) = \texttt{ceil}(C/h)$, where $C > 0$ is arbitrary.

(iv) The relation (72) obviously holds for this case as well. To show that the relation (73) also holds, we make use of the facts that (a)

(93)
$$\text{grad} f_h(v) = \mathcal{P}_h(\xi(u_h(v) - u_{dh}))|_S, \quad \text{grad}_N f_h(v) = \Pi_h(\xi(u_{h,N}(v) - u_d))|_S,$$

(b) both $\mathcal{P}_h$ and Π_h tend to the identity operator at the rate $O(h)$ at least, and (c) the bound functions $\Delta(h)$, $\varphi(h, N)$, and $N^*(h)$ have the required properties.

5.4. A numerical example. In this example, Ω_1 is the unit circle centered at the origin, and Ω_2 is the rectangle $(0, 1) \times (0, 1)$ minus the unit triangle with vertices (0,0),(0,1),(1,0) and minus a disk of boundary S. The control boundary is S.

The function which is to be recovered by the optimization process is $u_d = e^{-x\sqrt{2}} \sin(y)$ over the whole domain Ω. The weight on the control has been deliberately chosen to have oscillations $\xi = \sin(30 * (x - 1.15)) + \sin(30 * (y - 0.5))$. We have used an automatic mesh generator controlled by a parameter n, the number of vertices on the boundaries, so, for practical reasons, we initialized $h = 1/(8n)$. The number of Schwarz iterations was initialized at 1.

The tests (26) (in Algorithm Model 2) and (51) (in Algorithm Model 3) for increasing the number of Schwarz iterations were determined by setting $\varphi(h, N) = (0.8)^N$ and $C_1 = 0.1$. The mesh refinement test (31) in Algorithm Model 2 was implemented with the right-hand side set to $-0.001[10^{-4}h^2 + (0.8)^{1/(8h)}]$, which corresponds to $N^*(h) = 0.1/(8h)$, $\varphi(h, N) = 0.8^N$, $\Delta(h) = h^2$. Naturally other choices of coefficients and bound functions are possible.

The mesh refinement test (52) in Algorithm Model 3 was implemented by setting $\epsilon(n) = 10^{-n}$, where $n = 1/8h$.

We have used the code freefem+ [2], which is a matlab-like environment for PDEs developed for the purpose of testing parallel algorithms, among other things.

In Figure 5, we plot the values of the cost function $f(\cdot)$ versus the iteration number for two cases. The first corresponds to optimization using a fixed mesh and a fixed number of Schwarz iterations, i.e., without adaptive precision refinement (curve "criter0"), and the second one was obtained using adaptive refinement based either on the norm of the gradient (case (i), curve "criter1"), or on the decrease of the cost function (case (ii), curve "criter").

Figure 6 shows the number of Schwarz iterations N and the mesh parameter n versus the iteration number for case (i). After 30 iterations, the gradient is 10^{-6} times its initial value, while, without mesh refinement, it is only 10^{-2} times its initial value (multilevel effect).

The solution and the precision are shown in Figures 7 and 8, respectively.

6. Conclusion. We have developed algorithm models based on the consistent approximations approach for solving infinite dimensional problems with two independent precision parameters. We have applied it to two optimal control problems with ODE and PDE dynamics each having two precision parameters: the step-size and an iteration loop count in the solvers. Our numerical results show that the algorithms are effective. The numerical study was done using the method of steepest descent, but the models and the proofs are general and can probably be used with Newton's

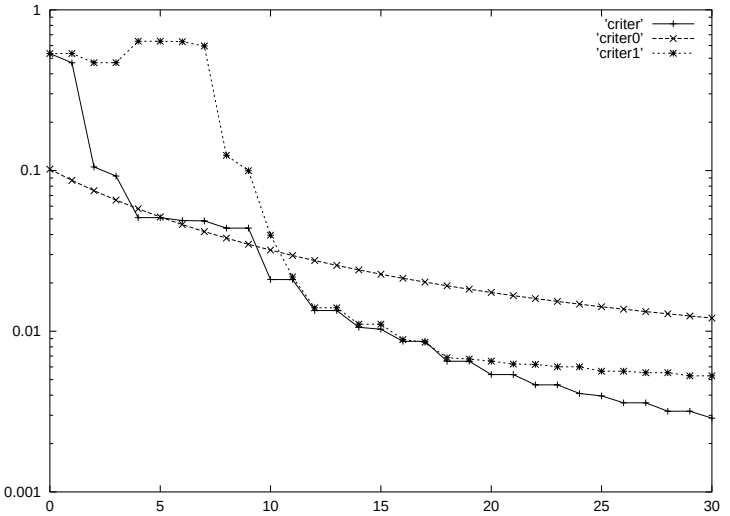


FIG. 5. *Cost function versus iterations for without* $(-\times-)$ *and with mesh adaptation with Algorithm Model* 1 $(-+-)$ *and* 3 $(-*-)$ *for problem* P_2. *Here again, although the general behavior is similar, there is an order of magnitude decrease in the computing time with mesh adaptation because the first fifteen iterations are essentially instantaneous in this case.*

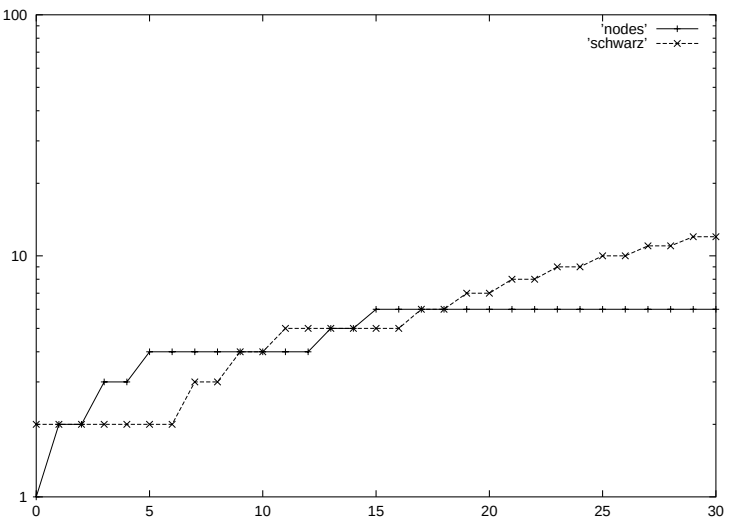


FIG. 6. *Changes in the number of mesh points* $(-+-)$ *and Schwarz iterations* $(-\times-)$ *versus iteration count when adaptation is used to solve problem* P_2 *by Algorithm Model* 3.

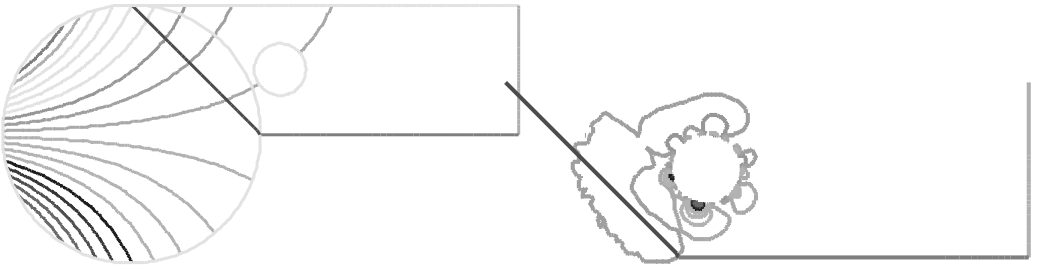


FIG. 7. *Left: Level lines of the computed solution* u (20 *lines equally spaced ranging from* -1.77 *to* 1.77). *Right: Level lines of the error* $u - u_d$ (20 *lines equally spaced ranging from* -0.0199 *to* 0.0150).

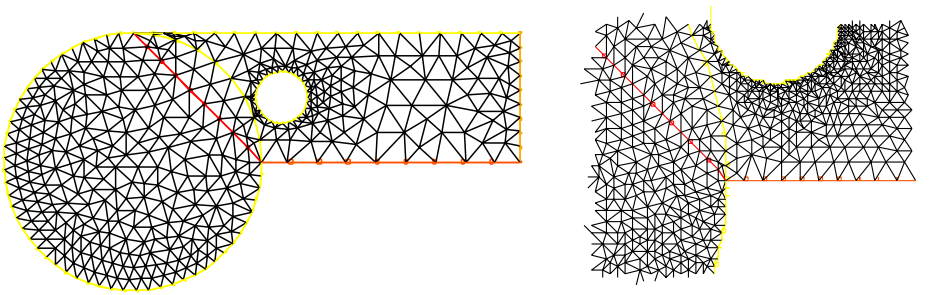


FIG. 8. *Shown here are the* 2*nd finite element mesh on the left and a zoom of part of the* 7*th finite element mesh (the last is the* 9*th) on the right. Both are generated automatically by a Delaunay–Voronoi mesh generator from a uniform distribution of points on the boundaries.*

method, conjugate gradient methods, etc. Other applications are under way, in cooperation with G. Lemarchand and Y. Achdou, using the conjugate gradient method for two problems:

- Optimal shape design of wing profiles with mesh adaptation and for which the iteration number for the solvers of the Euler or compressible Navier–Stokes equations was determined in the same way as the number of Schwarz iterations was determined in this paper.
- Volatility smile in financial modeling of option pricing, where one parameter function (the volatility) in the Black–Scholes PDE is adjusted by least square/optimal control to fit the market observations; there, incomplete gradients are due to short cycles in the solvers, and mesh adaptation is used.

In these two important applications, we have reduced the computing time by an order of magnitude, as compared to the times with fixed discretization parameters, and we had no stability problems even though we had much less information as to the rate of convergence of the solvers to guide the choice of the various parameters of our algorithm models.

REFERENCES

[1] R. BECKER, H. KAPP, AND R. RANNACHER, *Adaptive finite element methods for optimal control of partial differential equations: Basic concept*, SIAM J. Control Optim., 39 (2000), pp. 113–132.

[2] D. Bernardi, F. Hecht, K. Otsuka, and O. Pironneau, freefem+, a finite element software to handle several meshes, 1999. Available via ftp. from ftp://ftp.ann.jussieu.fr/pub/soft/pironneau/.

[3] J. T. Betts and W. P. Huffman, *Mesh refinement in direct transcription methods for optimal control*, Optimal Control Appl. Methods, 19 (1998), pp. 1–21.

[4] R. G. Carter *On the global convergence of trust region algorithms using inexact gradient information*, SIAM J. Numer. Anal., 28 (1991), pp. 251–265.

[5] R. G. Carter, *Numerical experience with a class of algorithms for nonlinear optimization using inexact function and gradient information*, SIAM J. Sci. Comput., 14, (1993), pp. 368–388.

[6] P. G. Ciarlet, *The Finite Element Method*, Prentice–Hall, Englewood Cliffs, NJ, 1977.

[7] P. Deuflhard. *A modified Newton method for the solution of ill-conditioned systems of nonlinear equations with application to multiple shooting*, Numer. Math., 22 (1974), pp. 289–315.

[8] P. Deuflhard, *A relaxation strategy for the modified Newton method. Optimization and optimal control*, in Proceedings of the Conference on Optimization and Optimal Control, Oberwolfach, West Germany, 1974, R. Bulirsch, W. Oettli, and J. Stoer, eds., Springer-Verlag, Berlin, 1975, pp. 59–73.

[9] P. Deuflhard, *Global inexact Newton methods for very large scale nonlinear problems*, Impact of Computing in Science and Engineering, 3 (1991), pp. 366–393.

[10] J. C. Dunn and E. W. Sachs, *The effect of perturbations on the convergence rates of optimization algorithms*, Appl. Math. Optim., 10 (1983), pp. 143–147.

[11] C. T. Kelley and E. W. Sachs, *Fast algorithms for compact fixed point problems with inexact function evaluations*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 725–742.

[12] C. T. Kelley, and E. W. Sachs, *A trust region method for parabolic boundary control problems*, SIAM J. Optim., 9 (1999), pp. 1064–1081.

[13] J.-L. Lions, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.

[14] P. L. Lions, *On the Schwarz alternating method* I, in Proceedings of the International Symposium on Domain Decomposition Methods for Partial Differential Equations, SIAM, Philadelphia, 1988, pp. 1–42.

[15] P. L. Lions, *On the Schwarz alternating method* II, in Proceedings of the International Symposium on Domain Decomposition Methods for Partial Differential Equations, SIAM, Philadelphia, 1989, pp. 47–70.

[16] P. L. Lions, *On the Schwarz alternating method* III, in Proceedings of the International Symposium on Domain Decomposition Methods for Partial Differential Equations, SIAM, Philadelphia, 1990, pp. 202–223.

[17] J.-L. Lions and O. Pironneau, *Algorithmes parallèles pour la solution de problèmes aux limites*, C. R. Acad. Sci. Paris Sér. I Math., 327 (1998), pp. 947–952.

[18] D. Q. Mayne and E. Polak, *A feasible directions algorithm for optimal control problems with terminal inequality constraints*, IEEE Trans. Automat. Control, 22 (1977), pp. 741–751.

[19] E. Polak and D. Q. Mayne, *An algorithm for optimization problems with functional inequality constraints*, IEEE Trans. Automat. Control, 21 (1976), pp. 184–193.

[20] E. Polak, *On the use of consistent approximations in the solution of semi-infinite optimization and optimal control problems*, Math. Programming, 62 (1993), pp. 385–414.

[21] E. Polak, *Optimization: Algorithms and Consistent Approximations*, Springer-Verlag, New York, 1997.

[22] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Heidelberg, 1997.

[23] E. Sachs, *Rates of convergence for adaptive Newton methods*, J. Optim. Theory Appl., 48 (1986), pp. 175–190.

[24] A. L. Schwartz and E. Polak, *Consistent approximations for optimal control problems based on Runge–Kutta integration*, SIAM J. Control Optim., 34 (1996), pp. 1235–1269.

[25] A. L. Schwartz, *RIOTS The Most Powerful Optimal Control Problem Solver*, 1995. Available from http://www.accesscom.com/ adam/RIOTS/.

[26] A. L. Schwartz and E. Polak, *Consistent approximations for optimal control problems based on Runge–Kutta integration*, SIAM J. Control Optim., 34 (1996), pp. 1235–1269.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 2, pp. 511–541

INDIRECT BOUNDARY STABILIZATION OF WEAKLY COUPLED HYPERBOLIC SYSTEMS*

FATIHA ALABAU-BOUSSOUIRA†

Abstract. This work is concerned with the boundary stabilization of an abstract system of two coupled second order evolution equations wherein only one of the equations is stabilized (indirect damping; see, e.g., *J. Math. Anal. Appl.*, 173 (1993), pp. 339–358). We show that, under a condition on the operators of each equation and on the boundary feedback operator, the energy of smooth solutions of this system decays polynomially at ∞. We then apply this abstract result to several systems of partial differential equations (wave-wave systems, Kirchhoff–Petrowsky systems, and wave-Petrowsky systems).

Key words. boundary stabilization, indirect damping, hyperbolic systems, abstract linear evolution equations

AMS subject classifications. 34G10, 35B35, 35B37, 35L90, 93D15, 93D20

PII. S0363012901385368

1. Introduction. Motivations. It is well known that the energy of the solutions of the wave equation in a bounded open domain $\Omega \subset \mathbb{R}^N$,

$$\begin{cases} u_{tt} - \Delta u = 0 & \text{in } \Omega \times (0,\infty), \\ u(.,0) = u_0(.), u_t(.,0) = u_1(.) & \text{in } \Omega, \end{cases} \tag{1}$$

is dissipated when a boundary feedback of the form

$$\frac{\partial u}{\partial \nu} + au + lu_t = 0 \quad \text{on} \quad \Sigma_1 = \Gamma_1 \times (0,\infty) \tag{2}$$

is applied on a part Γ_1 of the boundary Γ of Ω that satisfies certain geometric conditions (see [5], [27], [18]), whereas no feedback is applied on the other part of the boundary, i.e.,

$$u = 0 \quad \text{on} \quad \Sigma_0 = (\Gamma - \Gamma_1) \times (0,\infty). \tag{3}$$

Here Δ stands for the Laplacian with respect to the spatial variables, and the subscript t stands for the partial derivative with respect to the t-variable. We recall that the energy of a solution u of the wave equation is defined by

$$E(u(t)) = \frac{1}{2}\int_\Omega (|u_t|^2 + |\nabla u|^2)\, dx,$$

and that, formally, the dissipation of energy is given by the relation

$$E'(u(t)) = -\int_{\Gamma_1} \ell |u_t|^2 \, d\gamma \le 0 \quad \text{for } \ell \ge 0 \text{ on } \Gamma_1.$$

Moreover, if the feedback coefficients a and l are suitably chosen (see, e.g., [18] and the references therein), the dissipation of the energy through the part Γ_1 of the boundary

*Received by the editors February 16, 2001; accepted for publication (in revised form) November 30, 2001; published electronically June 26, 2002.
http://www.siam.org/journals/sicon/41-2/38536.html
†MMAS, Université de Metz and CNRS (FRE 2344), 57045 Metz, France (alabau@poncelet.univ-metz.fr).

is sufficient to lead to exponential decay of the solutions; i.e., there exist positive constants M and ω such that

$$E(u(t)) \leq M \exp(-\omega t) E(u(0))$$

for all initial data (u_0, u_1) of finite energy. On the other hand, when no feedback is applied on the boundary, i.e., when

$$u = 0 \quad \text{on} \quad \Sigma = \Gamma \times (0, \infty), \tag{4}$$

then the energy of the solutions is conserved, that is, $E(u(t)) = E(u(0))$ for all $t \geq 0$.

One question of interest then is *how the stability properties are affected if we couple the exponentially stable wave equations* (1)–(3) *to the conservative wave equations* (1) and (4). That is, we wonder how these properties are affected if we consider the following system:

$$\begin{cases} u_{1,tt} - \Delta u_1 + \alpha u_2 = 0 & \text{in } \Omega \times (0, \infty), \\ u_{2,tt} - \Delta u_2 + \alpha u_1 = 0 & \text{in } \Omega \times (0, \infty), \\ \dfrac{\partial u_1}{\partial \nu} + a u_1 + l u_{1,t} = 0 & \text{on } \Sigma_1 = \Gamma_1 \times (0, \infty), \\ u_1 = 0 \quad \text{on } \Sigma_0 = \Gamma_0 \times (0, \infty), \qquad u_2 = 0 \quad \text{on } \Sigma = \Gamma \times (0, \infty), \\ u_i(0) = u_{i0}, u_i'(0) = u_{i1}, \end{cases} \tag{5}$$

where α is a coupling parameter. For this model problem, one can remark that a boundary feedback is applied directly to the first component of the solution, whereas no direct feedback is applied to the second component. Also of interest is the question *is it possible to obtain a somehow general result for abstract systems of second order evolution equations coupling a conservative equation to an exponentially stable one*? The abstract model that we refer to in this paper is

$$\begin{cases} u_1'' + A_1 u_1 + B u_1' + \alpha P u_2 = 0 \quad \text{in } V_1', \\ u_2'' + A_2 u_2 + \alpha P^\star u_1 = 0 \quad \text{in } V_2', \\ (u_1, u_1')(0) = (u_1^0, u_1^1) = U_1^0 \in V_1 \times H, \\ (u_2, u_2')(0) = (u_2^0, u_2^1) = U_2^0 \in V_2 \times H, \end{cases} \tag{6}$$

where H, $V_1 \subset H$ and $V_2 \subset H$ are separable Hilbert spaces; A_1, A_2 are coercive self-adjoint unbounded operators in H; B is unbounded symmetric in H, whereas the coupling operator P is assumed to be bounded in H; $P^\star$ is the adjoint operator of P; and α is a coupling parameter. The total energy of a solution (u_1, u_2) is defined by

$$E(u_1(t), u_2(t)) = \frac{1}{2}\left(||u_1'(t)||_H^2 + ||u_2'(t)||_H^2 + ||A_1^{1/2} u_1(t)||_H^2 + ||A_2^{1/2} u_2(t)||_H^2 \right) + \alpha (u_1, P u_2)_H,$$

where $||\cdot||_H$ and $(\cdot)_H$ denote, respectively, the norm and scalar product in H, and $A_i^{1/2}$, $i \in \{1, 2\}$, denotes the usual fractional power of a coercive self-adjoint operator A_i in H (see [29]).

Now the questions of interest are *is the full above system stable and, if so, at which rate*? We can first remark that if $\alpha = 0$, then the solutions of both the above

model problem and the abstract one are not even strongly stable. Hence, the results we are looking for cannot be obtained by a perturbation argument with respect to the case $\alpha = 0$. Moreover, in [2] we have studied the above abstract system in the case of bounded feedback operators B (internal stabilization case). Using a general result of [2, section 1] concerning systems of first evolution equations coupling a conservative equation to a nonconservative one (the coupling operator in the conservative equation being compact in the product space), we can deduce that our abstract system (6) is never exponentially stable. Hence if stability holds for such systems, it must be a weaker stability criterion than the exponential one. Indeed, our main result (see Theorem 3.3 in section 3 below) shows that if $|\alpha|$ is sufficiently small (but not equal to 0), the total energy of the smooth solutions of (6) decays polynomially at ∞, provided mainly that the semigroup generated by A_1 is exponentially stable (see hypothesis $(H1)$), and the two operators A_0 (where A_0 is defined from A_1 in section 2) and A_2 satisfy a compatibility condition (see $(H3)$). From this result, we deduce, thanks to the density of the domains of powers of the generator of the associated first order evolution equation and to the contractivity of the corresponding semigroup, that any solution of (6) is strongly stable. The main requirement for proving the above result is to obtain a generalized integral inequality of the form

$$\int_S^T E(u_1(t), u_2(t))dt \le c \sum_{p=0}^{k} E(u_1^{(p)}(S), u_2^{(p)}(S)) \quad \forall 0 \le S \le T$$

for smooth initial data, where $k = 2$ in our case and where E stands for the total energy of the system. Then a general lemma (introduced in [1] and valid for arbitrary nonzero integer k) shows that if in addition E is a nonincreasing function, it has to decay polynomially at ∞. The main technical point is then to prove this generalized inequality. This is done by the use of appropriate multipliers, one of the obvious difficulties being that, with the second equation not directly stabilized, we miss some information that we must get back from the system.

We then apply our abstract result to several systems of partial differential equations. In particular, we obtain a polynomial decay rate for the energy of smooth solutions of (5) under classical (multiplier-type) geometric conditions on the boundary where the feedback is active. Similar results can be obtained for a Kirchhoff–Petrowsky system. For these two examples, the operators A_0 and A_2 coincide. We further give two examples for which these two operators do not coincide, namely, the case of two coupled wave equations with different speeds of propagation, and a wave-Petrowsky coupling. In these two latter cases, we have to restrict our analysis to the situation in which the spatial domain is an n-dimensional interval.

These results can also be interpreted in the framework of indirect stabilization, which, as far as we know, was introduced by Russell [32]. Indeed, since strong stability holds for abstract systems of the form (6), the first equation of this system can be viewed as an *indirect* stabilizer for the second equation.

We will now give some references to the existing literature on this subject, referring the reader to those papers for further references. A large number of papers (see, e.g., [10], [28], [26], [24], [3], [30]) concern the stabilization of hyperbolic-parabolic coupled systems, such as thermoelasticity, thermoplates, etc. For such systems, the main goal is to determine whether the dissipation induced by the heat-type equation is sufficient for stabilizing the full system obtained by coupling it to a hyperbolic-type equation. Exponential stability results for coupled hyperbolic-hyperbolic systems via two feedback operators (i.e., each equation of the system is directly stabilized if the

coupling parameter is set equal to 0; this is the *direct* stabilization case) can be found in [17], [20], [23]. In the case of coupled wave-wave systems subjected to only one internal feedback operator (this is the *indirect* internal stabilization situation), positive and negative exponential stability results have been obtained in [16], [4]. In [2] we have proved polynomial decay estimates in the indirect internal stabilization case. These results were extended to several cases (wave-wave, Petrowsky-Petrowsky coupling) for locally distributed indirect stabilization in [8]. In the one-dimensional case, another approach, based on the use of a Riesz basis, leads to stability results with an optimal decay rate (see, e.g., [31], [14]). This technique is based on determination of precise asymptotic estimates of the eigenvalues of the involved operators for large frequencies. Such precise estimates can be obtained only in the one-dimensional case and require a careful analysis of the associated spectral problems. Moreover, this analysis has to be performed for each new system under consideration. The main advantage of the method presented in our paper is that it is valid in any dimension and for a wide class of weakly coupled systems, without requiring the performance of all computations for each new system. However, the stability results obtained by this method are probably not optimal, in contrast to the one-dimensional results derived by the use of the Riesz basis.

Complete and partial observability (respectively, controllability) results for coupled systems of either hyperbolic-hyperbolic type or of hyperbolic-parabolic type can be found in [27]. These results assume that the coupling parameter is sufficiently small. They have been extended in [19] to the cases of arbitrary coupling parameters (assuming bounded coupling operators). For both references, the multiplier method was the main requirement for obtaining the desired estimates. Complete observability (respectively, controllability) results have also been obtained in [25] for systems of coupled second order hyperbolic equations containing first order terms in both the original and the coupled unknowns. These results are based on Carleman estimates.

The paper is organized as follows. In section 2, we give the abstract framework for system (6) and establish the well-posedness of both the uncoupled and coupled abstract systems. In section 3, we establish a polynomial decay lemma for a nonincreasing nonnegative functional satisfying a generalized integral inequality. We then prove our main stability result and some useful corollaries. Finally, in section 4 we give several applications of our main results to systems of partial differential equations.

2. Abstract coupled model.

2.1. Introduction. Let V_i, $i = 1, 2$, and H be separable real Hilbert spaces such that the injections $V_i \subset H$ are dense, compact, and continuous for $i = 1, 2$, and the injection $V_2 \subset V_1$ is continuous.

In all of what follows we identify H with its dual space, so that the injections $V_i \subset H \subset V_i'$ hold and are continuous, dense, and compact. The scalar products on V_i, $i = 1, 2$, and H are respectively denoted by $(,)_{V_i}$ and $(,)_H$, whereas the corresponding norms are respectively denoted by $||\ ||_{V_i}$ and $||\ ||_H$. Moreover, we denote by $\langle , \rangle_{V_i', V_i}$ the duality product, and by A_i, $i = 1, 2$, the duality mapping from V_i to V_i' defined by

$$\langle A_i w, z\rangle_{V_i', V_i} = (w, z)_{V_i} \qquad \forall w, z \in V_i.$$

It will also be useful (see subsection 3.2) to assume that V_1 contains a closed subspace V_0, equipped with the norm and scalar product induced by those of V_1. Then denoting by i the canonical injection from V_0 in V_1, and by P_0 the operator of

projection from V_1 on V_0, we recall that for any $u_1 \in V_1$, $P_0 u_1$ is characterized by

$$\begin{cases} \langle A_1 i(P_0 u_1), i(\phi)\rangle_{V_1',V_1} = \langle A_1 u_1, i(\phi)\rangle_{V_1',V_1} \quad \forall\ \phi \in V_0,\ u_1 \in V_1, \\ P_0 u_1 \in V_0. \end{cases} \tag{7}$$

We define an operator A_0 from V_0 to V_0' by

$$\langle A_0\phi, \psi\rangle_{V_0',V_0} = \langle A_1 i(\phi), i(\psi)\rangle_{V_1',V_1} \quad \forall\ \phi, \psi \in V_0. \tag{8}$$

Note that A_0 is, indeed, the duality mapping from V_0 to V_0'.

Let B be a given linear continuous operator from V_1 to V_1' (it will serve as the abstract formulation of the boundary conditions), which further satisfies

$$\langle Bu, u\rangle_{V_1',V_1} \geq 0, \quad \langle Bu, z\rangle_{V_1',V_1} = \langle Bz, u\rangle_{V_1',V_1} \quad \forall u, z \in V_1. \tag{9}$$

Moreover, let P be a given linear continuous operator on H, and α be a given nonzero parameter. For the sake of clarity we will assume that α is positive; nevertheless, the results in this paper are valid for negative α as well. We consider the following weakly coupled system:

$$\begin{cases} u_1'' + A_1 u_1 + Bu_1' + \alpha P u_2 = 0 \quad \text{in } V_1', \\ u_2'' + A_2 u_2 + \alpha P^\star u_1 = 0 \quad \text{in } V_2', \\ (u_1, u_1')(0) = (u_1^0, u_1^1) = U_1^0 \in V_1 \times H, \\ (u_2, u_2')(0) = (u_2^0, u_2^1) = U_2^0 \in V_2 \times H. \end{cases} \tag{10}$$

In order to study this coupled system, we need to establish basic results on the decoupled system obtained when $\alpha = 0$.

2.2. The decoupled system. We consider the decoupled system

$$\begin{cases} u_1'' + A_1 u_1 + Bu_1' = 0 \quad \text{in } V_1', \\ (u_1, u_1')(0) = U_1^0 \in V_1 \times H, \end{cases} \tag{11}$$

and

$$\begin{cases} u_2'' + A_2 u_2 = 0 \quad \text{in } V_2', \\ (u_2, u_2')(0) = U_2^0 \in V_2 \times H. \end{cases} \tag{12}$$

We set $\mathcal{H}_i = V_i \times H$ for $i = 1, 2$. This space is equipped with the scalar product

$$((u_i, v_i), (\tilde{u_i}, \tilde{v_i}))_{\mathcal{H}_i} = (u_i, \tilde{u_i})_{V_i} + (v_i, \tilde{v_i})_H \qquad \forall (u_i, v_i), (\tilde{u_i}, \tilde{v_i}) \in \mathcal{H}_i$$

and the corresponding norm $||\ ||_{\mathcal{H}_i}$. We define two linear unbounded operators $\mathcal{A}_i$ on $\mathcal{H}_i$ for $i = 1, 2$ by

$$\mathcal{A}_1(u_1, v_1) = (-v_1, A_1 u_1 + Bv_1), \quad D(\mathcal{A}_1) = \{(u_1, v_1)\ \in V_1 \times V_1,\ A_1 u_1 + Bv_1\ \in H\},$$

$$\mathcal{A}_2(u_2, v_2) = (-v_2, A_2 u_2), \quad D(\mathcal{A}_2) = \{(u_2, v_2)\ \in V_2 \times V_2,\ A_2 u_2\ \in H\}.$$

Then the decoupled system (11)–(12) can be reformulated as

$$\begin{cases} (u_i, v_i)' + \mathcal{A}_i(u_i, v_i) = 0, \quad i = 1, 2, \\ (u_i, v_i)(0) = U_i^0\ \in \mathcal{H}_i, \quad i = 1, 2. \end{cases} \tag{13}$$

PROPOSITION 2.1. *We assume the above hypotheses on V_i, H, A_i, and B. Then $\mathcal{A}_i$ is a maximal monotone operator on $\mathcal{H}_i$ for $i = 1, 2$, so that for every $U_i^0 = (u_i^0, u_i^1) \in \mathcal{H}_i$ problem (13) has a unique solution $(u_i, v_i) \in \mathcal{C}([0, +\infty); \mathcal{H}_i)$, $i = 1, 2$. Moreover, the energy of the solution, defined by*

$$e_i(t) = \frac{1}{2} \, ||(u_i, v_i)||^2_{\mathcal{H}_i}, \tag{14}$$

is locally absolutely continuous for $i = 1, 2$, and e_1 is nonincreasing, whereas e_2 is conserved through time. If, in addition, $U_i^0 \in D(\mathcal{A}_i^k)$ for $k \in \mathbb{N}^\star$, then the solution is in $\mathcal{C}^{k-j}([0, +\infty); D(\mathcal{A}_i^j))$ for $j = 0, \ldots, k$ and $i = 1, 2$.

Proof. For $(u_1, v_1) \in D(\mathcal{A}_1)$ we have

$$(\mathcal{A}_1(u_1, v_1), (u_1, v_1))_{\mathcal{H}_1} = (-v_1, u_1)_{V_1} + (A_1 u_1 + B v_1, v_1)_H = \langle B v_1, v_1 \rangle_{V_1', V_1} \geq 0,$$

since B satisfies (9). In the same way, for $(u_2, v_2) \in D(\mathcal{A}_2)$ we obtain $(\mathcal{A}_2(u_2, v_2), (u_2, v_2))_{\mathcal{H}_2} = 0$. Hence $\mathcal{A}_i$ is a monotone operator for $i = 1, 2$. We now prove that $\mathcal{A}_i$ is onto. For this, it is sufficient to prove that $I + A_1 + B$ and $I + A_2$ are onto. Let $h_1 \in V_1'$ be given arbitrarily. We consider, as usual (see, e.g., [18]), the map F_1 defined from V_1 on $\mathbb{R}$ by

$$F_1(u_1) = \frac{1}{2}||u_1||^2_H + \frac{1}{2}||u_1||^2_{V_1} + \frac{1}{2}\langle B u_1, u_1 \rangle_{V_1', \, V_1} - \langle h_1, u_1 \rangle_{V_1', \, V_1}.$$

Then F_1 is continuously differentiable and

$$F_1'(u_1) \cdot \phi_1 = \langle (I + A_1 + B) u_1 - h_1, \phi_1 \rangle_{V_1', V_1} \quad \forall u_1, \phi_1 \in V_1.$$

Moreover, F_1 is convex and coercive, i.e, $F_1(u_1) \to +\infty$ if $||u_1||_{V_1} \to +\infty$. Therefore, F_1 attains its minimum at some point $u_1 \in V_1$ for which $F_1'(u_1) = 0$. Hence we have $(I + A_1 + B)u_1 = h_1$. We then deduce easily that $I + \mathcal{A}_1$ is onto on $\mathcal{H}_1$. In a similar way, we prove that $I + \mathcal{A}_2$ is onto on $\mathcal{H}_2$. Hence the operators $\mathcal{A}_i$, $i = 1, 2$, are maximal monotone. We easily conclude, using well-known properties of maximal monotone linear operators (see, e.g., [7]).

2.3. Abstract formulation of the coupled system. We now turn back to the weakly coupled system (10). We set $V = V_1 \times V_2$. This space is equipped with the usual scalar product $(u, \tilde{u})_V = (u_1, \tilde{u_1})_{V_1} + (u_2, \tilde{u_2})_{V_2}$ and the corresponding norm $|| \, ||_V$, where $u = (u_1, u_2) \in V$ and $\tilde{u} = (\tilde{u_1}, \tilde{u_2}) \in V$. We have $V \subset H \times H \subset V'$ with continuous, dense, and compact injections. We also define a linear continuous operator A_α from V on V' by

$$A_\alpha u = (A_1 u_1 + \alpha P u_2, A_2 u_2 + \alpha P^\star u_1), \qquad u = (u_1, u_2) \in V.$$

Moreover, we consider on V the continuous bilinear form

$$(u, \tilde{u})_\alpha = (u, \tilde{u})_V + \alpha(P u_2, \tilde{u_1})_H + \alpha(P^\star u_1, \tilde{u_2})_H, \qquad u = (u_1, u_2), \tilde{u} = (\tilde{u_1}, \tilde{u_2}) \in V.$$

PROPOSITION 2.2. *Assume the hypotheses of Proposition* 2.1. *Then there exists $\alpha_0 > 0$ such that for all $0 \leq |\alpha| < \alpha_0$ there exist constants $c_1(\alpha) > 0$ and $c_2(\alpha) > 0$ such that*

$$c_1(\alpha)||u||_V \leq ((u, u)_\alpha)^{1/2} \leq c_2(\alpha)||u||_V \quad \forall \; u \in V.$$

Hence, for all $0 \leq |\alpha| < \alpha_0$, *the application*

$$u \in V \mapsto ||u||_\alpha = ((u,u)_\alpha)^{1/2}$$

defines a norm on V *which is equivalent to the norm* $||\ ||_V$. *Moreover, for all* $0 \leq |\alpha| < \alpha_0$, A_α *is the duality mapping from* V *on* V' *when* V *is equipped with the scalar product* $(,)_\alpha$.

Proof. Let us denote by β_i, $i = 1, 2$, the smallest positive constants such that

$$||u_i||_H \leq \beta_i ||u_i||_{V_i} \quad \forall\ u_i \in V_i.$$

Then we have

$$(u,u)_\alpha \geq ||u_1||^2_{V_1} + ||u_2||^2_{V_2} - 2\alpha||P||\beta_1\beta_2||u_1||_{V_1}||u_2||_{V_2} \geq ||u||^2_V(1 - \alpha||P||\beta_1\beta_2).$$

Hence, setting

$$\alpha_0 = (||P||\beta_1\beta_2)^{-1}, \qquad c_1(\alpha) = \sqrt{1 - \alpha||P||\beta_1\beta_2},$$

we have

$$\sqrt{(u,u)_\alpha} \geq c_1(\alpha)||u||_V \quad \forall\ u \in V.$$

In a similar way, we have

$$\sqrt{(u,u)_\alpha} \leq c_2(\alpha)||u||_V \quad \forall\ u \in V,$$

where

$$c_2(\alpha) = \sqrt{1 + \alpha||P||\beta_1\beta_2}.$$

We now prove that A_α is the duality mapping for the scalar product $(,)_\alpha$. For $u = (u_1, u_2)$ and $u = (\tilde{u}_1, \tilde{u}_2)$ given in V we have

$$\langle A_\alpha u, \tilde{u}\rangle_{V',V} = (u,\tilde{u})_V + \alpha\langle Pu_2, \tilde{u_1}\rangle_{V_1',V_1} + \alpha\langle P^\star u_1, \tilde{u_2}\rangle_{V_2',V_2}.$$

Now, since $Pu_2 \in H$ and $\tilde{u_1} \in V_1$, we have

$$\langle Pu_2, \tilde{u_1}\rangle_{V_1',V_1} = (Pu_2, \tilde{u_1})_H.$$

In a similar way, we have

$$\langle P^\star u_1, \tilde{u_2}\rangle_{V_2',V_2} = (P^\star u_1, \tilde{u_2})_H.$$

Hence

$$\langle A_\alpha u, \tilde{u}\rangle_{V',V} = (u,\tilde{u})_\alpha. \qquad \square$$

We now set $\mathcal{H} = V \times H^2$. This space is equipped with the scalar product

$$(U,\tilde{U})_{\mathcal{H}} = (u,\tilde{u})_\alpha + (v,\tilde{v})_{H\times H}$$

and the corresponding norm

$$||U||_{\mathcal{H}} = (||u||^2_\alpha + ||v||^2_{H\times H})^{1/2},$$

where $U = (u, v) \in \mathcal{H}$, with $u = (u_1, u_2)$, $v = (v_1, v_2)$. We also define the unbounded linear operator $\mathcal{A}_\alpha$ on $\mathcal{H}$ by

$$\mathcal{A}_\alpha U = (-v, A_\alpha u + (Bv_1, 0)),$$

$$D(\mathcal{A}_\alpha) = \{U = (u, v) = ((u_1, u_2), (v_1, v_2)) \in V \times V,\ A_\alpha u + (Bv_1, 0) \in H \times H\}.$$

One can easily prove that

$$U = ((u_1, u_2), (v_1, v_2)) \in D(\mathcal{A}_\alpha) \Longleftrightarrow (u_1, v_1) \in D(\mathcal{A}_1), (u_2, v_2) \in D(\mathcal{A}_2).$$

We can now reformulate the system (10) as the abstract first order equation

$$\begin{cases} U' + \mathcal{A}_\alpha U = 0, \\ U(0) = U^0 \in \mathcal{H}. \end{cases} \tag{15}$$

PROPOSITION 2.3. *Assume the hypotheses of Proposition* 2.1*, and let* α_0 *be given as in Proposition* 2.2*. Then, for all* $0 \leq |\alpha| < \alpha_0$*,* $\mathcal{A}_\alpha$ *is a maximal monotone linear operator on* $\mathcal{H}$*, so that for every* $U^0 \in \mathcal{H}$ *problem* (15) *has a unique solution* $U = (u, v) \in \mathcal{C}([0, +\infty); \mathcal{H})$*. If, in addition,* $U^0 \in D({\mathcal{A}_\alpha}^k)$ *for* $k \in \mathbb{N}^\star$*, then the solution is in* $\mathcal{C}^{k-j}([0, +\infty); D(\mathcal{A}_\alpha^j))$ *for* $j = 0, \ldots, k$*. Moreover, the energy of the solution defined by*

$$E(U(t)) = \frac{1}{2}\ ||U||^2_{\mathcal{H}} \tag{16}$$

is locally absolutely continuous, and for strong solutions, i.e., when $U^0 \in D(\mathcal{A}_\alpha)$*, we have*

$$E'(U(t)) = -\langle Bu_1', u_1' \rangle_{V_1', V_1}; \tag{17}$$

here $'$ *denotes the derivative with respect to time* t*. Hence the energy of any solution of* (15) *is a nonincreasing function of time.*

Proof. The proof is similar to that of Proposition 2.1 and is left to the reader.

Remark. The well-posedness of problem (15) holds true for any α, since $\mathcal{A}_\alpha$ is a compact perturbation of the corresponding decoupled operator (obtained by setting $\alpha = 0$). Of course, in this case V should be equipped with the norm $||\ ||_V$, and $\mathcal{H}$ with the scalar product $(U, \tilde{U}) = (u, \tilde{u})_V + (v, \tilde{v})_{H \times H}$ and the corresponding norm.

For what follows, we will also need the following result whose (easy) proof is left to the reader.

PROPOSITION 2.4. *Assume the hypotheses of Proposition* 2.1*, and let* α_0 *be given as in Proposition* 2.2*. Then, for all* $0 \leq |\alpha| < \alpha_0$*, there exist constants* $c_3(\alpha) > 0$ *and* $c_4(\alpha) > 0$ *such that for all* $U = (u_1, u_2, v_1, v_2) \in \mathcal{H}$ *we have*

$$\begin{aligned} \frac{c_3(\alpha)}{2}\ (||(u_1, v_1)||^2_{\mathcal{H}_1} + ||(u_2, v_2)||^2_{\mathcal{H}_2}) &\leq ||U||^2_{\mathcal{H}} \\ &\leq \frac{c_4(\alpha)}{2}(||(u_1, v_1)||^2_{\mathcal{H}_1} + ||(u_2, v_2)||^2_{\mathcal{H}_2}). \end{aligned} \tag{18}$$

3. Main results.

3.1. A generalized integral inequality. We prove below a generalized integral inequality that will be useful in what follows for obtaining polynomial decay rates for smooth solutions of coupled equations when only one of the equations is stabilized. Let A be the infinitesimal generator of a continuous semigroup $\exp(tA)$ on a Hilbert space $\mathcal{H}$, and $D(A)$ its domain. For U^0 in $\mathcal{H}$ we set $U(t) = \exp(tA)U^0$ in all of what follows.

THEOREM 3.1. *Assume that there exists a functional E defined on $C([0,+\infty),\mathcal{H})$ such that for every U^0 in $\mathcal{H}$, $E(\exp(.A))$ is a nonincreasing, locally absolutely continuous function from $[0,+\infty)$ on $[0,+\infty)$. Assume, moreover, that there exist an integer $k \in \mathbb{N}^\star$ and nonnegative constants c_p for $p = 0,\dots,k$ such that*

$$\int_S^T E(U(t))\ dt \le \sum_{p=0}^{k} c_p E(U^{(p)}(S)) \qquad \forall\ 0 \le S \le T,\ U^0 \in D(A^k). \tag{19}$$

Then the following inequalities hold for every U^0 in $D(A^{kn})$, where n is any positive integer:

$$\int_S^T E(U(\tau))\frac{(\tau - S)^{n-1}}{(n-1)!}\ d\tau \le c\sum_{p=0}^{kn} E(U^{(p)}(S)) \qquad \forall\, 0 \le S \le T,\ U^0 \in D(A^{kn}), \tag{20}$$

and

$$E(U(t)) \le c\sum_{p=0}^{kn} E(U^{(p)}(0))t^{-n} \qquad \forall t > 0,\ U^0 \in D(A^{kn}),$$

where c is a constant that depends on n.

Proof. We first prove (20) by induction on n. For $n = 1$, it reduces to the hypothesis (19). Assume now that (20) holds for n, and let U^0 be given in $D(A^{k(n+1)})$. Then we have

$$\int_S^T \int_t^T E(U(\tau))\frac{(\tau - t)^{n-1}}{(n-1)!}\ d\tau\ dt \le c\sum_{p=0}^{kn}\int_S^T E(U^{(p)}(t))\ dt$$

$$\forall 0 \le S \le T,\ U^0 \in D(A^{kn}).$$

Since U^0 is in $D(A^{k(n+1)})$, we deduce that $U^{(p)}(0) = A^p U^0$ is in $D(A^k)$ for $p \in \{0,\dots,kn\}$. Hence we can apply assumption (19) to the initial data $U^{(p)}(0)$. This, together with Fubini's theorem applied on the left-hand side of the above inequality, gives (20) for $n+1$. Using the property that $E(U(t))$ is nonincreasing in (20), we easily obtain the last desired inequality. ∎

In all of what follows, we write U instead of $U(t)$ in the expressions involving the energy, and S will denote a nonnegative real number.

3.2. Polynomial decay of weakly coupled systems. Let us first prove that, under the hypotheses of Proposition 2.1, the semigroup generated by the operator $-\mathcal{A}_\alpha$ is not exponentially stable.

PROPOSITION 3.2. *We assume the hypotheses of Proposition* 2.1 *and that the space* $V_2 \times H$ *is infinite-dimensional. Then the semigroup generated by the operator* $-\mathcal{A}_\alpha$ *is not exponentially stable.*

Proof. We recall that $\mathcal{H}_i = V_i \times H$ and $\mathcal{H} = V_1 \times V_2 \times H^2$, and we define the operators $\mathcal{A}_i$ for $i = 1, 2$ and their domains as in section 2. Now let U_0 be given in $\mathcal{H}$, and $U = ((u_1, u_2), (v_1, v_2))$ be the solution of (15). Setting $w_1 = (u_1, v_1)$ and $w_2 = (u_2, v_2)$, we can reformulate the system (15) in the following form:

$$\text{(21)} \qquad \begin{cases} w_1'(t) = L_1 w_1(t) + K_1 w_2(t), \\ w_2'(t) = L_2 w_2(t) + K_2 w_1(t), \\ w_1(0) = (u_1^0, u_1^1) = U_1^0 \in \mathcal{H}_1, \\ w_2(0) = (u_2^0, u_2^1) = U_2^0 \in \mathcal{H}_2, \end{cases}$$

where the operators L_i are unbounded operators in $\mathcal{H}_i$ for $i = 1, 2$ and are defined by $L_i = -\mathcal{A}_i$, $D(L_i) = D(\mathcal{A}_i)$ for $i = 1, 2$. In contrast, the operators K_1 and K_2 are the bounded linear operators acting, respectively, from $\mathcal{H}_2$ on $\mathcal{H}_1$ and $\mathcal{H}_1$ on $\mathcal{H}_2$ defined by $K_1 = (0, -\alpha P)$ and $K_2 = (0, -\alpha P^\star)$. Now, from our hypotheses and thanks to Proposition 2.1, we know that L_i generates a strongly continuous semigroup of bounded linear operators $\exp(tL_i)$ on $\mathcal{H}_i$ for $i = 1, 2$, and that $||\exp(tL_2)w|| = ||w||$ for all $w \in \mathcal{H}_2$, where $||\cdot||$ denotes the norm on $\mathcal{H}_2$. Moreover, one can easily notice that, thanks to the compact imbedding of V_2 in H, K_2 is a compact operator. Hence if $V_2 \times H$ is infinite-dimensional, we can apply the results of [2, section 1], so that the semigroup generated by $-\mathcal{A}_\alpha$ is not exponentially stable. □

Hence if the system (15) is stable, it must be stable in a weaker sense than the exponential one. Indeed, we now want to prove that, under additional hypotheses on the operators $(\mathcal{A}_1, \mathcal{A}_2)$, the operator $-\mathcal{A}_\alpha$ generates a polynomially decaying semigroup.

We first assume that the operator $\mathcal{A}_1$ satisfies the following hypotheses (H1)–(H2):

$$\text{(H1)} \begin{cases} \exists \gamma_i > 0,\ i = 1, 2, 3, \text{ such that } \forall f_1 \in \mathcal{C}^1([0, +\infty); H) \text{ and } 0 \le S \le T, \\ \text{the solution } (u_1, v_1) \text{ of} \\ \quad (u_1, v_1)' + \mathcal{A}_1(u_1, v_1) = (0, f_1), \\ \quad (u_1, v_1) = (u_1^0, v_1^0) \in D(\mathcal{A}_1) \\ \text{satisfies} \\ \int_S^T e_1(t)\, dt \le \gamma_1(e_1(S) + e_1(T)) + \gamma_2 \int_S^T ||f_1(t)||_H^2\, dt + \gamma_3 \int_S^T \langle Bv_1, v_1\rangle_{V_1', V_1}\, dt; \end{cases}$$

and

$$\text{(H2)} \begin{cases} \langle Bu_1, i(\phi)\rangle_{V_1', V_1} = 0 \quad \forall\ \phi \in V_0,\ u_1 \in V_1, \\ \text{and} \\ \exists \beta > 0, \qquad ||u_1 - P_0 u_1||_H^2 \le \beta\ \langle Bu_1, u_1\rangle_{V_1', V_1} \quad \forall u_1 \in V_1. \end{cases}$$

Remarks. The assumption (H1) implies, in particular, that $-\mathcal{A}_1$ generates an exponentially stable semigroup, since for $f_1 = 0$ we deduce that e_1, which is locally absolutely continuous and nonincreasing (see Proposition 2.1), satisfies the classical integral inequality

$$\text{(22)} \qquad \int_S^T e_1(t)\ dt \le (2\gamma_1 + \gamma_3) e_1(S) \quad \forall\ 0 \le S \le T,$$

so that $-\mathcal{A}_1$ generates an exponentially stable semigroup (see [15], [18]). As will be seen later (in section 4), this property is satisfied for most systems (e.g., wave, Kirchhoff, etc.).

The hypothesis (H2) implies that B satisfies a "weak" coercivity property (since the norm on the left-hand side of the second inequality in (H2) is the weaker H-norm) in the subspace orthogonal to the closed subspace V_0. As will be seen later (in section 4), this property is satisfied for most systems (e.g., wave, Kirchhoff, etc.).

We now state the next hypothesis, which gives the "authorized" couplings in the abstract system (10). For the sake of simplicity, we still denote by A_i the unbounded operator on H defined by the restriction of A_i to $D(A_i) = \{\phi \in V_i,\ A_i\phi \in H\}$ for $i = 0$ and $i = 2$. We now assume the following properties on the coupling:

$$\text{(H3)} \begin{cases} V_2 \subset V_0 \text{ with continuous imbedding and} \\ \exists \text{ a bounded operator } C \text{ on } H \text{ such that} \\ CV_2 \subset V_0, CD(A_2) \subset D(A_0), \text{ and } A_0 C u_2 = C A_2 u_2 \quad \forall u_2 \in D(A_2). \end{cases}$$

Furthermore, we assume that

$$\text{(H4)} \qquad \exists \gamma > 0 \text{ such that } (Pu_2, Cu_2)_H \geq \gamma ||u_2||_H^2 \quad \forall u_2 \in H.$$

We now state the main result of this paper.

THEOREM 3.3. *Assume that A_1, A_2, and B satisfy the hypotheses of section* 2 *and assumptions* (H1)–(H4). *Then there exists an $\alpha_1 \in (0, \alpha_0]$ such that for all $0 < |\alpha| < \alpha_1$ the solution $U(t) = \exp(-\mathcal{A}_\alpha t)U^0$ of* (10) *satisfies*

$$E(U(t)) \leq \frac{c}{t^n} \sum_{p=0}^{2n} E(U^{(p)}(0)) \qquad \forall\ t > 0,\ U^0 \in D(\mathcal{A}_\alpha^{2n}),$$

where c is a constant depending on α and n. Moreover, if $U^0 \in \mathcal{H}$, then $E(U(t))$ converges to zero as t goes to infinity (this is strong stability).

As will be seen in what follows (in section 4), assumptions (H1) and (H2) will be satisfied for many operators $\mathcal{A}_1$. Assumption (H3), which concerns the coupling between the two operators involved in the full system, is more restrictive. We give below two abstract examples for which assumption (H3) is satisfied.

Example 1. Case $A_0 = A_2$. We give a first example for which assumption (H3) is trivially satisfied.

PROPOSITION 3.4. *Assume that*

$$\begin{aligned} &\text{(H3)}' \quad V_0 = V_2, \quad A_0 = A_2, \quad D(A_0) = D(A_2), \\ &\text{(H4)}' \quad \exists \gamma > 0,\ (Pu_2, u_2)_H \geq \gamma ||u_2||_H^2 \quad \forall\ u_2 \in H, \end{aligned}$$

where the equality between Banach spaces E and F has to be understood as E and F are isomorphic. Then assumptions (H3)–(H4) *are trivially satisfied with $C = \mathrm{Id}_H$.*

Therefore, from Theorem 3.3 and Proposition 3.4, we easily deduce the following corollary.

COROLLARY 3.5. *Assume that A_1, A_2, A_0, and B satisfy the hypotheses of section* 2, *and take assumptions* (H1)–(H2) *together with* (H3)′ *and* (H4)′. *Then there exists $\alpha_1 \in (0, \alpha_0]$ such that for all $0 < |\alpha| < \alpha_1$ the solution $U(t) = \exp(-\mathcal{A}_\alpha t)U^0$ of* (10) *satisfies*

$$E(U(t)) \leq \frac{c}{t^n} \sum_{p=0}^{2n} E(U^{(p)}(0)) \qquad \forall\ t > 0,\ U^0 \in D(\mathcal{A}_\alpha^{2n}).$$

Moreover, if $U^0 \in \mathcal{H}$, then $E(U(t))$ converges to zero as t goes to infinity.

Example 2. Case $A_0 \neq A_2$. The second example allows us a more general situation, even though it is still restrictive. We define the spaces H, V_1, V_2, V_0 and the unbounded operators $A_0 : D(A_0) \subset H \mapsto H$, $A_2 : D(A_2) \subset H \mapsto H$ as in section 2, where $D(A_i) = \{u \in V_i, A_i u \in H\}$ for $i = 0, 2$.

LEMMA 3.6. *Assume that $V_2 \subset V_0$ and that there exists a common orthonormal basis $\{e_k\}_{k=1}^{\infty}$ of eigenfunctions of the operators A_i in H, for $i = 0, 2$, with*

$$A_i e_k = \lambda_{i,k} e_k, \qquad k = 1, \dots, \ i = 0, 2.$$

Assume, moreover, that the following hypothesis holds:

$$\text{(H5)} \left\{ \begin{array}{l} \exists r : \mathbb{N}^\star \mapsto \mathbb{N}^\star, \ \textit{one-to-one, such that} \\ \lambda_{2,k} = \lambda_{0,r(k)} \quad \forall k \in \mathbb{N}^\star. \end{array} \right.$$

Then, there exists a bounded linear operator C in H that satisfies assumption (H3) *and $||Cu||_H = ||u||_H$ for all $u \in H$.*

Proof. Assume that there exists an application r as above. For $u \in H$, u has a unique orthonormal expansion $u = \sum_{k=1}^{\infty} u_k e_k$, where $\sum_{k=1}^{\infty} |u_k|^2 < +\infty$. In the remainder of the proof, we will assume that $u \in H$ is written under this form. Moreover, we recall that for $i = 0$ or $i = 2$, $u \in V_i$ if and only if

$$\sum_{k=1}^{\infty} \lambda_{i,k} |u_k|^2 < +\infty,$$

whereas $u \in D(A_i)$ if and only if

$$\sum_{k=1}^{\infty} |\lambda_{i,k}|^2 |u_k|^2 < +\infty.$$

We define C as follows:

$$Cu = \sum_{k=1}^{\infty} u_k e_{r(k)}.$$

Hence, we have $Cu = \sum_{l=1}^{\infty} v_\ell e_\ell$, where $v_\ell = u_k$ if $k \in \{1, \dots, \infty\}$ exists such that $\ell = r(k)$, and $v_\ell = 0$ otherwise. This implies that $||Cu||_H = ||u||_H$ for all $u \in H$. We now check that this operator C satisfies assumption (H3). Let u be given in V_2. Then, thanks to the assumption on r, we have

$$\sum_{k=1}^{\infty} \lambda_{0,r(k)} |u_k|^2 < +\infty.$$

Defining v_ℓ as above, we deduce that

$$\sum_{\ell=1}^{\infty} \lambda_{0,\ell} |v_\ell|^2 = \sum_{k=1}^{\infty} \lambda_{0,r(k)} |u_k|^2 < +\infty.$$

Hence, we have $Cu \in V_0$, and thanks to the assumption on r, $||Cu||_{V_0} = ||u||_{V_2}$ holds.

Now let u be given in $D(A_2)$, and define v_ℓ as above. Then, thanks to the assumption on r, we prove, as above, that

$$\sum_{\ell=1}^{\infty} |\lambda_{0,\ell}|^2 |v_\ell|^2 = \sum_{k=1}^{\infty} |\lambda_{0,r(k)}|^2 |u_k|^2 < +\infty.$$

Hence, we have $Cu \in D(A_0)$. Finally, let u be given in $D(A_2)$. Then

$$A_2 u = \sum_{k=1}^{\infty} \lambda_{2,k} u_k e_k,$$

so that, thanks to the assumption on r,

$$CA_2 u = \sum_{k=1}^{\infty} \lambda_{0,r(k)} u_k e_{r(k)}.$$

On the other hand, we have

$$A_0 Cu = \sum_{\ell=1}^{\infty} \lambda_{0,\ell} v_l e_\ell = CA_2 u,$$

thanks to the above definition of v_ℓ. □

Thanks to Theorem 3.3 and to Lemma 3.6, we also deduce the following corollary of our main result.

COROLLARY 3.7. *Assume that A_1, A_2, A_0, and B satisfy the hypotheses of section* 2. *Moreover, assume that* (H1)–(H2), (H4), *and the assumptions of Lemma* 3.6 *hold. Then there exists $\alpha_1 \in (0, \alpha_0]$ such that for all $0 < |\alpha| < \alpha_1$ the solution $U(t) = \exp(-\mathcal{A}_\alpha t)U^0$ of* (10) *satisfies*

$$E(U(t)) \le \frac{c}{t^n} \sum_{p=0}^{2n} E(U^{(p)}(0)) \qquad \forall\ t > 0,\ U^0 \in D(\mathcal{A}_\alpha^{2n}).$$

Moreover, if $U^0 \in \mathcal{H}$, then $E(U(t))$ converges to zero as t goes to infinity.

Proof of the main result. We now turn to the proof of Theorem 3.3. However, we first need to prove the following lemmas.

LEMMA 3.8. *Assume the hypotheses of Theorem* 3.3. *Then for all $0 < |\alpha| < \alpha_0$ and all $U^0 = (u_1^0, u_2^0, u_1^1, u_2^1) \in D(\mathcal{A}_\alpha)$ the solution $U(t) = \exp(-t\mathcal{A}_\alpha)U^0 = (u_1, u_2, v_1, v_2)$ of* (10) *satisfies*

$$\frac{\alpha\gamma}{2} \int_S^T ||u_2||_H^2\ dt \le \alpha\ c \int_S^T ||u_1||_H^2\ dt + c(E(U'(S)) + E(U(S))) \quad \forall\ 0 \le S \le T,$$

where c is a constant depending on α.

Proof. Assume first that $U^0 \in D(\mathcal{A}_\alpha^2)$; then we know that the solution $U(t) = \exp(-t\mathcal{A}_\alpha)U^0 = (u_1, u_2, v_1, v_2)$ of (10) is in $\mathcal{C}([0, +\infty); D(\mathcal{A}_\alpha^2)) \cap \mathcal{C}^1([0, +\infty); D(\mathcal{A}_\alpha)) \cap \mathcal{C}^2([0, +\infty); \mathcal{H})$. Hence $U = (u_1, u_2, v_1, v_2)$ satisfies

$$\begin{cases} v_1 = u_1', \quad v_2 = u_2', \\ u_1'' + A_1 u_1 + B u_1' + \alpha P u_2 = 0 \quad \text{in } H, \\ u_2'' + A_2 u_2 + \alpha P^\star u_1 = 0 \quad \text{in } H. \end{cases} \tag{23}$$

We now evaluate the term

$$\begin{aligned} K = \int_S^T & (u_1'' + A_1 u_1 + B u_1' + \alpha P u_2,\ Cu_2)_H \\ & - (Cu_2'' + CA_2 u_2 + \alpha C P^\star u_1,\ P_0 u_1)_H\ dt = 0. \end{aligned} \tag{24}$$

Then we have $K = K_1 + K_2 + K_3$, where

$$K_1 = \int_S^T (u_1'', Cu_2)_H - (P_0u_1, Cu_2'')_H \ dt,$$
$$K_2 = \int_S^T (A_1u_1 + Bu_1', Cu_2)_H - (CA_2u_2, P_0u_1)_H \ dt,$$
$$K_3 = \alpha \int_S^T (Pu_2, Cu_2)_H - (CP^\star u_1, P_0u_1)_H \ dt.$$

We first consider the term K_1.

Thanks to the regularity of the solution, we have that $u_i \in \mathcal{C}^2([0,+\infty); V_i)$ for $i = 1, 2$. Hence, $P_0(u_1'') = (P_0u_1)''$, so that we can rewrite K_1 as

$$K_1 = \int_S^T (u_1'' - P_0u_1'', Cu_2)_H \ dt + [(P_0u_1', Cu_2)_H - (P_0u_1, Cu_2')_H]_S^T.$$

Hence we have for all $\varepsilon > 0$

$$|K_1| \leq \frac{1}{2\varepsilon} \int_S^T ||u_1'' - P_0u_1''||_H^2 \ dt + \frac{\varepsilon}{2} \int_S^T ||Cu_2||_H^2 \ dt$$
$$+ c \sum_{i=1,2} (||u_i'(T)||_H^2 + ||u_i(T)||_H^2) + c \sum_{i=1,2} (||u_i'(S)||_H^2 + ||u_i(S)||_H^2).$$

On the other hand, since we have

$$||u_i'||_H^2 + ||u_i||_H^2 \leq ||u_i'||_H^2 + \beta_i^2 ||u_i||_{V_i}^2,$$

we deduce, also using Proposition 2.2, that

$$\sum_{i=1,2} (||u_i'(\cdot)||_H^2 + ||u_i(\cdot)||_H^2) \leq 2 \max\left(1, \frac{\max(\beta_1^2, \beta_2^2)}{c_1^2(\alpha)}\right) E(U(\cdot)),$$

where the constant $c_1(\alpha)$ is defined as in Proposition 2.2. Together with hypothesis (H2), this implies that for all $\varepsilon > 0$

$$|K_1| \leq \frac{\beta}{2\varepsilon} \int_S^T \langle Bu_1'', u_1'' \rangle_{V_1', V_1} \ dt + \frac{\varepsilon}{2} \int_S^T ||Cu_2||_H^2 \ dt + c(E(U(T)) + E(U(S))). \tag{25}$$

On the other side, we know from Proposition 2.3 that

$$E'(U(t)) = -\langle Bu_1', u_1' \rangle_{V_1', V_1}$$

for all $U^0 \in D(\mathcal{A}_\alpha)$, so that we have

$$E'(U'(t)) = -\langle Bu_1'', u_1'' \rangle_{V_1', V_1}$$

for all $U^0 \in D(\mathcal{A}_\alpha^2)$. Hence, using this last relation in (25), together with the fact that $E(U(\cdot))$ and $E(U'(\cdot))$ are nonnegative and nonincreasing, we obtain

$$|K_1| \leq \frac{\beta}{2\varepsilon} E(U'(S)) + \frac{\varepsilon}{2} \int_S^T ||Cu_2||_H^2 \ dt + cE(U(S)). \tag{26}$$

We now consider the term K_2. Since $U^0 \in D(\mathcal{A}_\alpha^2)$, and thanks to the hypothesis (H3), we have $u_2 \in D(A_2) \subset V_0$ and $Cu_2 \in D(A_0)$. Therefore, since (H2) holds, we have

$$(A_1u_1 + Bu_1', Cu_2)_H = \langle A_1u_1 + Bu_1', i(Cu_2)\rangle_{V_1',V_1} = \langle A_1u_1, i(Cu_2)\rangle_{V_1',V_1}.$$

On the other hand, by definition of P_0 and A_0, we have

$$\begin{aligned}\langle A_1u_1, i(Cu_2)\rangle_{V_1',V_1} &= \langle A_1 i(P_0u_1), i(Cu_2)\rangle_{V_1',V_1} \\ &= \langle A_0P_0u_1, Cu_2\rangle_{V_0',V_0} = \langle A_0Cu_2, P_0u_1\rangle_{V_0',V_0}.\end{aligned}$$

Moreover, since $Cu_2 \in D(A_0)$, we have $A_0Cu_2 \in H$. Hence

$$\langle A_0Cu_2, P_0u_1\rangle_{V_0',V_0} = (A_0Cu_2, P_0u_1)_H,$$

so that

$$(A_1u_1 + Bu_1', Cu_2)_H = (A_0Cu_2, P_0u_1)_H.$$

Hence, thanks once again to assumption (H3), we have finally

$$K_2 = 0. \tag{27}$$

We now turn to the term K_3. Thanks to assumption (H4), we have

$$K_3 \geq \alpha\gamma \int_S^T ||u_2||_H^2 \, dt - \alpha c \int_S^T ||u_1||_H^2 \, dt. \tag{28}$$

Using (26)–(28) in (24), we obtain for all $\varepsilon > 0$

$$\begin{aligned}\alpha\gamma \int_S^T ||u_2||_H^2 \, dt \leq \alpha c \int_S^T ||u_1||_H^2 \, dt + \frac{\beta}{2\varepsilon} E(U'(S)) \\ + \frac{\varepsilon ||C||^2}{2} \int_S^T ||u_2||_H^2 \, dt + cE(U(S)).\end{aligned}$$

Choosing $\varepsilon = \alpha\gamma/||C||^2$ and using assumption (H4), we obtain

$$\begin{aligned}&\frac{\alpha\gamma}{2} \int_S^T ||u_2||_H^2 \, dt \leq \alpha c \int_S^T ||u_1||_H^2 \, dt + c(E(U'(S)) + E(U(S))) \\ &\quad \forall\ 0 \leq S \leq T,\ U^0 \in D(\mathcal{A}_\alpha^2),\end{aligned} \tag{29}$$

where c is a constant depending on α. By a density argument, we deduce that (29) holds for every $U^0 \in D(\mathcal{A}_\alpha)$. □

From now on, for $U(t) = \exp(-t\mathcal{A}_\alpha)U^0 = (u_1, u_2, v_1, v_2)$ a solution of (10), we set $e_i((u_i, v_i)(t)) = e_i(t)$ for $i = 1, 2$, where $e_i(t)$ is defined by (14).

LEMMA 3.9. *Assume the hypotheses of Theorem* 3.3. *Then there exists* $\alpha_1 \in (0, \alpha_0]$ *such that for all* $0 < |\alpha| < \alpha_1$ *and all* $U^0 = (u_1^0, u_2^0, u_1^1, u_2^1) \in D(\mathcal{A}_\alpha)$ *the solution* $U(t) = \exp(-t\mathcal{A}_\alpha)U^0 = (u_1, u_2, v_1, v_2)$ *of* (10) *satisfies*

$$\int_S^T e_1((u_1, v_1)(t)) \, dt \leq c(\alpha)(E(U(S)) + E(U'(S))) \quad \forall\ 0 \leq S \leq T.$$

Proof. Assume that $U^0 \in D(\mathcal{A}_\alpha)$; then we know that the solution $U(t)$ of (10) is in $\mathcal{C}([0,+\infty);D(\mathcal{A}_\alpha)) \cap \mathcal{C}^1([0,+\infty);\mathcal{H})$. Hence $U=(u_1,u_2,v_1,v_2)$ satisfies

$$\begin{cases} (u_1,v_1)' + \mathcal{A}_1(u_1,v_1) = -\alpha(0,Pu_2), \\ (u_1,v_1)(0) = (u_1^0,u_1^1) \in \mathcal{A}_1. \end{cases}$$

Moreover, we have, in particular, that $Pu_2 \in \mathcal{C}^1([0,+\infty);H)$. Hence, thanks to assumption (H1), we have

$$\begin{aligned} \int_S^T e_1((u_1,v_1)(t))\ dt &\le \gamma_1(e_1((u_1,v_1)(S)) + e_1((u_1,v_1)(T))) \\ &\quad + \alpha^2\gamma_2||P||^2 \int_S^T ||u_2||_H^2\ dt + \gamma_3 \int_S^T \langle Bv_1,v_1\rangle_{V_1',V_1}\ dt. \end{aligned} \tag{30}$$

Using (29) and (17) in this last relation, we obtain

$$\begin{aligned} &\int_S^T e_1((u_1,v_1)(t))\ dt \\ &\quad \le \gamma_1(e_1((u_1,v_1)(S)) + e_1((u_1,v_1)(T))) + \gamma_3(E(U(S)) - E(U(T))) \\ &\quad + \frac{2\alpha\gamma_2||P||^2}{\gamma}\left(\alpha\ c\ \int_S^T ||u_1||_H^2\ dt + cE(U'(S)) + cE(U(S))\right) \quad \forall\ 0 \le S \le T, \end{aligned} \tag{31}$$

so that for

$$0 < |\alpha| < \min\left(\alpha_0, \sqrt{\frac{\gamma}{4\gamma_2 c\beta_1^2||P||^2}}\right) = \alpha_1$$

we have

$$\begin{aligned} 0 &\le \left(1 - \frac{4\gamma_2\alpha^2 c\beta_1^2||P||^2}{\gamma}\right)\int_S^T e_1(u_1,v_1)(t))\ dt \\ &\le \gamma_1(e_1((u_1,v_1)(S)) + e_1((u_1,v_1)(T))) \\ &\quad + c(\alpha)(E(U'(S)) + E(U(S))) \quad \forall\ 0 \le S \le T. \end{aligned}$$

Hence, thanks to Proposition 2.4, we have

$$\int_S^T e_1((u_1,v_1)(t))\ dt \le c(\alpha)(E(U(S)) + E(U'(S))) \quad \forall\ 0 \le S \le T, \tag{32}$$

where $c(\alpha)$ is a constant that depends on α but is bounded with respect to it for any $\alpha \in [r_1,r_2]$, for any $0 < r_1 < r_2 < \alpha_1$. In all of what follows, we will denote by $c(\alpha)$ a generic constant verifying these properties. ∎

LEMMA 3.10. *Assume the hypotheses of Theorem* 3.3. *Then for all* $0 < |\alpha| < \alpha_0$ *and all* $U^0 = (u_1^0,u_2^0,u_1^1,u_2^1) \in D(\mathcal{A}_\alpha^2)$ *the solution* $U(t) = \exp(-t\mathcal{A}_\alpha)U^0 = (u_1,u_2,v_1,v_2)$ *of* (10) *satisfies*

$$\int_S^T e_2((u_2,v_2)(t))\ dt \le c(\alpha)(E(U(S)) + E(U'(S)) + E(U''(S))) \quad \forall\ 0 \le S \le T.$$

Proof. Since (29) holds for every $U^0 \in D(\mathcal{A}_\alpha)$, we deduce that for any $U^0 \in D(\mathcal{A}_\alpha^2)$ we have

$$\begin{aligned} \frac{\alpha\gamma}{2}\int_S^T ||u_2'||_H^2\ dt &\le \alpha c \int_S^T ||u_1'||_H^2\ dt + c(E(U'(S)) + E(U''(S))) \\ &\quad \forall\ 0 \le S \le T,\ U^0 \in D(\mathcal{A}_\alpha^2). \end{aligned} \tag{33}$$

On the other hand, thanks to the third relation in (23), we have

$$\int_S^T (u_2'' + A_2 u_2 + \alpha P^\star u_1, u_2)_H \; dt = 0,$$

so that we have

$$\int_S^T e_2((u_2, v_2)(t)) \; dt = \int_S^T ||u_2'||_H^2 \; dt - \frac{1}{2}[(u_2', u_2)_H]_S^T - \frac{\alpha}{2} \int_S^T (u_1, Pu_2)_H \; dt.$$

Hence, using (32) and (33) in this last relation, we obtain

$$\int_S^T e_2((u_2, v_2)(t)) \; dt \le c(\alpha) \; (E(U(S)) + E(U'(S)) + E(U''(S))) \quad \forall \; 0 \le S \le T. \tag{34}$$

□

Proof of Theorem 3.3. Combining (32) and (33) with the inequality of Proposition 2.4, we finally obtain for $0 < |\alpha| < \alpha_1$

$$\int_S^T E(U(t))\, dt \le c(\alpha)(E(U(S)) + E(U'(S)) + E(U''(S))) \quad \forall\, 0 \le S \le T, \; U^0 \in D(\mathcal{A}_\alpha^2).$$

Now applying Theorem 3.1 with $k = 2$, we deduce that

$$E(U(t)) \le \frac{c(\alpha)}{t^n} \left(\sum_{p=0}^{2n} E(U^{(p)}(0)) \right) \quad \forall \; t > 0, \; U^0 \in D(\mathcal{A}_\alpha^{2n}).$$

The strong stability result follows easily, thanks to the above inequality for $n = 1$ and since $\exp(-t\mathcal{A}_\alpha)$ is a contraction semigroup on $\mathcal{H}$, and $D(\mathcal{A}_\alpha^2)$ is dense in $\mathcal{H}$. □

4. Applications.

4.1. The case $A_0 = A_2$. In all of what follows, Ω is a nonempty bounded open set in $\mathbb{R}^N$ having a boundary Γ of class C^2, $\{\Gamma_0, \Gamma_1\}$ is a partition of Γ such that $\overline{\Gamma}_0 \cap \overline{\Gamma}_1 = \emptyset$, and x_0 is a point in $\mathbb{R}^N$ such that $m \cdot \nu \le 0$ on Γ_0 and $m \cdot \nu \ge \beta > 0$ on Γ_1, where $m(x) = x - x_0$. We set $\sup_\Omega ||m|| = R$. Moreover, we set $H^p_{\Gamma_0}(\Omega) = \{u \in H^p(\Omega), u = \cdots = \frac{\partial^p u}{\partial \nu^p} = 0 \text{ on } \Gamma_0\}$.

Coupled wave equations with the same speed of propagation. We consider the system

$$\begin{cases} u_{1,tt} - \Delta u_1 + \alpha u_2 = 0 \quad \text{in} \quad \Omega \times (0, \infty), \\ u_{2,tt} - \Delta u_2 + \alpha u_1 = 0 \quad \text{in} \quad \Omega \times (0, \infty), \\ u_1 = u_2 = 0 \quad \text{on} \quad \Sigma_0 = \Gamma_0 \times (0, \infty), \\ \dfrac{\partial u_1}{\partial \nu} + a u_1 + \ell u_{1,t} = 0, u_2 = 0 \quad \text{on} \quad \Sigma_1 = \Gamma_1 \times (0, \infty), \\ (u_1, u_{1,t})(0) = (u_1^0, u_1^1), (u_2, u_{2,t})(0) = (u_2^0, u_2^1) \quad \text{on} \quad \Omega, \end{cases} \tag{35}$$

where

$$a = (N-1)m \cdot \frac{\nu}{2R^2}, \qquad l = m \cdot \frac{\nu}{R}. \tag{36}$$

For clarity, we will assume that $a \neq 0$ or $\text{meas}(\Gamma_0) > 0$, where the measure stands for the Lebesgue measure. We set $H = L^2(\Omega)$ and $V_1 = H^1_{\Gamma_0}(\Omega)$, equipped, respectively, with the L^2 scalar product and the scalar product $(u,z)_{V_1} = \int_\Omega \nabla u \cdot \nabla z + \int_{\Gamma_1} auz$ and the corresponding norms. Moreover, we set $V_2 = H^1_0(\Omega)$, equipped with the scalar product $(u,z)_{V_2} = \int_\Omega \nabla u \cdot \nabla z$ and the associated norm. We define the duality mappings A_1 and A_2 as in section 2. Moreover, we define a continuous linear operator B from V_1 to V_1' by

$$\langle Bu,\ z\rangle_{V_1',V_1} = \int_{\Gamma_1} \ell uz \, d\gamma.$$

Then B satisfies (9). We also set $P = P^\star = \text{Id}_H$. Then system (35) can be rewritten under the form (10) with the above notation. The energy of a solution $U = (u_1, u_2, v_1, v_2)$ is then given by

$$E(U(t)) = \frac{1}{2}(||u_1||^2_{V_1} + ||u_2||^2_{V_2} + ||u_{1,t}||^2_H + ||u_{2,t}||^2_H) + \alpha(u_1,u_2)_H. \tag{37}$$

To prove polynomial decay of the solutions, we need only to check that the assumptions of Corollary 3.5 are satisfied. For the sake of simplicity we will assume that $N \geq 3$. We begin with assumption (H1). Let (u_1, v_1) be a solution of the system considered in hypothesis (H1); then u_1 satisfies

$$\begin{cases} u_{1,tt} - \Delta u_1 = f_1 \quad \text{in} \quad \Omega \times (0,\infty), \\ u_1 = 0 \quad \text{on} \quad \Sigma_0 = \Gamma_0 \times (0,\infty), \\ \dfrac{\partial u_1}{\partial \nu} + au_1 + \ell u_{1,t} = 0 \quad \text{on} \quad \Sigma_1 = \Gamma_1 \times (0,\infty), \\ (u_1, u_{1,t})(0) = (u_1^0, u_1^1) \in \mathcal{H}_1 \quad \text{on} \quad \Omega. \end{cases} \tag{38}$$

Then, proceeding as in [18, Theorem 8.6], we use the multiplier $Mu_1 = m \cdot \nabla u_1 + \frac{(N-1)}{2} u_1$ in the first equation of (38). We obtain the identity

$$\begin{aligned} &\int_S^T e_1((u_1,v_1)(t))\, dt \\ &\quad = \int_S^T (f_1, Mu_1)_H \, dt + [(u_{1,t}, Mu_1)_H]_T^S + \frac{1}{2}\int_S^T \int_{\Gamma_0} m \cdot \nu \left|\frac{\partial u_1}{\partial \nu}\right|^2 d\gamma \, dt \\ &\quad + \int_S^T \int_{\Gamma_1} \frac{m\cdot\nu}{2}(|u_{1,t}|^2 - |\nabla u_1|^2 + bu_1^2 - 2(bu_1 + ku_{1,t})Mu_1)\, d\gamma \, dt. \end{aligned} \tag{39}$$

Now, recalling that the following three inequalities hold for all $t \geq 0$ (see [18, pp. 106–108]),

$$\begin{aligned} |(u_{1,t}, Mu_1)_H| &\leq e_1((u_1,v_1)(t)), \\ |u_{1,t}|^2 - |\nabla u_1|^2 + bu_1^2 - 2(bu_1 + ku_{1,t})Mu_1 &\leq 2|u_{1,t}|^2 \text{ on } \Gamma_1, \\ ||Mu_1||^2_H &\leq 2R^2 e_1((u_1,v_1)(t)), \end{aligned}$$

and using them in (39), we deduce that for all $\varepsilon > 0$ we have

$$\int_S^T e_1((u_1,v_1),t)\,dt \le \frac{R}{1-\varepsilon R^2}(e_1((u_1,v_1)(S)) + e_1((u_1,v_1)(T))) + \frac{1}{2\varepsilon(1-\varepsilon R^2)}\int_S^T ||f_1||_H^2\,dt + \frac{R}{1-\varepsilon R^2}\int_S^T \langle Bu_{1,t}, u_{1,t}\rangle_{V_1',V_1}\,dt. \tag{40}$$

Now choosing $\varepsilon = 1/R^2$, we deduce that the hypothesis (H1) is satisfied with $\gamma_1 = \gamma_3 = 2R$ and $\gamma_2 = 1/2R^2$.

We now check hypothesis (H2). We set $V_0 = V_2$. For the sake of clarity, we identify $i(\phi)$ with ϕ for $\phi \in V_0$ (where i is the canonical injection from V_0 in V_1). We remark that the first equality in assumption (H2) is trivially satisfied, thanks to the definition of B and V_0. We define P_0 and A_0 as in section 2. Then P_0u_1 is the weak solution of

$$\begin{cases} -\Delta P_0 u_1 = -\Delta u_1 \quad \text{in} \quad \Omega, \\ P_0 u_1 \in V_0, \end{cases}$$

and A_0 is defined by

$$\langle A_0\phi, \psi\rangle_{V_0',V_0} = \int_\Omega \nabla\phi\cdot\nabla\psi\,dx \qquad \forall\psi,\ \phi\in V_0. \tag{41}$$

We now check the second relation in (H2). For this, we set $z = u_1 - P_0u_1$. Therefore, z is the weak solution of

$$\begin{cases} -\Delta z = 0 \quad \text{in} \quad \Omega, \\ z = u_1 \quad \text{on} \quad \Gamma. \end{cases}$$

Notice that this function has been introduced by [9] in a different framework. By classical results on elliptic theory, we deduce that there exists a constant $c > 0$ such that

$$||z||_H \le c||u_{1|\Gamma_1}||_{L^2(\Gamma_1)},$$

so that, since $m\cdot\nu \ge \delta > 0$ and $\ell = \frac{m\cdot\nu}{R}$ on Γ_1, there exists $\beta > 0$ such that

$$||z||_H^2 \le \beta\ \langle Bu_1, u_1\rangle_{V_1',V_1} \qquad \forall u_1 \in V_1.$$

On the other hand, since $V_0 = V_2$ and (41) holds, we deduce that assumption (H3)$'$ is satisfied. In addition, since $P = \mathrm{Id}_H$, assumption (H4)$'$ is verified with $\gamma = 1$. Now applying Corollary 3.5, we deduce the following result.

THEOREM 4.1. *There exists an $\alpha_1 \in (0,\alpha_0]$ such that for all $0 < |\alpha| < \alpha_1$ the solution $U(t) = \exp(-\mathcal{A}_\alpha t)U^0$ of (35) satisfies*

$$E(U(t)) \le \frac{c}{t^n}\sum_{p=0}^{2n} E(U^{(p)}(0)) \qquad \forall\ t > 0,\ U^0 \in D(\mathcal{A}_\alpha^{2n}).$$

Moreover, strong stability holds in the energy space $\mathcal{H} = V_1 \times V_2 \times H^2$.

Coupled Kirchhoff–Petrowsky plates. Let $N = 2$, and assume that Ω is a nonempty bounded open set in $\mathbb{R}^N$ having a boundary Γ of class C^4. We assume as before that

$\{\Gamma_0, \Gamma_1\}$ is a partition of Γ such that $\overline{\Gamma}_0 \cap \overline{\Gamma}_1 = \emptyset$, and x_0 is a point in $\mathbb{R}^N$ such that $m \cdot \nu \leq 0$ on Γ_0 and $m \cdot \nu \geq \beta > 0$ on Γ_1, where $m(x) = x - x_0$. We set $\sup_\Omega ||m|| = R$.

We consider the system

$$
\left\{
\begin{array}{l}
u_{1,tt} + \Delta^2 u_1 + \alpha u_2 = 0 \quad \text{in} \quad \Omega \times (0, +\infty), \\
u_{2,tt} + \Delta^2 u_2 + \alpha u_1 = 0 \quad \text{in} \quad \Omega \times (0, +\infty), \\
u_1 = u_2 = 0 = \dfrac{\partial u_1}{\partial \nu} = \dfrac{\partial u_2}{\partial \nu} \quad \text{on} \quad \Sigma_0 = \Gamma_0 \times (0, +\infty), \\
\Delta u_1 + (1-\mu) B_1 u_1 = -km \cdot \nu \dfrac{\partial u_{1,t}}{\partial \nu}, \qquad u_2 = 0 = \dfrac{\partial u_2}{\partial \nu} \quad \text{on} \quad \Sigma_1, \\
\dfrac{\partial \Delta u_1}{\partial \nu} + (1-\mu) \dfrac{\partial B_2 u_1}{\partial \tau} = \ell m \cdot \nu u_{1,t} \quad \text{on} \quad \Sigma_1, \\
(u_1, u_{1,t})(0) = (u_1^0, u_1^1), \quad (u_2, u_{2,t})(0) = (u_2^0, u_2^1) \quad \text{on} \quad \Omega.
\end{array}
\right. \tag{42}
$$

We assume that $\text{meas}(\Gamma_0) > 0$, that the functions k and ℓ are continuous functions on Γ_1, and that there exist constants k_0, k_1 ℓ_0, and ℓ_1 such that

$$k_1 \geq k \geq k_0 > 0, \quad \ell_1 \geq \ell \geq \ell_0 > 0 \quad \text{on} \quad \Gamma_1.$$

We refer the reader to [22] for more details on this model. The constant $\mu \in (0, 1/2)$ is the Poisson coefficient, and the boundary operators B_1 and B_2 are defined by

$$
\begin{aligned}
B_1 v &= 2\nu_1 \nu_2 v_{xy} - \nu_1^2 v_{yy} - \nu_2^2 v_{xx}, \\
B_2 v &= (\nu_1^2 - \nu_2^2) v_{xy} + \nu_1 \nu_2 (v_{xx} - v_{yy}),
\end{aligned}
$$

where the subscripts x and y denote the partial derivatives with respect to the first and second components of the space variable. We set $H = L^2(\Omega)$ (equipped with the usual norm and scalar product) and $V_1 = H^2_{\Gamma_0}(\Omega)$, equipped with the scalar product

$$(u, z)_{V_1} = \int_\Omega (u_{xx} z_{xx} + u_{yy} z_{yy} + \mu(u_{xx} z_{yy} + u_{yy} z_{xx}) + 2(1-\mu) u_{xy} z_{xy})\, dx\, dy$$

and the associated norm. Moreover, we set $V_2 = H_0^2(\Omega)$, equipped with the scalar product

$$(u, z)_{V_2} = \int_\Omega \Delta u \Delta z \, dx \, dy$$

and the associated norm. We define the duality mappings A_1, A_2 as in section 2. Moreover, we define a linear continuous operator B from V_1 to V_1' by

$$\langle Bu, z \rangle_{V_1', V_1} = \int_{\Gamma_1} m \cdot \nu \left(\ell \, uz + k \, \frac{\partial u}{\partial \nu} \frac{\partial z}{\partial \nu} \right) d\gamma.$$

Then B satisfies (9). We also set $P = P^\star = \text{Id}_H$. Then the system (42) can be rewritten under the form (10) with the above notation. The energy of a solution $U = (u_1, u_2, v_1, v_2)$ is then defined as in (37). We now check the assumptions of Corollary 3.5. We begin with assumption (H1). Let (u_1, v_1) be a solution of the system considered in hypothesis (H1); then u_1 satisfies

$$
\left\{
\begin{array}{l}
u_{1,tt} + \Delta^2 u_1 = f_1 \quad \text{in} \quad \Omega \times (0, \infty), \\
u_1 = \dfrac{\partial u_1}{\partial \nu} = 0 \quad \text{on} \quad \Sigma_0 = \Gamma_0 \times (0, \infty), \\
\Delta u_1 + (1-\mu) B_1 u_1 = -km \cdot \nu \dfrac{\partial u_{1,t}}{\partial \nu} \quad \text{on} \quad \Sigma_1, \\
\dfrac{\partial \Delta u_1}{\partial \nu} + (1-\mu) \dfrac{\partial B_2 u_1}{\partial \tau} = \ell m \cdot \nu u_{1,t} \quad \text{on} \quad \Sigma_1, \\
(u_1, u_{1,t})(0) = (u_1^0, u_1^1) \quad \text{on} \quad \Omega.
\end{array}
\right. \tag{43}
$$

We then proceed as in [22]. Consider the relation

$$\int_S^T (u_{1,tt} + \Delta^2 u_1, m \cdot \nabla u_1)_H \, dt = \int_S^T (f_1, m \cdot \nabla u_1)_H \, dt. \tag{44}$$

This leads to the inequality (see [22])

$$\begin{aligned} 2 \int_S^T & e_1((u_1, v_1)(t)) \, dt \\ &= \int_S^T (f_1, m \cdot \nabla u_1)_H \, dt + [(u_{1,t}, m \cdot \nabla u_1)_H]_T^S + \frac{1}{2} \int_S^T \int_{\Gamma_0} m \cdot \nu (\Delta u_1)^2 \, d\gamma \, dt \\ &\quad + \frac{1}{2} \int_S^T \int_{\Gamma_1} m \cdot \nu |u_{1,t}|^2 \, d\gamma \, dt \\ &\quad - \int_S^T \int_{\Gamma_1} m \cdot \nu \left(\ell u_{1,t} m \cdot \nabla u_1 + k \frac{\partial u_{1,t}}{\partial \nu} \frac{\partial m \cdot \nabla u_1}{\partial \nu} \right) d\gamma \, dt \\ &\quad - \frac{1}{2} \int_S^T \int_{\Gamma_1} m \cdot \nu (u_{1xx}^2 + u_{1yy}^2 + 2\mu u_{1xx} u_{1yy} + 2(1-\mu) u_{1xy}^2) \, d\gamma \, dt. \end{aligned} \tag{45}$$

We now estimate the terms on the right-hand side of the above inequality. We first denote by λ_0 the smallest positive constant such that

$$||\nabla u||_H^2 \le \lambda_0^2 ||u||_{V_1}^2 \qquad \forall u \in H_{\Gamma_0}^2(\Omega), \tag{46}$$

and by μ_0 the smallest positive constant such that

$$\int_{\Gamma_1} |\nabla u|^2 \, d\gamma \le \mu_0^2 ||u||_{V_1}^2 \qquad \forall u \in H_{\Gamma_0}^2(\Omega). \tag{47}$$

Now, using (46) and for all $\varepsilon > 0$, we estimate the first term of (45) as follows:

$$\begin{aligned} \int_S^T (f_1, m \cdot \nabla u_1)_H \, dt &\le \frac{1}{2\varepsilon} \int_S^T ||f_1||_H^2 \, dt + \frac{\varepsilon R^2}{2} \int_S^T ||\nabla u_1||_H^2 \, dt \\ &\le \frac{1}{2\varepsilon} \int_S^T ||f_1||_H^2 \, dt + \varepsilon R^2 \lambda_0^2 \int_S^T e_1((u_1, v_1)(t)) \, dt. \end{aligned} \tag{48}$$

We now consider the second term of (45). Using (46) once again, we deduce that

$$|[(u_{1,t}, m \cdot \nabla u_1)_H]_T^S| \le R\lambda_0 (e_1((u_1, v_1)(S)) + e_1((u_1, v_1)(T))). \tag{49}$$

Since $m \cdot \nu \le 0$ on Γ_0, we estimate the third term of (45) as follows:

$$\frac{1}{2} \int_S^T \int_{\Gamma_0} m \cdot \nu (\Delta u_1)^2 \, d\gamma \, dt \le 0. \tag{50}$$

We now consider the fourth term in (45). We have, thanks to our assumptions on ℓ and k,

$$\begin{aligned} \frac{1}{2} \int_S^T \int_{\Gamma_1} m \cdot \nu |u_{1,t}|^2 \, d\gamma \, dt &\le \frac{1}{2\ell_0} \int_S^T \int_{\Gamma_1} m \cdot \nu \left(\ell |u_{1,t}|^2 + k \left| \frac{\partial u_{1,t}}{\partial \nu} \right|^2 \right) d\gamma \, dt \\ &= \frac{1}{2\ell_0} \int_S^T \langle B u_{1,t}, u_{1,t} \rangle_{V_1', V_1} \, dt. \end{aligned} \tag{51}$$

We now turn to the fifth term in (45). Using well-known estimates (see [22]), we obtain that for all $\delta > 0$

$$\begin{aligned}
&\left| \int_S^T \int_{\Gamma_1} m \cdot \nu \left(\ell u_{1,t} m \cdot \nabla u_1 + k \frac{\partial u_{1,t}}{\partial \nu} \frac{\partial m \cdot \nabla u_1}{\partial \nu} \right) d\gamma \ dt \right| \\
&\quad \leq \frac{1}{2\delta} \int_S^T \langle B u_{1,t}, u_{1,t} \rangle_{V_1',V_1} \ dt + \frac{\delta}{2} \int_S^T \int_{\Gamma_1} m \cdot \nu (R^2 \ell + 2k) |\nabla u_1|^2 \ d\gamma \ dt \\
&\quad + \frac{\delta R^2}{1-\mu} \int_S^T \int_{\Gamma_1} k m \cdot \nu (u_{1xx}^2 + u_{1yy}^2 + 2\mu u_{1xx} u_{1yy} + 2(1-\mu) u_{1xy}^2) d\gamma \ dt \qquad (52) \\
&\quad \leq \frac{1}{2\delta} \int_S^T \langle B u_{1,t}, u_{1,t} \rangle_{V_1',V_1} \ dt \ + \frac{\delta}{2} (R^2 \ell_1 + 2k_1) R \mu_0^2 \int_S^T ||u_1||_{V_1}^2 \ dt \\
&\quad + \frac{\delta R^2 k_1}{1-\mu} \int_S^T \int_{\Gamma_1} m \cdot \nu \left(u_{1xx}^2 + u_{1yy}^2 + 2\mu u_{1xx} u_{1yy} + 2(1-\mu) u_{1xy}^2 \right) d\gamma \ dt.
\end{aligned}$$

Using the estimates (48)–(52) in (45), we obtain for all $\varepsilon > 0$ and all $\delta > 0$

(53)

$$\begin{aligned}
&(2 - \varepsilon R^2 \lambda_0^2 - \delta (R^2 \ell_1 + 2k_1) R \mu_0^2) \int_S^T e_1((u_1, v_1), t) \ dt \leq \frac{1}{2\varepsilon} \int_S^T ||f_1||_H^2 \ dt \\
&\quad + R \lambda_0 (e_1((u_1, v_1)(S)) + e_1((u_1, v_1)(T))) + \left(\frac{1}{2\ell_0} + \frac{1}{2\delta} \right) \int_S^T \langle B u_{1,t}, u_{1,t} \rangle_{V_1',V_1} \ dt \\
&\quad + \left(\frac{\delta R^2 k_1}{1-\mu} - \frac{1}{2} \right) \int_S^T \int_{\Gamma_1} m \cdot \nu (u_{1xx}^2 + u_{1yy}^2 + 2\mu u_{1xx} u_{1yy} + 2(1-\mu) u_{1xy}^2) \ d\gamma \ dt.
\end{aligned}$$

Now choosing any δ such that

$$0 < \delta < \min \left(\frac{1-\mu}{2R^2 k_1}, \frac{2}{R \mu_0^2 (R^2 \ell_1 + 2k_1)} \right)$$

and then

$$0 < \varepsilon < \frac{2 - \delta R \mu_0^2 (R^2 \ell_1 + 2k_1)}{R^2 \lambda_0^2},$$

we prove that the desired inequality in assumption (H1) is satisfied with the corresponding γ_i, for $i = 1, 2, 3$. We now check assumption (H2). We set $V_0 = V_2$, and for the sake of clarity, we identify, as before, $i(\phi)$ with ϕ for any $\phi \in V_0$. Thanks to the definition of B and V_0, the first relation of (H2) is trivially satisfied. On the other hand, since the set of $\mathcal{C}^\infty$-functions on Ω with compact support in Ω is dense in V_0, we deduce that

$$\langle A_1 u, z \rangle_{V_1',V_1} = \int_\Omega \Delta u \Delta z \ dx \ dy \qquad \forall z \in V_0. \qquad (54)$$

Hence, for every $u_1 \in V_1$, $P_0 u_1$ is the weak solution of

$$\begin{cases} \Delta^2 (P_0 u_1) = \Delta^2 (u_1), \\ P_0 u_1 \in H_0^2(\Omega). \end{cases}$$

We set $z = u_1 - P_0 u_1$; then z is the weak solution of

$$\begin{cases} \Delta^2 z = 0, \\ z = u_1, \quad \dfrac{\partial z}{\partial \nu} = \dfrac{\partial u_1}{\partial \nu} \quad \text{on } \Gamma. \end{cases}$$

From classical results, based on elliptic theory, we know that there exists a constant $c > 0$ such that, for every $u_1 \in V_1$, we have

$$||z||_H^2 \le c \int_{\Gamma_1} \left(|u_1|^2 + \left| \frac{\partial u_1}{\partial \nu} \right|^2 \right) \, d\gamma.$$

Since $m \cdot \nu \ge \delta > 0$, $\ell \ge \ell_0 > 0$, and $k \ge k_0 > 0$ on Γ_1, we deduce that

$$||z||_H^2 \le C \langle Bu_1, u_1 \rangle_{V_1', V_1},$$

so that the second relation in (H2) is satisfied. Moreover, thanks to (54) and since $V_0 = V_2$ holds, assumption (H3)$'$ is satisfied. Finally, since $P = \mathrm{Id}_H$, assumption (H4)$'$ is trivially satisfied with $\gamma = 1$. Hence, we can apply Corollary 3.5. We deduce the following result.

THEOREM 4.2. *There exists an $\alpha_1 \in (0, \alpha_0]$ such that for all $0 < |\alpha| < \alpha_1$ the solution $U(t) = \exp(-\mathcal{A}_\alpha t) U^0$ of (42) satisfies*

$$E(U(t)) \le \frac{c}{t^n} \sum_{p=0}^{2n} E(U^{(p)}(0)) \qquad \forall\ t > 0,\ U^0 \in D(\mathcal{A}_\alpha^{2n}).$$

Moreover, strong stability holds in the energy space $\mathcal{H} = V_1 \times V_2 \times H^2$.

4.2. Case of different operators A_0 and A_2. In order to avoid loss of regularity of the solutions, we assume in this subsection that $\Gamma_0 = \emptyset$. Hence we do not treat the case of mixed boundary conditions. Nevertheless, the results are still valid for a more general situation (see [11, 12, 13]).

Coupled wave equations with different speed of propagation. We consider the following system:

$$\begin{cases} u_{1,tt} - c_1 \Delta u_1 + \alpha P u_2 = 0 \quad \text{in} \quad \Omega \times (0, \infty), \\ u_{2,tt} - c_2 \Delta u_2 + \alpha P^\star u_1 = 0 \quad \text{in} \quad \Omega \times (0, \infty), \\ \dfrac{\partial u_1}{\partial \nu} + a u_1 + \ell u_{1,t} = 0, \quad u_2 = 0 \quad \text{on} \quad \Sigma = \Gamma \times (0, \infty), \\ (u_1, u_{1,t})(0) = (u_1^0, u_1^1), \quad (u_2, u_{2,t})(0) = (u_2^0, u_2^1) \quad \text{on} \quad \Omega, \end{cases} \tag{55}$$

where $c_i > 0$, $i = 1, 2$, and

$$a = (N-1) m \cdot \frac{\nu}{2R^2}, \qquad l = \frac{m \cdot \nu}{R\sqrt{c_1}}. \tag{56}$$

We mainly keep the notation of section 4.1. We set $H = L^2(\Omega)$ and $V_1 = H^1(\Omega)$, equipped, respectively, with the L^2 scalar product and the scalar product $(u, z)_{V_1} = c_1 \int_\Omega \nabla u \cdot \nabla z + c_1 \int_\Gamma auz$ and the corresponding norms. Moreover, we set $V_2 = H_0^1(\Omega)$, equipped with the scalar product $(u, z)_{V_2} = c_2 \int_\Omega \nabla u \cdot \nabla z$ and the associated norm.

We define the duality mappings A_1 and A_2 as in section 2. Moreover, we define a continuous linear operator B from V_1 to V_1' by

$$\langle Bu,\ z\rangle_{V_1',V_1} = c_1 \int_\Gamma \ell u z \ d\gamma.$$

Then B satisfies (9). We assume for the moment that P is a given bounded operator in H. Then the system (55) can be rewritten under the form (10) with the above notation. The energy of a solution $U = (u_1, u_2, v_1, v_2)$ is then given by

$$E(U(t)) = \frac{1}{2}(||u_1||^2_{V_1} + ||u_2||^2_{V_2} + ||u_{1,t}||^2_H + ||u_{2,t}||^2_H) + \alpha(u_1, u_2)_H. \tag{57}$$

To prove polynomial decay of the solutions, we need to check only that the assumptions of Lemma 3.6 are satisfied. We begin with assumption (H1). Let (u_1, v_1) be a solution of the system considered in hypothesis (H1); then u_1 satisfies

$$\begin{cases} u_{1,tt} - c_1\Delta u_1 = f_1 \quad \text{in} \quad \Omega \times (0,\infty), \\ \dfrac{\partial u_1}{\partial \nu} + a u_1 + \ell u_{1,t} = 0 \quad \text{on} \quad \Sigma = \Gamma \times (0,\infty), \\ (u_1, u_{1,t})(0) = (u_1^0, u_1^1) \in \mathcal{H}_1 \quad \text{on} \quad \Omega. \end{cases} \tag{58}$$

Then, making the change of time variable $s = \sqrt{c_1}t$, we obtain a system similar to system (38), where f_1 is replaced by f_1/c_1, and ℓ is replaced by $\sqrt{c_1}\ell$. However, as will be seen later, hypothesis (H3) can be checked only for domains Ω which are N-dimensional intervals $\prod_{i=1}^N (a_i, b_i)$, where $a_i < b_i$, $i = 1, \dots, N$, and $N \le 3$. Of course, for such domains the boundary is no longer of class $\mathcal{C}^2$. Therefore we need to check that all the computations performed earlier are still valid in this case. For this, we make the following statement.

Important Remark. We recall that when Ω is a convex polygon (case $N = 2$) or polyhedron (case $N = 3$), Grisvard's results (see [11], [12], and also [6]) on the solution of the Poisson equation

$$-\Delta u = f \in L^2(\Omega) \quad \text{in } \Omega,$$

subjected to either Dirichlet, Neumann, or oblique boundary conditions in such domains, say that these solutions are in $H^2(\Omega)$. Moreover, thanks to the above regularity result, all the classical results for the wave equation, subjected to either Dirichlet, Neumann, or oblique boundary conditions (in particular, the well-known multiplier identity, which leads to the required estimate in our hypothesis (H1)), are still valid for convex polygons or polyhedra. Hence our hypothesis (H1) is verified when Ω is an N-dimensional interval $\prod_{i=1}^N (a_i, b_i)$, where $a_i < b_i$, $i = 1, \dots, N$, with $N \le 3$. (Let us remark that in that case the angles between corners as defined in Grisvard's results are all equal to $\pi/2$.)

We now check hypothesis (H2), assuming as before that Ω is an N-dimensional interval $\prod_{i=1}^N (a_i, b_i)$, where $a_i < b_i$, $i = 1, \dots, N$, and $N \le 3$. We set $V_0 = V_2$. For the sake of clarity, we identify $i(\phi)$ with ϕ for $\phi \in V_0$ (where i is the canonical injection from V_0 in V_1). We remark that the first equality in assumption (H2) is trivially satisfied, thanks to the definition of B and V_0. We define P_0 and A_0 as in section 2. Then $P_0 u_1$ is the weak solution of

$$\begin{cases} -\Delta P_0 u_1 = -\Delta u_1 \quad \text{in} \quad \Omega, \\ P_0 u_1 \in V_0. \end{cases}$$

We now check that the second relation in (H2) holds. We set $z = u_1 - P_0 u_1$. Then, thanks to the above important remark on the regularity of solutions of the Poisson equation subjected to oblique boundary conditions in convex polygons or polyhedra, we know that $u_1(t)$ is in $H^2(\Omega)$ for any positive t. On the other hand, thanks to similar results (the Dirichlet case), we know that $P_0 u_1$ is also in $H^2(\Omega)$. Therefore, $z(\cdot)$ is in $H^2(\Omega)$ for any positive t. Moreover, z is a strong solution of

$$\begin{cases} -\Delta z = 0 \quad \text{in} \quad \Omega, \\ z = u_1 \quad \text{on} \quad \Gamma. \end{cases}$$

In order to establish the second relation in (H2), we introduce the solution of the following Poisson equation:

$$\begin{cases} -\Delta w = z \quad \text{in} \quad \Omega, \\ w = 0 \quad \text{on} \quad \Gamma. \end{cases}$$

Then since $z \in L^2(\Omega)$ and thanks to Grisvard's results, we have that w is in $H^2(\Omega)$, so that the trace of the normal derivative of w on Γ is well defined as a function of $H^{1/2}(\Gamma)$. Moreover, we have

$$||w||_{V_1} \leq c||z||_H. \tag{59}$$

Multiplying the above equation by z, integrating by parts, and using the relation $\Delta z = 0$ in Ω and $z = u_1$ on Γ, we obtain the following equality:

$$\int_\Omega z^2 dx = \int_\Gamma u_1 \frac{\partial w}{\partial \nu} d\gamma. \tag{60}$$

Now using the classical multiplier $Mw = m \cdot \nabla w + \frac{N-1}{2} w$ and integrating by parts the expression

$$-\int_\Omega \Delta w M w dx,$$

we obtain that

$$\int_\Gamma m \cdot \nu \left| \frac{\partial w}{\partial \nu} \right|^2 d\gamma = \int_\Omega |\nabla w|^2 dx + 2 \int_\Omega z M w dx.$$

Using (59) in the above equality and since $m \cdot \nu \geq \beta > 0$ on Γ, we deduce that there exists a positive constant c such that

$$\int_\Gamma \left| \frac{\partial w}{\partial \nu} \right|^2 d\gamma \leq c||z||_H^2.$$

Using this last inequality in (60), we obtain

$$||z||_H \leq c||u_1||_{L^2(\Gamma)}.$$

Hence the second relation of (H2) is satisfied.

We now want to check assumption (H3). For this, we will use Lemma 3.6. As will be seen below, we can check this restrictive hypothesis only when Ω is an N-dimensional interval $\prod_{i=1}^N (a_i, b_i)$, where $a_i < b_i$, $i = 1, \ldots, N$, with $N \leq 3$.

We first note that we have, thanks to Grisvard's regularity results,

$$A_0 = -c_1\Delta, \qquad D(A_0) = H^2(\Omega) \cap H_0^1(\Omega),$$
$$A_2 = -c_2\Delta, \qquad D(A_2) = H^2(\Omega) \cap H_0^1(\Omega).$$

We denote by $A = -\Delta$ the unbounded operator in H with domain $D(A) = H^2(\Omega) \cap H_0^1(\Omega)$. Then, it is well known that there exists an orthonormal basis in H of eigenfunctions e_k of A associated to the eigenvalues $\lambda_k > 0$, for $k = 1, \dots$. Moreover, λ_k and e_k are such that

$$\begin{cases} -\Delta e_k = \lambda_k e_k & \text{in } \Omega, \\ e_k \in H_0^1(\Omega) \cap \mathcal{C}^\infty(\Omega). \end{cases}$$

Then, $\{e_k\}_{k=1}^\infty$ also forms an orthonormal basis of the unbounded operators A_0 and A_2, respectively. We denote by $\lambda_{i,k}$ the eigenvalue of the operator A_i associated to the eigenfunction e_k for $i = 0, 2$. When Ω is an N-dimensional interval, it is more convenient, as will be seen below, to consider $k = (k_1, \dots, k_N) \in (\mathbb{N}^\star)^N$. This notation will be used in all of what follows. We have the following result.

THEOREM 4.3. *Assume now that Ω is an N-dimensional interval $\prod_{i=1}^N (a_i, b_i)$, where $a_i < b_i$, $i = 1, \dots, N$, with $N \le 3$, and that there exists a positive integer k_0 such that $c_2 = k_0^2 c_1$. Moreover, assume that for $u = \sum_{k \in (N^\star)^N} u_k e_k$ in H, Pu is defined by*

$$Pu = \sum_{k \in (N^\star)^N} \delta_k u_k e_{k_0 k},$$

where the sequence of real numbers $(\delta_k)_k$ is such that there exist $\gamma > 0$ and δ with $0 < \gamma \le \delta_k \le \delta$. Then there exists an $\alpha_1 \in (0, \alpha_0]$ such that for all $0 < |\alpha| < \alpha_1$ the solution $U(t) = \exp(-\mathcal{A}_\alpha t)U^0$ of (55) satisfies

$$E(U(t)) \le \frac{c}{t^n} \sum_{p=0}^{2n} E(U^{(p)}(0)) \qquad \forall\ t > 0,\ U^0 \in D(\mathcal{A}_\alpha^{2n}).$$

Moreover, strong stability holds in the energy space $\mathcal{H} = V_1 \times V_2 \times H^2$.

Proof. Since Ω is an N-dimensional interval, we have for all $k = (k_1, \dots, k_N) \in (\mathbb{N}^\star)^N$

$$e_k(x_1, \dots, x_N) = \prod_{i=1}^N \sin\left(\frac{k_i \pi (x_i - a_i)}{b_i - a_i}\right),$$

whereas

$$\lambda_{i,k} = c_i \pi^2 \left(\sum_{i=1}^N \frac{k_i^2}{(b_i - a_i)^2}\right).$$

Thanks to the relation $c_2 = k_0^2 c_1$, we deduce that $\lambda_{2,k} = \lambda_{0,k_0 k}$ for all $k \in (N^\star)^N$, where $k_0 k = (k_0 k_1, \dots, k_0 k_N)$. Hence the assumptions of Lemma 3.6 are satisfied, and, in particular, assumption (H5), with the one-to-one application r defined on $(\mathbb{N}^\star)^N$ by $r(k) = k_0 k$ for k in $(\mathbb{N}^\star)^N$. We define C as in the proof of Lemma 3.6, that is

$$Cu = \sum_{k \in (N^\star)^N} u_k e_{r(k)}.$$

Then, thanks to the assumption on the sequence δ_k, we deduce that P satisfies (H4). Now applying Corollary 3.7, we conclude the proof. □

Remark. If the condition on the ratio of the two speeds of propagation is violated, one can show that, in some situations, the total energy does not decay to 0 at infinity. Indeed, assume that there exists a bounded operator C such that (H3) holds, but that C is not one-to-one, so that the assumption (H4) is not verified. This situation occurs, for instance, in the above example of two coupled wave equations, when $c_2/c_1 = p_0^2/q_0^2$, where $p_0 \in \mathbb{N}^\star$ and $q_0 > 1$ are integers with no common divisors. In this case, one can take the operator C defined by

$$Cu = \sum_{n \in (N^\star)^N} u_{nq_0} e_{np_0} \qquad \forall\ u = \sum_{k \in (N^\star)^N} u_k e_k \in H.$$

If P is any bounded operator on H that is not one-to-one (for instance, one can take $P = C$), then, setting $u_1 = 0$ and choosing any smooth function u_2 with values in the kernel of P such that

$$u_2'' + A_2 u_2 = 0,$$

the couple (u_1, u_2) is a solution of the full coupled system (10). Its energy is conserved, so it does not decay to zero at ∞ unless the initial conditions on u_2 and its first time derivative are zero. With the above choice of C, one can take, for instance, $u_2 = y(t)e_{(p_0,\dots,p_0)}$, where y satisfies the ordinary differential equation

$$y'' + \lambda_{2,(p_0,\dots,p_0)} y = 0$$

with nonzero initial data.

Coupled wave-Petrowsky equations. We consider the following system:

$$\text{(61)} \qquad \begin{cases} u_{1,tt} - \Delta u_1 + \alpha P u_2 = 0 \quad \text{in} \quad \Omega \times (0,\infty), \\ u_{2,tt} + \Delta^2 u_2 + \alpha P^\star u_1 = 0 \quad \text{in} \quad \Omega \times (0,\infty), \\ \dfrac{\partial u_1}{\partial \nu} + a u_1 + \ell u_{1,t} = 0, \quad u_2 = \Delta u_2 = 0 \quad \text{on} \quad \Sigma = \Gamma \times (0,\infty), \\ (u_1, u_{1,t})(0) = (u_1^0, u_1^1), \quad (u_2, u_{2,t})(0) = (u_2^0, u_2^1) \quad \text{on} \quad \Omega, \end{cases}$$

where

$$\text{(62)} \qquad a = (N-1)m \cdot \frac{\nu}{2R^2}, \qquad \ell = \frac{m \cdot \nu}{R}.$$

We mainly keep the notation of section 4.1. We set $H = L^2(\Omega)$ and $V_1 = H^1(\Omega)$, equipped, respectively, with the L^2 scalar product and the scalar product $(u,z)_{V_1} = \int_\Omega \nabla u \cdot \nabla z + \int_\Gamma auz$ and the corresponding norms. Moreover, we set $V_2 = H^2(\Omega) \cap H_0^1(\Omega)$, equipped with the scalar product $(u,z)_{V_2} = \int_\Omega \Delta u \cdot \Delta z$ and the associated norm. We define the duality mappings A_1 and A_2 as in section 2. Moreover, we define a continuous linear operator B from V_1 to V_1' by

$$\langle Bu,\ z\rangle_{V_1',V_1} = \int_\Gamma \ell u z \, d\gamma.$$

Then B satisfies (9). We assume for the moment that P is a given bounded operator in H. Then the system (61) can be rewritten under the form (10) with the above notation. The energy of a solution $U = (u_1, u_2, v_1, v_2)$ is then given by

$$\text{(63)} \qquad E(U(t)) = \frac{1}{2}(||u_1||_{V_1}^2 + ||u_2||_{V_2}^2 + ||u_{1,t}||_H^2 + ||u_{2,t}||_H^2) + \alpha(u_1, u_2)_H.$$

To prove polynomial decay of the solutions, we need to check only that the assumptions of Lemma 3.6 are satisfied. We first need to characterize the domain of the operator $A_2 = \Delta^2$, viewed as an unbounded operator in H. Now let u be given in $D(A_2)$. Then $u \in V_2$ and there exists a $f \in L^2(\Omega)$ such that

$$\int_\Omega \Delta u \Delta v dx = \int_\Omega f v dx \qquad \forall v \in V_2. \tag{64}$$

We set $w = \Delta u$. Since $u \in V_2$, we have $w \in L^2(\Omega)$. Moreover, thanks to (64), we have

$$\int_\Omega w \Delta v dx = \int_\Omega f v dx \qquad \forall v \in V_2. \tag{65}$$

Hence, since the space of $\mathcal{C}^\infty$-functions on Ω with compact support in Ω is continuously imbedded in V_2, we deduce that $\Delta w = f$ almost everywhere in Ω, so that $\Delta w \in L^2(\Omega)$. We cannot conclude directly because we have to prove that $w \in H_0^1(\Omega)$. For that, we proceed as follows.

We introduce the variational solution $\Theta \in H_0^1(\Omega)$ of

$$\Delta \Theta = f \quad \text{ in } \Omega.$$

Then, thanks to the above-mentioned Grisvard's results, we know that $\Theta \in V_2$. Using (65) and the definition of Θ, we easily deduce that

$$\int_\Omega w \Delta v dx = -\int_\Omega \nabla \Theta \cdot \nabla v dx = \int_\Omega \Theta \Delta v dx \qquad \forall v \in V_2, \tag{66}$$

so that we have

$$\int_\Omega (w - \Theta) \Delta v dx = 0 \qquad \forall v \in V_2. \tag{67}$$

We now choose as a test function v, the variational solution $v \in H_0^1(\Omega)$ of

$$\Delta v = w - \Theta \quad \text{ in } \Omega.$$

Then, thanks to the above-mentioned Grisvard's results, we know that $v \in V_2$. Now using (67) with this specially chosen v, we obtain that

$$\int_\Omega |w - \Theta|^2 dx = 0. \tag{68}$$

Hence we have $w = \Theta$ almost everywhere on Ω, so that $w = 0$ on Γ. Hence, thanks again to Grisvard's regularity results for the Poisson equation subjected to homogeneous Dirichlet boundary conditions, we have $w \in H^2(\Omega) \cap H_0^1(\Omega)$. Hence, we have proved that when $u \in D(A_2)$, then $u \in V_2$ satisfies

$$\Delta u = w \in H^2(\Omega) \cap H_0^1(\Omega) \quad \text{ in } \Omega.$$

Then, thanks to Grisvard's results (see, for instance, [12, section 2.7]) for convex polygons or polyhedra, we know that u is indeed in $H^3(\Omega)$, since the involved angles in an N-dimensional interval are all equal to $\pi/2$. Hence the traces of all second derivatives of u on Γ are well defined, and, moreover, we have

$$\Delta u = w = 0 \quad \text{ on } \Gamma. \tag{69}$$

We will now prove that u is indeed in $H^4(\Omega)$. We proceed as follows. Let us, for instance, assume that $N = 3$, and let us choose one of the six faces, denoted by Γ_3, of the surface Γ, the normal to which is along the x_3-direction. Then the tangential directions to Γ_3 are in the x_i-directions for $i = 1, 2$ on Γ_3. Since $u = 0$ on Γ, we deduce that

$$\frac{\partial^2 u}{\partial x_i^2} = 0 \quad \text{on } \Gamma_3 \quad \text{for } i = 1, 2.$$

Then, thanks to (69), we deduce that

$$\frac{\partial^2 u}{\partial x_3^2} = 0 \quad \text{on } \Gamma_3.$$

For each other face of Γ–Γ_3, the normal to it is either in the x_1- or x_2-directions; hence x_3 is always a tangential direction to it. Since $u = 0$ on Γ, we deduce that

$$\frac{\partial^2 u}{\partial x_3^2} = 0 \quad \text{on } \Gamma - \Gamma_3.$$

Now we set $z = \partial^2 u/\partial x_3^2$. Then z is the solution of

$$\Delta z = \frac{\partial^2 w}{\partial x_3^2} \in L^2(\Omega) \quad \text{in } \Omega,$$

with homogeneous boundary conditions on Γ. Hence from Grisvard's regularity results we deduce that z is in $H^2(\Omega)$. We proceed in a similar way for the other fourth order derivatives of u and show in this way that u is in $H^4(\Omega)$. Hence we have proved that $D(A_2) = \{u \in H^4(\Omega) \cap H_0^1(\Omega),\ \Delta u = 0 \text{ on } \Gamma\}$. Hence all the integrations by parts required for the definition of weak solutions of the above coupled wave-Petrowsky system are justified for data in the domain of the operator.

For the sake of clarity, we identify $i(\phi)$ with ϕ for $\phi \in V_0$ (where i is the canonical injection from V_0 in V_1). We define P_0 and A_0 as in section 2. From the former example on wave-wave equations with different speeds of propagation in an N-dimensional interval, with $N \leq 3$, we already know that the hypotheses (H1) and (H2) are satisfied. We now want to check assumption (H3). For this, we will use Lemma 3.6. We first note that we have

$$\begin{aligned} &A_0 = -\Delta, \quad D(A_0) = H^2(\Omega) \cap H_0^1(\Omega), \\ &A_2 = \Delta^2, \quad D(A_2) = \{u \in H^4(\Omega),\ u = \Delta u = 0 \text{ on } \Gamma\}. \end{aligned}$$

We have already remarked that there exists an orthonormal basis in H of eigenfunctions e_k of A_0 associated to the eigenvalues $\lambda_{0,k} > 0$, for $k \in (N^\star)^N$. Moreover, $\lambda_{0,k} > 0$ and e_k are such that

$$\begin{cases} -\Delta e_k = \lambda_{0,k} e_k & \text{in } \Omega, \\ e_k \in H_0^1(\Omega) \cap \mathcal{C}^\infty(\Omega). \end{cases}$$

Then $\{e_k\}_{k \in (N^\star)^N}$ also forms an orthonormal basis of the unbounded operator A_2, with the corresponding eigenvalues $\lambda_{2,k} = \lambda_{0,k}^2$. Then assumption (H5) of Lemma 3.6 is satisfied if and only if there exists a one-to-one application r on $(\mathbb{N}^\star)^N$ such that,

for all $k \in (\mathbb{N}^\star)^N$, $\lambda_{0,k}^2 = \lambda_{0,r(k)}$. In the case of N-dimensional intervals, we know explicitly the functions e_k, and thus we can prove the following result.

THEOREM 4.4. *Let Ω be an N-dimensional interval $\prod_{i=1}^N (a_i, b_i)$, where $a_i < b_i$, $i = 1, \dots, N$, with $N \leq 3$. Assume, moreover, that there exists a $d > 0$ for which $b_i - a_i = d$ for all $i = 1, \dots, N$ and such that $\frac{\pi}{d\sqrt{N}} \in N^\star$. In addition, assume that for $u = \sum_{k \in (N^\star)^N} u_k e_k$ in H, Pu is defined by*

$$Pu = \sum_{k \in (N^\star)^N} \delta_k u_k e_{r(k)},$$

where the sequence of real numbers $(\delta_k)_k$ is such that there exist $\gamma > 0$ and δ with $0 < \gamma \leq \delta_k \leq \delta$, and where r is the one-to-one application defined by $r(k) = (\ell, \dots, \ell)$ for $k \in (N^\star)^N$, where $\ell = \frac{\pi}{d\sqrt{N}} \sum_{i=1}^N k_i^2$. Then there exists an $\alpha_1 \in (0, \alpha_0]$ such that for all $0 < |\alpha| < \alpha_1$ the solution $U(t) = \exp(-\mathcal{A}_\alpha t) U^0$ of (61) satisfies

$$E(U(t)) \leq \frac{c}{t^n} \sum_{p=0}^{2n} E(U^{(p)}(0)) \qquad \forall\ t > 0,\ U^0 \in D(\mathcal{A}_\alpha^{2n}).$$

Moreover, strong stability holds in the energy space $\mathcal{H} = V_1 \times V_2 \times H^2$.

Proof. Since Ω is an N-dimensional interval, we have for all $k \in (\mathbb{N}^\star)^N$

$$e_k(x_1, \dots, x_N) = \prod_{i=1}^N \sin\left(\frac{k_i \pi (x_i - a_i)}{b_i - a_i}\right),$$

whereas

$$\lambda_{0,k} = (d^{-1}\pi)^2 \sum_{i=1}^N k_i^2.$$

Since, by hypothesis, $\pi d^{-1} N^{-1/2} \in \mathbb{N}^\star$, we deduce that $\lambda_{2,k} = \lambda_{0,r(k)}$ for all $k \in (\mathbb{N}^\star)^N$. Hence the assumptions of Lemma 3.6 are satisfied, and in particular assumption (H5), with the application r defined above. We define C as in the proof of Lemma 3.6, that is,

$$Cu = \sum_{k \in (\mathbb{N}^\star)^N} u_k e_{r(k)}.$$

Then, thanks to the assumption on the sequence δ_k, we deduce that P satisfies the assumption (H4). Now applying Corollary 3.7, we conclude the proof. ∎

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

REFERENCES

[1] F. ALABAU, *Stabilisation frontière indirecte de systèmes faiblement couplés*, C. R. Acad. Sci. Paris Sér. I Math., 328 (1999), pp. 1015–1020.

[2] F. ALABAU, P. CANNARSA, AND V. KOMORNIK, *Indirect internal stabilization of weakly coupled systems*, J. Evolution Equations, to appear.

[3] F. AMMAR-KHODJA, A. BADER, AND A. BENABDALLAH, *Dynamic stabilization of systems via decoupling techniques*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 577–593.

[4] A. BADER, *Quelques Risultats sur la Stabilisation des Systemes Couplies*, Ph.D. thesis, l'Université de Franche-Comté, Besançon, France, 2000.

[5] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[6] R. Bey, J. P. Loheac, and M. Moussaoui, *Singularities of the solution of a mixed problem for a general second order elliptic equation and boundary stabilization of the wave equation*, J. Math. Pures Appl. (9), 78 (1999), pp. 1043–1067.

[7] H. L. Brézis, *Analyse Fonctionnelle: Théorie et Applications*, Masson, Paris, 1983.

[8] A. Beyrath, *Stabilisation indirecte intense par un feedback localement distribué de systemes d'equations couplies*, C. R. Acad. Sci. Paris Sér. I Math., 333 (2001), pp. 451–456.

[9] F. Conrad and B. Rao, *Decay of solutions of the wave equation in a star-shaped domain with nonlinear boundary feedback*, Asymptotic Anal., 7 (1993), pp. 159–177.

[10] C. M. Dafermos, *On the existence and the asymptotic stability of solutions to the equations of linear thermoelasticity*, Arch. Ration. Mech. Anal., 29 (1968), pp. 241–271.

[11] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Monographs and Studies in Math. 24, Pitman, London, 1985.

[12] P. Grisvard, *Singularities in Boundary Value Problems*, Research Notes in Appl. Math. 22, Masson and Springer-Verlag, Paris, Berlin, 1992.

[13] P. Grisvard, *Contrôlabilité exacte des solutions de l'équation des ondes en présence de singularités*, J. Math. Pures Appl. (9), 68 (1989), pp. 215–259.

[14] B. Z. Guo and K. Y. Chan, *Riesz basis generation, eigenvalues distribution, and exponential stability for an Euler-Bernoulli beam with joint feedback control*, Rev. Mat. Complut., 14 (2001), pp. 1–24.

[15] A. Haraux, *Semi-Groupes Linéaires et Équations d'Évolution Linéaires Périodiques*, Publication 78011, Laboratoire d'Analyse Numérique, Université Pierre et Marie Curie, Paris, 1978.

[16] B. V. Kapitonov, *Uniform stabilization and exact controllability for a class of coupled hyperbolic systems*, Comput. Appl. Math., 15 (1996), pp. 199–212.

[17] J. U. Kim and Y. Renardy, *Boundary control of the Timoshenko beam*, SIAM J. Control Optim., 25 (1987), pp. 1417–1429.

[18] V. Komornik, *Exact Controllability and Stabilization: The Multiplier Method*, Collection RMA 36, Masson and John Wiley, Paris, Chichester, UK, 1994.

[19] V. Komornik and P. Loreti, *Ingham-type theorems for vector-valued functions and observability of coupled linear systems*, SIAM J. Control Optim., 37 (1998–1999), pp. 461–485.

[20] V. Komornik and B. Rao, *Boundary stabilization of compactly coupled wave equations*, Asymptotic Anal., 14 (1997), pp. 339–359.

[21] V. Komornik and E. Zuazua, *A direct method for the boundary stabilization of the wave equation*, J. Math. Pures Appl. (9), 69 (1990), pp. 33–54.

[22] J. E. Lagnese, *Boundary Stabilization of Thin Plates*, Studies in Appl. Math. 10, SIAM, Philadelphia, 1989.

[23] J. E. Lagnese and J.-L. Lions, *Modelling Analysis and Control of Thin Plates*, Recherches en Mathematiques Appliquees 6, Masson, Paris, 1988.

[24] I. Lasiecka, *Uniform decay rates for full von Karman system of dynamic thermoelasticity with free boundary conditions and partial boundary dissipation*, Comm. Partial Differential Equations, 24 (1999), pp. 1801–1847.

[25] I. Lasiecka and R. Triggiani, *Carleman estimates and exact boundary controllability for a system of coupled, nonconservative second-order hyperbolic equations*, Lecture Notes in Pure and Appl. Math., 188 (1997), pp. 215–243.

[26] G. Lebeau and E. Zuazua, *Decay rates for the three-dimensional linear system of thermoelasticity*, Arch. Rational Mech. Anal., 148 (1999), pp. 179–231.

[27] J. L. Lions, *Contrôlabilité Exacte et Stabilisation de Systèmes Distribués*, Vols. 1–2, Masson, Paris, 1988.

[28] K. Liu and Z. Liu, *Exponential stability and analyticity of abstract linear thermoelastic systems*, Z. Angew. Math. Phys., 48 (1997), pp. 885–904.

[29] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer-Verlag, New York, 1983.

[30] J. E. Muñoz Rivera and R. Racke, *Smoothing properties, decay, and global existence of solutions to nonlinear coupled systems of thermoelastic type*, SIAM J. Math. Anal., 26 (1995), pp. 1547–1563.

[31] B. Rao, *A compact perturbation method for the boundary stabilization of the Rayleigh beam equation*, Appl. Math. Optim., 33 (1996), pp. 253–264.

[32] D. L. Russell, *A general framework for the study of indirect damping mechanisms in elastic systems*, J. Math. Anal. Appl., 173 (1993), pp. 339–358.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 2, pp. 542–562

AVERAGING AND VIBRATIONAL CONTROL OF MECHANICAL SYSTEMS*

FRANCESCO BULLO†

Abstract. This paper investigates averaging theory and oscillatory control for a large class of mechanical systems. A link between averaging and controllability theory is presented by relating the key concepts of averaged potential and symmetric product. Both analysis and synthesis results are presented within a coordinate-free framework based on the theory of affine connections.

The analysis focuses on characterizing the behavior of mechanical systems forced by high amplitude high frequency inputs. The averaged system is shown to be an affine connection system subject to an appropriate forcing term. If the input codistribution is integrable, the subclass of systems with Hamiltonian equal to "kinetic plus potential energy" is closed under the operation of averaging. This result precisely characterizes when the notion of averaged potential arises and how it is related to the symmetric product of control vector fields. Finally, a notion of vibrational stabilization for mechanical systems is introduced, and sufficient conditions are provided in the form of linear matrix equality and inequality tests.

Key words. mechanical system, averaging, vibrational stabilization, nonlinear controllability

AMS subject classifications. 34C29, 70Q05, 93B05, 93B29, 93D99

PII. S0363012999364176

1. Introduction.

This paper investigates the open loop response of nonlinear mechanical control systems. This topic is studied in different ways by the classic disciplines of averaging and controllability. Relying on tools from both fields, this work characterizes the response of a large class of mechanical systems to high amplitude high frequency forcing. The class of mechanical control systems we consider includes systems with integrable inputs (Hamiltonian systems with conservative forces) as well as systems with more general types of forces and nonholonomic constraints.

Averaging and vibrational stabilization techniques find useful applications in various areas. Within the context of mechanical systems, much recent interest has focused on the control of underactuated robotic manipulators and on the analysis and design of robotic locomotion devices. Underactuated robotic manipulators have fewer control inputs than their degrees of freedom due to either design or failure. In both cases, the objective is to control the system despite the lack of control authority. Examples of works in this area are [37, 25], where the authors investigate the control via oscillatory inputs for some two and three degrees of freedom planar manipulators.

Robotic locomotion studies the movement patterns that biological systems and mechanical robots undergo during locomotion; see [24]. Typically, cyclic motion in certain internal variables generates displacement in Euclidean space; consider the example of how a snake changes its shape to locomote. Computing the feasible trajectories of a locomotion system is an analytically untractable problem for any nontrivial example. Averaging provides a means of tackling such problems; see, for example,

*Received by the editors November 8, 1999; accepted for publication (in revised form) January 25, 2002; published electronically June 26, 2002. This research was supported by the Campus Research Board at the University of Illinois at Urbana-Champaign and by NSF grant CMS-0100162.
http://www.siam.org/journals/sicon/41-2/36417.html
†Coordinated Science Laboratory and General Engineering Department, The University of Illinois at Urbana-Champaign, 1308 W. Main St., Urbana, IL 61801 (bullo@uiuc.edu, http://motion.csl.uiuc.edu).

the contributions on motion planning and trajectory generation documented in [17, 6] and the references therein.

Finally, averaging analysis seems well suited to tackle novel applications in the field of microelectromechanical systems, and vibrational control is being investigated within the context of active control of fluids and separation control. Examples include [7] on the scale dependence in oscillatory control and [43] on unsteady flow control using oscillatory blowing. In these settings, vibrational stabilization schemes appear advantageous since they require no expensive or complicated sensing.

Literature review. Averaging theory is discussed in a number of textbooks [13, 42, 19]. The control relevance of averaging ideas was underlined in the work on vibrational control by Bellman, Bentsman, and Meerkov [9, 10] and by Bentsman [11]. These works introduce vibrational stabilization techniques under various types of input forcing (e.g., vector additive, linear, and nonlinear multiplicative forcing). The later work by Baillieul [3, 4, 5] and Baillieul and Lehman [6] extends these techniques to the context of mechanical systems described by specific Lagrangian and Hamiltonian models. In particular, the work in [3] presents two treatments of averaging for mechanical control systems. The first approach relies on a coordinate transformation to bring the system to standard averaging form. The second approach is based on directly averaging the Hamiltonian function and gives rise to the notion of *averaged potential*. Some assumptions restrict how applicable the latter approach is. For example, the control system is assumed to have a cyclic variable and to be single-input with the control input applied to the cyclic variable. Nonetheless, the notion of averaged potential has proven very successful in treating a number of important cases; see, for example, [50, 51, 7].

Another set of relevant results includes the work on small-time local controllability for mechanical systems. The main references are the original work in Lewis and Murray [32] and the advances in [31]. These works introduce the notion of configuration and equilibrium controllability and provide sufficient conditions to characterize them. The main technical tool is the notion of *symmetric product* as a way to represent certain Lie brackets. Control algorithms that exploit motions along the "symmetric product directions" are presented in [17].

Statement of contributions. This paper contains a number of novel results both on averaging analysis as well as on control design. One key technical contribution is the understanding of the relationship between the symmetric product [32] and the averaged potential [3]. We describe the contributions in the next three paragraphs.

We start by studying the behavior of a large class of mechanical systems forced by high amplitude high frequency inputs. We rely on the notion of a *system described by an affine connection* as a generalized way of describing mechanical control systems with simple Hamiltonian, generic nonintegrable (nonconservative) forces and nonholonomic constraints. Under mild assumptions, we show how the averaged system is again a system described by an affine connection and subject to an appropriate forcing. Since this forcing term is a certain symmetric product, the result illustrates an instructive connection between controllability and averaging. The averaging analysis relies on a careful application of the variation of constants formula and of the homogeneity property of mechanical systems. The theorem statement and proof are presented in a coordinate-free manner.

We then consider the set of simple mechanical systems, that is, systems with Hamiltonian equal to "kinetic plus potential," and we investigate when this subclass is closed under the operation of averaging. A sufficient condition is that the input

codistribution be integrable, or in other words, that the control forces be described by conservative fields. Under this assumption, the Hamiltonian function of the average system includes a generalized averaged potential. This result shows how the notion of averaged potential is applicable to a wider set of systems than those considered by Baillieul [3]. The proof relies on the observation that the averaged potential is related to a certain symmetric product of functions; see [20].

Finally, we focus on the design of open and closed loop controllers based on high amplitude high frequency forcing. We introduce an appropriate notion of vibrational stabilization for mechanical systems, where only the configuration variables are considered. We consider simple systems with integrable forces and assume that the control system is underactuated (i.e., fewer control inputs are available than degrees of freedom). We consider the point stabilization problem and design control Lyapunov functions via the "potential shaping" technique; see the original [46] and a modern account in [47]. Here the closed loop potential energy reflects the presence of both a proportional action as well an oscillatory action. We provide sufficient conditions for stabilizability in the form of a linear matrix equality and inequality test. We illustrate the control design by applying it to an underactuated two-link manipulator.

Organization. The paper is organized as follows. We present a quick summary of averaging and introduce some tools from chronological calculus in section 2. In section 3, we introduce a useful classification of mechanical systems and study their common homogeneous structure. Section 4 contains the averaging analysis. In section 5, we present the vibrational stabilization results and work out the example.

2. Averaging and the variation of constants formula. In this section, we present some basic results on averaging theory and their coordinate-free interpretation. The averaging results are taken from Sanders and Verhulst [42] and from Guckenheimer and Holmes [22].

Let x, y, x_0 belong to an open subset $D \subset \mathbb{R}^n$, let $t \in \mathbb{R}_+ = [0, \infty)$, and let the parameter ϵ vary in the range $(0, \epsilon_0]$ with $\epsilon_0 \ll 1$. Let $f, g : \mathbb{R}_+ \times D \to \mathbb{R}^n$ be smooth time-varying vector fields. Consider the initial value problem in *standard form*:

$$\frac{dx}{dt} = \epsilon f(t, x), \qquad x(0) = x_0. \tag{2.1}$$

If $f(t, x)$ is a T-periodic function in its first argument, we let *the averaged system* be the initial value problem

$$\frac{dy}{dt} = \epsilon f^0(y), \qquad y(0) = x_0, \tag{2.2}$$

$$f^0(y) = \frac{1}{T} \int_0^T f(t, y) dt.$$

We say that an estimate is *on the time scale* $\delta^{-1}(\epsilon)$ if the estimate holds for all times t such that $0 < \delta(\epsilon)t < L$ with L a constant independent of ϵ. From pages 39 and 71 in [42] and from page 168 in [22], we summarize as follows.

THEOREM 2.1 (first order periodic averaging). *There exists a positive* ϵ_0 *such that, for all* $0 < \epsilon \leq \epsilon_0$,

(i) $x(t) - y(t) = O(\epsilon)$ *as* $\epsilon \to 0$ *on the time scale* $1/\epsilon$, *and*

(ii) *if the origin is a hyperbolically stable critical point for* f^0, *then* $x(t) - y(t) = O(\epsilon)$ *as* $\epsilon \to 0$ *for all* $t \in \mathbb{R}_+$, *and the differential equation* (2.1) *possesses a unique periodic orbit which is hyperbolically stable and belongs to an* $O(\epsilon)$ *neighborhood of the origin.*

Next, consider the initial value problem

$$\frac{dx}{dt} = f\left(t/\epsilon, x\right), \qquad x(0) = x_0, \tag{2.3}$$

where $f(t,x)$ is a T-periodic function in its first argument. A time scaling argument shows that the averaged version of this problem is the same as in (2.2). Accordingly, Theorem 2.1 implies that $x(t) - y(t) = O(\epsilon)$ as $\epsilon \to 0$ *only* on the time scale 1 unless $y = 0$ is a hyperbolically stable point of f^0.

2.1. Variation of constants formula in coordinate-free terms. The variation of constants formula is a means to bring various systems into the standard form in (2.1). This tool originates in Lagrange's work (see [42, page 183]) and is presented here in a coordinate-free setting.

Given a diffeomorphism ϕ and a vector field g, the *pull-back of g along ϕ*, denoted $\phi^* g$, is the vector field

$$(\phi^* g)(x) \triangleq \left(\frac{\partial \phi^{-1}}{\partial x} \circ g \circ \phi\right)(x),$$

where the order of composition of functions is $(\varphi \circ \phi)(x) = \varphi(\phi(x))$. A useful diffeomorphism is the flow map $y(t) = \Phi^g_{0,T}(y_0)$ describing the solution at time T to the initial value problem

$$\dot{y} = g(t,y), \quad y(0) = y_0.$$

Next, consider the initial value problem

$$\dot{x}(t) = f(x,t) + g(x,t), \qquad x(0) = x_0. \tag{2.4}$$

We regard f as a perturbation to the vector field g, and we seek to characterize the flow map of $f+g$ in terms of the nominal flow map of g. The answer is provided by the *variation of constants* formula:

$$\Phi^{f+g}_{0,t} = \Phi^g_{0,t} \circ \Phi^{(\Phi^g_{0,t})^* f}_{0,t}. \tag{2.5}$$

In other words, if $z(t)$ is the solution to the initial value problem

$$\dot{z}(t) = \left((\Phi^g_{0,t})^* f\right)(z), \qquad z(0) = x_0, \tag{2.6}$$

the solution $x(t)$ to the initial value problem (2.4) satisfies

$$\dot{x}(t) = g(t,x), \qquad x(0) = z(t). \tag{2.7}$$

We illustrate the formula in Figure 2.1 and provide a self-contained proof in the appendix.

2.2. Formal expansions for the pull-back of a flow map. Here we study in more detail the differential geometry of the initial value problem (2.6). Such a system is referred to as the "pulled back" or the "adjoint" system; e.g., see [23].

If f and g are time-invariant vector fields, the infinitesimal Campbell–Baker–Hausdorff formula (see [26]) provides a means of computing the pull-back

$$(\Phi^g_{0,t})^* f(x) = \sum_{k=0}^{\infty} \mathrm{ad}^k_g f \, \frac{t^k}{k!},$$

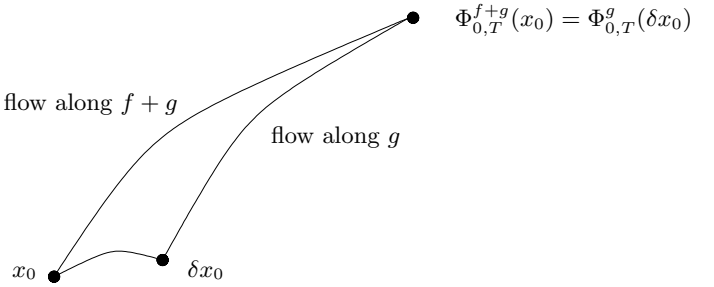


FIG. 2.1. *The flow along $f+g$ with initial condition x_0 equals the flow along g with initial condition δx_0. The variation δx_0 is computed via the variation of constants formula as the flow along $(\Phi^g_{0,t})^* f$ for time $[0,T]$ with initial condition x_0.*

where $\mathrm{ad}_g f(x) = [g,f](x)$ is the Lie bracket between g and f and $\mathrm{ad}^k_g f = \mathrm{ad}^{k-1}_g \mathrm{ad}_g f$.

If, instead, f is the time-invariant vector field and g is the time-varying vector field, we invoke a result from the chronological calculus formalism by Agrachev and Gamkrelidze [2]. It turns out that

$$((\Phi^g_{0,t})^* f)(t,x) = f(x) + \sum_{k=1}^{\infty} \int_0^t \dots \int_0^{s_{k-1}} \left(\mathrm{ad}_{g(s_k,x)} \dots \mathrm{ad}_{g(s_1,x)} f(x)\right) ds_k \dots ds_1. \tag{2.8}$$

The convergence properties for the series expansion in (2.8) are difficult to characterize; see, for example, a related discussion in [49] on the Campbell–Baker–Hausdorff formula. Nonetheless, sufficient conditions for local convergence are given in [2, Propositions 2.1 and 3.1]. For our analysis, the following simple statement suffices: if the terms $\mathrm{ad}_{g(s_k)} \dots \mathrm{ad}_{g(s_1)} f$ vanish for all k greater than a given N, then the series in (2.8) becomes a finite sum.

2.3. Averaging under high magnitude high frequency forcing. We return to the description of averaging results, and we focus on a setting of interest in vibrational stabilization problems [9, 10, 11]. Consider the initial value problem

$$\frac{dx}{dt} = f(x) + (1/\epsilon) g\left(t/\epsilon\,, x\right), \qquad x(0) = x_0, \tag{2.9}$$

where we assume that $g(t,x)$ is a T-periodic function in its first argument. Let $\Phi^g_{0,t}$ denote the flow map along $g(t,x)$, and define

$$F(t,x) = \left((\Phi^g_{0,t})^* f\right)(x), \tag{2.10}$$

$$F^0(x) = \frac{1}{T} \int_0^T F(\tau, x) d\tau. \tag{2.11}$$

Finally, let z and y be solutions to the initial value problems

$$\dot z = F(t/\epsilon, z), \quad z(0) = x_0, \tag{2.12}$$

$$\dot y = F^0(y), \quad y(0) = x_0. \tag{2.13}$$

LEMMA 2.2. *Let F be a T-periodic function in its first argument. For $t \in \mathbb{R}_+$, we have*

$$x(t) = \Phi^g_{0,t/\epsilon}(z(t)).$$

As $\epsilon \to 0$ on the time scale 1, we have

$$z(t) - y(t) = O(\epsilon).$$

If the origin is a hyperbolically stable critical point for F^0, then $z(t) - y(t) = O(\epsilon)$ as $\epsilon \to 0$ for all $t \in \mathbb{R}_+$, and the differential equation (2.12) possesses a unique periodic orbit which is hyperbolically stable and belongs to an $O(\epsilon)$ neighborhood of the origin.

Proof. As a first step, we change the time scale by setting $\tau = t/\epsilon$. Equation (2.9) becomes

$$\frac{d}{d\tau}x = \epsilon f(x) + g(\tau, x), \quad x(0) = x_0.$$

As a second step, we apply the variation of constants formula

$$\begin{aligned}\frac{d}{d\tau}x &= g(\tau, x), \quad x(0) = z(\tau),\\ \frac{d}{d\tau}z &= \epsilon F(\tau, z), \quad z(0) = x_0,\end{aligned}$$

where F is defined according to (2.10). As a third step, we average the initial value problem in z to obtain

$$\frac{d}{d\tau}y = \epsilon F^0(y), \quad y(0) = x_0,$$

where F^0 is defined according to (2.11) and F is assumed to be a T-periodic function. The averaged curve y approximates z over the time scale $\tau = 1/\epsilon$ and over all time according to Theorem 2.1. As a fourth step, we change the time scale back to $t = \epsilon\tau$ and compute

$$\begin{aligned}\frac{d}{dt}x &= (1/\epsilon)g(t/\epsilon, x), \quad x(0) = z(t),\\ \frac{d}{dt}z &= F(t/\epsilon, z), \quad z(0) = x_0,\\ \frac{d}{dt}y &= F^0(y), \quad y(0) = x_0.\end{aligned}$$

These are the definitions of z and y in (2.12) and (2.13). Finally, the equality in $x(t)$ follows by noting that the flow along $(1/\epsilon)g\,(t/\epsilon\,, x)$ for time 1 is equivalent to the flow along $g\,(t, x)$ for time $1/\epsilon$. □

This concludes our geometric presentation of averaging in systems with high magnitude high frequency inputs. These results on averaging and the variation of constants formula are known (see [10, Section III]), and they play a key role in the study of vibrational stabilization problems; see also [9, 11]. The presentation of these results in a coordinate-free fashion is novel: for a large class of mechanical control systems, an explicit expression will be provided for the infinite series describing the variation of constants formula.

3. Mechanical control systems and their homogeneous structure. In this section, we present three different types of mechanical systems and a geometric formalism that leads to a unified modeling framework. We also present some results on the Lie algebraic structure common to these systems and to generic second order control systems, where the input is an acceleration (alternatively, a force). To present

an accessible treatment, we assume the configuration space to be $Q = \mathbb{R}^n$. However, Remark 3.1 and section 3.1 provide the key ideas necessary to develop a coordinate-free treatment over manifolds.

Let $q = (q^1, \dots, q^n) \in \mathbb{R}^n$ be the configuration of the mechanical system. We consider the control system

$$\ddot{q}^i + \Gamma^i_{jk}(q)\dot{q}^j\dot{q}^k = Y^i_0(q) + Y^i_a(q)u^a(t) + R^i_j(q)\dot{q}^j, \tag{3.1}$$

where the summation convention is in place here and in what follows, the indices j, k run from 1 to n, the index a runs from 1 to m (the number of input fields), and where the following hold.

(i) The Γ^i_{jk} are $n^2(n+1)/2$ arbitrary scalar functions on $\mathbb{R}^n$ called the Christoffel symbols. (They satisfy the symmetric relationship $\Gamma^i_{jk} = \Gamma^i_{kj}$.)

(ii) $q \mapsto Y_a(q)$ for $a = 1, \dots, m$ are vector fields characterizing configuration-dependent forces applied to the system. Y_0, for example, might include the effect of a conservative force such as gravity.

(iii) The functions $t \mapsto u^a(t)$ are integrable and describe the control magnitude applied along the input Y_a. The ith component of Y_a is Y^i_a. We also let

$$Y(q,t) = Y_a(q)u^a(t).$$

(iv) $R(q)\dot{q}$ describes a generic force linearly proportional to velocity.

All quantities are assumed to be smooth functions of their arguments.

Equation (3.1) describes a large class of mechanical systems with Hamiltonian equal to kinetic plus potential energy, with symmetries and with nonholonomic constraints. A slightly loose but instructive classification follows.

Simple systems with integrable forces. These systems have Hamiltonian equal to "kinetic plus potential energy" and are subject to integrable (conservative) input forces. For example, should the mechanical system be a robotic manipulator with motors at joints, then the appropriate Christoffel symbols are computed via a well-known combination of partial derivatives of the inertia tensor; see the definition of the Coriolis matrix in [36], for example. Only for this kind of system can one write a Hamiltonian function that includes the effect of forces; the treatment in Chapter 14 of [38] relies on this assumption.

Simple systems with nonintegrable forces. This class is a superset of the previous class, where, however, *nonintegrable* input forces are allowed. For example, the force applied by a thruster of a satellite, hovercraft, or underwater vehicle is in general a nonintegrable force. Simplified equations of motion can be written if the system has symmetries, i.e., if the system's configuration belongs to the group of rigid displacements (or one of its subgroups) and its Hamiltonian is independent of the configuration.

Systems with nonholonomic constraints. This set includes systems from the previous two subclasses and is additionally subject to nonholonomic constraints. Two very interesting locomotion devices called snakeboard and roller racer are described in recent papers [40] and [29]. Two methodologies for writing the equations of motions for these systems into form (3.1) are discussed in [30, 31, 12]. While the description "nonholonomic" is commonly used to refer to wheeled robots and while such systems are usually driftless,[1] we consider here nonholonomic systems with drift.

[1] Driftless control systems have the characterizing property that $u_i = 0$ implies $\dot{x} = 0$, where x is the state and u_i are the inputs.

Three remarks are appropriate. First, the model relies on no specific structure on the Γ^i_{jk} functions. In the classic Hamiltonian system case, these functions are readily computed from the inertia matrix. By leaving these functions unspecified, our analysis includes systems with nonholonomic constraints. We refer to [12, 31] for a thorough treatment of this point.

Second, the distinctions between these three sets of mechanical systems have various instructive implications. For example, the notion of "actuated degree of freedom" is well defined only in systems subject to integrable forces. This simple fact is neglected even in recent literature on mechanical control systems.

Third, more complete definitions of the various quantities above should include transformation rules under changes of coordinates. For example, the Christoffel symbols $\{\Gamma^i_{jk}, i, j, k = 1, \dots, n\}$ obey relatively surprising transformation rules if the correct equations of motion are to be computed. If $\overline{q} = (\overline{q}^1, \dots, \overline{q}^n) \in \mathbb{R}^n$ are the transformed coordinates, the transformation rule for the Γ^i_{jk} is

$$\overline{\Gamma}^k_{ij} = \frac{\partial q^p}{\partial \overline{q}^i} \frac{\partial q^m}{\partial \overline{q}^j} \frac{\partial \overline{q}^k}{\partial q^r} \Gamma^r_{pm} + \frac{\partial \overline{q}^k}{\partial q^l} \frac{\partial^2 q^l}{\partial \overline{q}^i \partial \overline{q}^j}. \tag{3.2}$$

We refer to [35, section 7.5] for a more complete discussion.

3.1. Control systems described by an affine connection. Equations (3.1) are the Euler–Lagrange equations for a simple mechanical system. Numerous methodologies are available for writing these equations in vector or in abstract formats. The theory of affine connections is a convenient formalism that formalizes the Euler–Lagrange equations as well as more general second order control systems (including systems with nonholonomic constraints).

An easily accessible treatment of the theory of affine connections is given by Do Carmo [21]. An early reference on mechanical control systems on Riemannian manifolds is the work by Crouch [20]. The use of Riemannian concepts is encountering increasing success as testified by the contributions on modeling [12], decompositions [33], controllability [32], stabilization [28], tracking [18], interpolation [39], and (static and dynamic) feedback linearization [8, 41].

A smooth affine connection ∇ is a collection of n^3 smooth functions Γ^i_{jk} that satisfy the transformation rule in (3.2). An affine connection induces an operation between vector fields as follows. Let the vector fields X and Y have components

$$X(q) = X^i(q) \frac{\partial}{\partial q^i} \qquad \text{and} \qquad Y(q) = Y^i(q) \frac{\partial}{\partial q^i}.$$

The covariant derivative of Y along X is the vector field $\nabla_X Y$ defined by

$$\nabla_X Y = \left(\frac{\partial Y^i}{\partial q^j} X^j + \Gamma^i_{jk} X^j Y^k \right) \frac{\partial}{\partial q^i}.$$

Similarly, an affine connection induces an operation between a curve $\gamma : [0, 1] \mapsto \mathbb{R}^n$ and a vector field Y. The covariant derivative of Y along γ is a vector field along γ defined by

$$\nabla_{\dot\gamma} Y = \left(\frac{dY^i(\gamma(t))}{dt} + \Gamma^i_{jk} \dot\gamma^j Y^k \right) \frac{\partial}{\partial q^i}.$$

Whenever the reference curve is uniquely determined, we let $\nabla_{\dot\gamma} Y = \frac{DY}{dt}$. The two definitions of covariant derivative have similarities; however, $\frac{DY}{dt}$ is not a vector field

over $\mathbb{R}^n$, but it is only defined on the trajectory $\gamma : [0,1] \mapsto \mathbb{R}^n$. We refer to [21] for a more complete treatment of affine connections and of manifolds.

We are finally ready to rewrite (3.1) in a coordinate-free fashion. According to the definition of covariant derivative along a curve, the generalized Euler–Lagrange equations are

$$\frac{D\,\dot{q}}{dt} = Y_0(q) + R(q)\dot{q} + Y_a(q)u^a(t), \tag{3.3}$$

where the covariant derivative of $\dot{q}$ is computed along the curve $q(t)$, i.e., $D\dot{q}/dt = \nabla_{\dot{q}}\dot{q}$.

3.2. Lie algebraic structure. The fundamental structure of the control system in (3.1) (and, accordingly, (3.3)) is the polynomial dependence of the various vector fields on the velocity variable $\dot{q}$. This structure affects the Lie bracket computations involving input and drift vector fields; see related ideas in [32, 45]. We start by rewriting the system (3.1) as a first order differential equation. We write

$$\frac{d}{dt}\begin{bmatrix} q \\ \dot{q} \end{bmatrix} = \begin{bmatrix} \dot{q} \\ -\Gamma(q,\dot{q}) + Y_0(q) + R(q)\dot{q} \end{bmatrix} + \begin{bmatrix} 0 \\ Y_a \end{bmatrix} u^a(t),$$

where $\Gamma(q,\dot{q})$ is the vector with ith component $\Gamma^i_{jk}(q)\dot{q}^j\dot{q}^k$. Also, we let $x = (q,\dot{q})$,

$$Z_g(x) = \begin{bmatrix} \dot{q} \\ -\Gamma(q,\dot{q}) \end{bmatrix}, \quad Y_a^{\text{lift}}(x) \triangleq \begin{bmatrix} 0 \\ Y_a(q) \end{bmatrix}, \quad \text{and} \quad R^{\text{lift}}(x) \triangleq \begin{bmatrix} 0 \\ R(q)\dot{q} \end{bmatrix}$$

so that the control system is rewritten as

$$\dot{x} = Z_g(x) + Y_0^{\text{lift}}(x) + R^{\text{lift}}(x) + Y_a^{\text{lift}}(x)u^a(t).$$

Let $h_i(q,\dot{q})$ be the set of scalar functions on $\mathbb{R}^{2n}$, which are arbitrary functions of q and homogeneous polynomials in $\{\dot{q}^1,\dots,\dot{q}^n\}$ of degree i. Let $\mathcal{P}_i$ be the set of vector fields on $\mathbb{R}^{2n}$ whose first n components belong to h_i and whose second n components belong to h_{i+1}. It is easily seen that

$$Z_g \in \mathcal{P}_1, \quad R^{\text{lift}} \in \mathcal{P}_0, \quad \text{and} \quad Y_a^{\text{lift}} \in \mathcal{P}_{-1}.$$

Direct computations show that the sets $\{\mathcal{P}_i\}$ have the following properties:

(i) $[\mathcal{P}_i, \mathcal{P}_j] \subset \mathcal{P}_{i+j}$, i.e., the Lie bracket between a vector field in $\mathcal{P}_i$ and a vector field in $\mathcal{P}_j$ belongs to $\mathcal{P}_{i+j}$.
(ii) $\mathcal{P}_k = \{0\}$ for all $k \le -2$.
(iii) if $k \ge 1$, then $X(q,0) = 0$ for all $X(q,\dot{q}) \in \mathcal{P}_k$.

Given these properties, we investigate the Lie brackets between the vector fields Z_g and Y_a^{lift}. A few useful brackets are

$$[Z_g, Y_a^{\text{lift}}] \in \mathcal{P}_0, \qquad [Y_a^{\text{lift}}, Y_b^{\text{lift}}] = 0,$$
$$[Y_b^{\text{lift}}, [Z_g, Y_a^{\text{lift}}]] \in \mathcal{P}_{-1}.$$

Of particular interest is the Lie bracket $[Y_b^{\text{lift}}, [Z_g, Y_a^{\text{lift}}]]$. Since this vector field belongs to $\mathcal{P}_{-1}$, there must exist a vector field on $\mathbb{R}^n$, which we denote $\langle Y_a : Y_b \rangle$, such that

$$\langle Y_a : Y_b \rangle^{\text{lift}} = [Y_b^{\text{lift}}, [Z_g, Y_a^{\text{lift}}]].$$

We call this vector field the *symmetric product* between Y_b and Y_a. Some straightforward computations in coordinates show that $\langle Y_a : Y_b\rangle = \langle Y_b : Y_a\rangle$ and that

$$\langle Y_b : Y_a\rangle^i = \frac{\partial Y_a^i}{\partial q^j} Y_b^j + \frac{\partial Y_b^i}{\partial q^j} Y_a^j + \Gamma^i_{jk}(Y_a^j Y_b^k + Y_a^k Y_b^j),$$
$$\langle Y_b : Y_a\rangle = \nabla_{Y_a} Y_b + \nabla_{Y_b} Y_a.$$

REMARK 3.1. *While the results in this section are presented in coordinates, it is possible to turn them into coordinate-free statements on manifolds. The enabling concepts are the operation of vertical lift and symmetric product between vector fields (see* [32]*), the notion of geometric homogeneity (see* [27]*), and the intrinsic definition of the Liouville vector field (see* [34, page 64]*).*

4. Averaging for mechanical systems under high amplitude high frequency forcing. This section contains the main result of the paper. We consider systems described by an affine connection and subject to high amplitude high frequency forcing. We show how the average system is again described by the same affine connection subject to an appropriate forcing term. Additionally, we show how the subclass of systems subject to integrable forces and without nonholonomic constraints is also closed under the operation of averaging.

The approach we take differs substantially from the classic averaging of Hamiltonian systems; see Chapter 4 in [22]. In that setting, the Hamiltonian system is integrable, and the variation of constants formula is applied by treating the ϵ size forcing as perturbation. In our setting, it is the Hamiltonian dynamics that plays the role of the perturbation to the dominant high amplitude high frequency forcing. Finally, it is important to note that, while the accelerations driving the systems are high amplitude, the generated displacements are typically small in magnitude.

4.1. Systems described by affine connections. Consider a control system described by an affine connection as in (3.3):

$$\begin{aligned} &\frac{D\,\dot q}{dt} = Y_0(q) + R(q)\dot q \;+\; Y_a(q)(1/\epsilon)v^a(t/\epsilon),\\ &q(0) = q_0, \quad \dot q(0) = v_0, \end{aligned} \tag{4.1}$$

where $u^a(t) = v^a(t/\epsilon)/\epsilon$, and $\{v^1, \dots, v^m\}$ are T-periodic functions that satisfy

$$\int_0^T v^a(s_1)ds_1 = 0, \tag{4.2}$$

$$\int_0^T \int_0^{s_2} v^a(s_1)ds_1 ds_2 = 0. \tag{4.3}$$

Also, let $v(t) = [v^1(t), \dots, v^m(t)]'$ and define the matrix Λ according to

$$\Lambda = \frac{1}{2T}\int_0^T \left(\int_0^{s_1} v(s_2)ds_2\right)\left(\int_0^{s_1} v(s_2)ds_2\right)' ds_1. \tag{4.4}$$

Finally, define the time-varying vector field as

$$\Xi(t,q) = \left(\int_0^t v^a(s)ds\right) Y_a(q)$$

and the curve as

$$z(t) = \big(q(t),\ \dot{q}(t) - \Xi\big(t/\epsilon, q(t)\big)\big)\,. \tag{4.5}$$

THEOREM 4.1. *Let $q(t)$ be the solution to the initial value problem in* (4.1)*, and let $r(t)$ be the solution to*

$$\begin{aligned} \frac{D\,\dot{r}}{dt} &= Y_0(r) + R(r)\dot{r} - \sum_{a,b=1}^{m} \Lambda_{ab}\,\langle Y_a : Y_b\rangle\,(r), \\ r(0) &= q_0, \quad \dot{r}(0) = v_0. \end{aligned} \tag{4.6}$$

There exists a positive ϵ_0 such that, for all $0 < \epsilon \leq \epsilon_0$,

$$\begin{aligned} q(t) &= r(t) + O(\epsilon), \\ \dot{q}(t) &= \dot{r}(t) + \Xi(t/\epsilon, q(t)) + O(\epsilon) \end{aligned} \tag{4.7}$$

as $\epsilon \to 0$ on the time scale 1*.*

Furthermore, let $(r, \dot{r}) = (q_1, 0)$ be a hyperbolically stable critical point for (4.6)*, and let its region of attraction contain the initial condition (q_0, v_0). Then the approximations in* (4.7) *are valid for all $t \in \mathbb{R}_+$, and the curve $z(t)$ is the solution to an initial value problem which possesses a unique, hyperbolically stable, periodic orbit belonging to an $O(\epsilon)$ neighborhood of $(q_1, 0)$.*

Justified by the approximations in (4.7), we call the initial value problem in (4.6) the *averaged mechanical system* of the initial value problem in (4.1).

Proof. The proof brings together the analysis in subsections 2.3 and 3.2. As a first step, we translate the second order (4.1) into the first order format in (2.9). We let $x = (q, \dot{q})$ and

$$\begin{aligned} f(x) &= Z_g(x) + Y_0^{\text{lift}}(x) + R^{\text{lift}}(x), \\ g(t,x) &= \sum_{a=1}^{m} Y_a^{\text{lift}}(x) v^a(t). \end{aligned}$$

Next, we compute the vector field F according to (2.10):

$$F(t,y) = \big((\Phi_{0,t}^g)^* f\big)\,(y) = \Big(\Phi_{0,t}^{\sum Y_a^{\text{lift}}(y) v^a(t)}\Big)^* \,(Z_g(y) + Y_0^{\text{lift}}(y) + R^{\text{lift}}(y)).$$

We study its expression according to the series expansion in section 2.2:

$$(\Phi_{0,t}^g)^* f = f + \sum_{k=1}^{\infty} \int_0^t \dots \int_0^{s_{k-1}} \left(\mathrm{ad}_{g(s_k)} \dots \mathrm{ad}_{g(s_1)}\, f\right)\, ds_k \dots ds_1.$$

The Lie algebraic structure unveiled in section 3.2 leads to remarkable simplifications:

$$\begin{aligned} \mathrm{ad}^k_{Y_a^{\text{lift}}}(Z_g(y) + Y_0^{\text{lift}}(y) + R^{\text{lift}}(y)) &= 0 \qquad \forall\ k \geq 3, \\ \mathrm{ad}_{Y_b^{\text{lift}}}\, \mathrm{ad}_{Y_a^{\text{lift}}}(Z_g(y) + Y_0^{\text{lift}}(y) + R^{\text{lift}}(y)) &= -\,\langle Y_a : Y_b\rangle^{\text{lift}}\,. \end{aligned}$$

With a little bookkeeping, we exploit these equalities and compute

$$\begin{aligned}
&\left(\Phi_{0,t}^{\sum Y_a^{\text{lift}}(y)v^a(t)}\right)^* \left(Z_g(y)+Y_0^{\text{lift}}(y)+R^{\text{lift}}(y)\right)\\
&\qquad=\left(Z_g+Y_0^{\text{lift}}+R^{\text{lift}}\right)+\sum_{a=1}^{m}\left(\int_0^t v^a(s_1)ds_1\right)\left[Y_a^{\text{lift}},\left(Z_g+Y_0^{\text{lift}}+R^{\text{lift}}\right)\right]\\
&\qquad\quad+\sum_{a,b=1}^{m}\left(\int_0^t\int_0^{s_b} v^b(s_b)v^a(s_a)ds_ads_b\right)\left[Y_b^{\text{lift}},\left[Y_a^{\text{lift}},\left(Z_g+Y_0^{\text{lift}}+R^{\text{lift}}\right)\right]\right]\\
&\qquad=\left(Z_g+Y_0^{\text{lift}}+R^{\text{lift}}\right)+\sum_{a=1}^{m}\left(\int_0^t v^a(s_1)ds_1\right)\left[Y_a^{\text{lift}},\left(Z_g+R^{\text{lift}}\right)\right]\\
&\qquad\qquad-\sum_{a,b=1}^{m}\left(\int_0^t\int_0^{s_b} v^b(s_b)v^a(s_a)ds_ads_b\right)\langle Y_a:Y_b\rangle^{\text{lift}}.
\end{aligned}$$

An integration by parts and the symmetry of the symmetric product lead to

$$\begin{aligned}
&\sum_{a,b=1}^{m}\left(\int_0^t\int_0^{s_b} v^b(s_b)v^a(s_a)ds_ads_b\right)\langle Y_a:Y_b\rangle\\
&\qquad=\frac{1}{2}\sum_{a,b=1}^{m}\left(\int_0^t v^b(s_b)ds_b\int_0^t v^a(s_a)ds_a\right)\langle Y_a:Y_b\rangle
\end{aligned}$$

so that we have

(4.8)

$$\begin{aligned}
F(t,y)=&\left(Z_g+Y_0^{\text{lift}}+R^{\text{lift}}\right)+\sum_{a=1}^{m}\left(\int_0^t v^a(s_1)ds_1\right)\left[Y_a^{\text{lift}},\left(Z_g+R^{\text{lift}}\right)\right]\\
&-\frac{1}{2}\sum_{a,b=1}^{m}\left(\int_0^t v^b(s_b)ds_b\int_0^t v^a(s_a)ds_a\right)\langle Y_a:Y_b\rangle^{\text{lift}}.
\end{aligned}$$

Assumption (4.2) implies that the function F is T-periodic so that we can compute its average F^0 according to (2.11). Given the assumption on v^a in (4.3) and the definition of Λ in (4.4), we have

$$F^0(y)=\left(Z_g+Y_0^{\text{lift}}+R^{\text{lift}}\right)-\sum_{a,b=1}^{m}\Lambda_{ab}\langle Y_a:Y_b\rangle^{\text{lift}}.$$

This is precisely the vector field that describes the evolution of $(r,\dot r)$. This proves that $y=(r,\dot r)$. Let $\widehat z=(p,\dot p)$ be the flow of the vector field F starting from (q_0,v_0). Lemma 2.2 implies that, over the appropriate time scale,

$$\begin{aligned}
x(t)&=\Phi^g_{0,t/\epsilon}(\widehat z(t)),\\
\widehat z(t)&=y(t)+O(\epsilon),
\end{aligned}$$

and, should $(q_1,0)$ be a hyperbolically stable critical point for F^0, the vector field F possesses a unique, hyperbolically stable, periodic orbit in an $O(\epsilon)$ neighborhood of $(q_1,0)$.

Finally, we verify that the curve $\widehat{z}$ defined via the equality $x(t) = \Phi^g_{0,t/\epsilon}(\widehat{z}(t))$ is equal to the curve z defined in (4.5). In coordinates, we have

$$\frac{d}{ds}\begin{bmatrix} q(s) \\ \dot{q}(s) \end{bmatrix} = \begin{bmatrix} 0 \\ Y_a(q(s))v^a(s) \end{bmatrix}, \qquad \big(q(0), \dot{q}(0)\big) = \Phi^g_{0,t/\epsilon}\big(p(t), \dot{p}(t)\big)$$

so that, at final time $s = t/\epsilon$, we compute $q(t) = q(0) = p(t)$ and

$$\dot{q}(t) = Y_a(q(0)) \int_0^{t/\epsilon} v^a(s)ds + \dot{q}(0) = \Xi(t/\epsilon, q(t)) + \dot{p}(t). \qquad \square$$

The coordinate-free treatment and the use of the Lie algebraic structure underline the connection between these results on averaging and the treatment on controllability in [32] and on motion planning in [17]. To quickly recall the first of these references, consider the control system in (3.3), where $Y_0 = R = 0$. If the family of vector fields $\{Y_a, \langle Y_a : Y_b\rangle, a, b = 1, \dots, m\}$ is full rank in a neighborhood of q_0, then the control system (3.3) is small-time locally accessible from $(q_0, 0)$. Similar in these works is the key observation that a mechanical control system subject to a force Y moves approximately in the direction spanned by $\langle Y : Y\rangle$.

The novel proof methodology should facilitate further research into higher order averaging. Indeed, the work in [16] indicates that the exact solution of a mechanical control system can be written as a series expansion with terms including iterated symmetric products and time integrals.

4.2. Averaged potential for simple systems with integrable inputs. The textbook [22] presents the classic result that "the average of a Hamiltonian system forced by a bounded high frequency perturbation can be computed by averaging its Hamiltonian." For the case of high magnitude high frequency forces, the various insightful works by Baillieul [3, 4] and Baillieul and Lehman [6] introduce the notion of *averaged potential*[2] as a means to characterize the average behavior.

In this section, we assume that the original forced system is "simple," i.e., that no nonholonomic constraints are present, and we answer the questions "when is the averaged system again simple?" and "what assumptions lead to the definition of an averaged potential?" Incidentally, the answer to these questions involves the relationships between various definitions of symmetric product that go back to the early treatment by Crouch [20].

We quickly review some basic concepts in simple mechanical control systems and refer to the textbooks [21, 35] for a more detailed presentation. In a mechanical system without constraints, the total energy is defined as the sum of potential $V(q)$ and kinetic $\frac{1}{2}\langle\langle \dot{q}\,,\, \dot{q}\rangle\rangle = \frac{1}{2}\dot{q}^T M(q)\dot{q}$, where we denote with both $\langle\langle \cdot\,,\, \cdot\rangle\rangle$ and M the metric associated with the kinetic energy. The tensor R is weakly dissipative if $\langle\langle \dot{q}\,,\, R\dot{q}\rangle\rangle \leq 0$; it is strictly quadratically dissipative if there exists a positive constant β such that

$$\langle\langle \dot{q}\,,\, R\dot{q}\rangle\rangle \leq -\beta\langle\langle \dot{q}\,,\, \dot{q}\rangle\rangle. \tag{4.9}$$

If integrable forces are present, they are written as $Y_a(q) = \operatorname{grad}\varphi_a(q)$ for $a = 1, \dots, m$, where a gradient vector field reads, in coordinates, as

$$(\operatorname{grad}\varphi_a)^i = M^{ij}\frac{\partial \varphi_a}{\partial q_j}.$$

[2]More precisely, in Baillieul's work, the inputs are assumed to be high frequency bounded magnitude velocities. It is therefore very similar to our setting with high magnitude high frequency accelerations.

According to the treatment in [38, Chapter 12], the controlled Hamiltonian is

$$H(q,p,u) = V(q) + \frac{1}{2}p' M(q)^{-1} p - \sum_{a=1}^{m} \varphi_a(q) u^a, \tag{4.10}$$

where the momentum $p = M(q)\dot{q}$. The affine connection is the Levi–Civita connection of the metric M. The Christoffel symbols are computed according to the usual

$$\Gamma_{ij}^k = \frac{1}{2} M^{mk} \left(\frac{\partial M_{mj}}{\partial q^i} + \frac{\partial M_{mi}}{\partial q^j} - \frac{\partial M_{ij}}{\partial q^m} \right).$$

The equations of motion (4.6) take the specific form

$$\frac{D\,\dot{q}}{dt} = -\operatorname{grad} V(q) + R(q)\dot{q} + \operatorname{grad} \varphi_a(q)\, u^a(t). \tag{4.11}$$

Next we present a useful result on the symmetric product of gradient vector fields.

LEMMA 4.2 (symmetric products of functions). *Let φ_1, φ_2 be two smooth scalar functions. The symmetric product $\langle \operatorname{grad}\varphi_1 : \operatorname{grad}\varphi_2 \rangle$ is again a gradient vector field. Additionally, if one defines a symmetric product of functions according to*

$$\langle \varphi_i : \varphi_j \rangle \triangleq \frac{\partial \varphi_i}{\partial q} M^{-1} \frac{\partial \varphi_j}{\partial q} = \langle\!\langle \operatorname{grad}\varphi_i \,,\, \operatorname{grad}\varphi_j \rangle\!\rangle, \tag{4.12}$$

then

$$\langle \operatorname{grad}\varphi_1 : \operatorname{grad}\varphi_2 \rangle = \operatorname{grad} \langle \varphi_1 : \varphi_2 \rangle .$$

This result was originally proven by Crouch in [20], where this symmetric product of functions was presented under the name of the Beltrami bracket. It is interesting to note how, in contrast to the treatment in [20], this symmetric operation is relevant here in a Hamiltonian system context.

Finally, we are ready to apply Theorem 4.1 to the setting of simple systems.

THEOREM 4.3. *Consider the simple mechanical control system in* (4.11) *with Hamiltonian in* (4.10). *Let $u^a(t) = v^a(t/\epsilon)/\epsilon$, and let the functions v^a satisfy the condition in* (4.3). *It follows that the averaged system is a simple mechanical system subject to no force and with Hamiltonian*

$$H_{\text{averaged}}(q,p) = V_{\text{averaged}}(q) + \frac{1}{2} p' M(q)^{-1} p,$$

where the averaged potential *is defined as*

$$V_{\text{averaged}}(q) \triangleq V(q) + \sum_{a,b=1}^{m} \Lambda_{ab} \langle \varphi_a : \varphi_b \rangle (q). \tag{4.13}$$

Accordingly, the equations of motion for the averaged system are

$$\frac{D\,\dot{q}}{dt} = -\operatorname{grad}(V_{\text{averaged}}) + R(q)\dot{q}.$$

The result follows directly from Lemma 4.2 and Theorem 4.1. Theorem 4.3 can be used as follows. In order to stabilize a mechanical control system, we design oscillatory inputs that render V_{averaged} positive definite about the desired equilibrium point. The next section presents this idea in detail.

5. Vibrational stabilization of mechanical systems. In this section, we apply the averaging results to stabilization problems. We focus on simple mechanical systems, consider the point stabilization problem via oscillatory inputs, and rely on the averaged Hamiltonian as a candidate control Lyapunov function; see [44].

We start by presenting the notion of vibrational stabilization according to the treatments in [9, 10, 11]. Consider the control system

$$\frac{dx}{dt} = f(x) + g_a(x)u^a(t). \tag{5.1}$$

A critical point x_1 of f is said to be *vibrationally stabilizable*[3] if, for any $\delta > 0$, there exist almost-periodic zero-average inputs $u^a(t)$ such that the system in (5.1) has an asymptotically stable almost periodic solution $x^*(t)$ characterized by

$$\|\overline{x}^* - x_1\| \leq \delta, \qquad \overline{x}^* = \lim_{T\to\infty} \frac{1}{T} \int_0^T x^*(s)ds.$$

REMARK 5.1. *We refer to* [9, 10, 11] *for the vibrational stabilization theory for systems controlled by vector additive, linear, and nonlinear multiplicative forcing. Adopting these definitions, the vibrational stabilization problem we consider corresponds to a nonlinear multiplicative setting; see* [11]. *In that paper, the ith component of the vibrational forcing depends only on the ith state variable. This requirement is removed here, and the structure of the nonlinearities we consider is more general.*

5.1. Stabilization in systems with integrable inputs. Once more, consider the control system in (4.11):

$$\frac{D\,\dot{q}}{dt} = -\operatorname{grad} V(q) + R(q)\dot{q} \;+\; \operatorname{grad} \varphi_a(q)u^a(t). \tag{5.2}$$

We present a notion of vibrational stabilization tailored to mechanical systems. A configuration q_1 is said to be *vibrationally stabilizable* if, for any $\delta > 0$, there exist almost-periodic zero-average inputs $u^a(t)$ such that the system in (5.2) has an asymptotically stable almost-periodic solution $q^*(t)$ characterized by

$$\|\overline{q}^* - q_1\| \leq \delta, \qquad \overline{q}^* = \lim_{T\to\infty} \frac{1}{T} \int_0^T q^*(s)ds. \tag{5.3}$$

This definition is weaker than the general one above since no requirement is imposed on the behavior of the velocity variables $\dot{q}$.

Next, we design vibrationally stabilizing control laws. The following useful lemma focuses on "inverting" the definition of $\Lambda = \Lambda(v^1, \ldots, v^m)$ in (4.4).

LEMMA 5.2 (design of vibrations). *Let $t \in [0, T]$, and define a vector-valued function of time $v(t) = [v^1(t), \ldots, v^m(t)]'$ that satisfies* (4.2) *and* (4.3). *Any matrix Λ computed according to* (4.4) *is symmetric and positive semidefinite. Vice versa, given any symmetric positive semidefinite matrix Λ, there exists a vector-valued function of time v that satisfies* (4.2), (4.3), *and* (4.4).

Proof. Obviously, Λ is symmetric, and, for any vector $x \in \mathbb{R}^m$, one has

$$x'\Lambda x = \int_0^T \left(\int_0^{s_1} \big(x'v(s_2)\big) ds_2 \right)^2 ds_1 \geq 0.$$

[3] Baillieul and Lehman [6] assume both the inputs and the asymptotically stable solution x^* to be T-periodic.

Given any symmetric positive semidefinite Λ, we design inputs that satisfy (4.2), (4.3), and (4.4). First, we introduce the T-periodic base functions

$$\psi_i(t) = \frac{4\pi i}{T} \cos\left(\frac{2\pi\, i}{T} t\right), \quad i \in \mathbb{N}.$$

Any linear combination of the $\{\psi_i\}$ satisfies (4.2), (4.3), and

$$\frac{1}{2T} \int_0^T \left(\int_0^{s_1} \psi_i(s_2) ds_2 \right) \left(\int_0^{s_1} \psi_j(s_2) ds_2 \right) ds_1 = \delta_{ij},$$

where δ_{ij} is the Kronecker delta. Next, we diagonalize Λ via an orthogonal similarity transformation W. Assuming the rank of Λ is $p \leq m$, we have

$$\Lambda = W \operatorname{diag}([\lambda_1, \ldots, \lambda_p, 0, \ldots, 0])\, W' = \sum_{i=1}^{p} (\sqrt{\lambda_i} W e_i)(\sqrt{\lambda_i} W e_i)',$$

where $\operatorname{diag}([\lambda_1, \ldots, \lambda_p, 0, \ldots, 0])$ is the diagonal matrix with nonvanishing elements $\{\lambda_1, \ldots, \lambda_p\}$, and where $\{e_i, \ldots, e_n\}$ is the usual basis for $\mathbb{R}^n$. Since the vectors $(\sqrt{\lambda_i} W e_i)$ are uniquely determined by Λ, we define

$$w(t, \Lambda) = \sum_{i=1}^{p} (\sqrt{\lambda_i} W e_i)\, \psi_i(t). \tag{5.4}$$

By construction, $v(t) = w(t, \Lambda)$ satisfies (4.2), (4.3), and (4.4). □

Introduce the control gains $k_1 \in \mathbb{R}^m$, $K_2, K_3 \in \mathbb{R}^{m \times m}$, subject to $K_2 = K_2' \geq 0$ and $K_3 = K_3' \geq 0$. To simplify notation, let $\varphi = [\varphi^1, \ldots, \varphi^m]$, and let the $m \times m$ matrix $\langle \varphi : \varphi \rangle (q)$ have (a, b) component $\langle \varphi_a : \varphi_b \rangle (q)$. Let the control input be the sum of open (feedforward) and closed loop (feedback) terms

$$u(t, \epsilon) = -k_1 - K_2 \varphi + (1/\epsilon) w(t/\epsilon, K_3), \tag{5.5}$$

where w is as defined in (5.4). According to Theorem 4.3 and to Lemma 5.2, the averaged controlled system is Hamiltonian with potential energy given by

$$V_{\text{control}}(q) = V(q) + k_1' \varphi(q) + \frac{1}{2} \varphi(q)' K_2 \varphi(q) + \operatorname{Trace}\big(K_3 \langle \varphi : \varphi \rangle\big), \tag{5.6}$$

where the Trace operation is equivalent to the summation in (4.13). It is useful to note that V_{control} depends linearly on the control gains k_1, K_2, K_3.

Existence and stability of equilibrium points are analyzed according to the classic potential energy criterion. The configuration q_1 is an equilibrium point if it is a critical point for the averaged controlled potential energy V_{control}; it is locally/globally stable if V_{control} has a local/global minimum at q_1. Of course, the point is stable only in the average approximation. We make this point precise in the following theorem.

THEOREM 5.3 (vibrational stabilization of configurations). *Consider the control system in* (5.2)*, and assume the tensor R is strictly quadratically dissipative. Let $q_1 \in \mathbb{R}^n$, and consider the following set of linear matrix equality and inequalities in the free variables k_1, K_2, K_3:*

$$\begin{aligned} K_2 = K_2' \geq 0, \qquad & K_3 = K_3' \geq 0, \\ \frac{\partial V_{\text{control}}}{\partial q}(q_1) = 0, \qquad & \frac{\partial^2 V_{\text{control}}}{\partial q^2}(q_1) > 0. \end{aligned} \tag{5.7}$$

If the convex problem (5.7) *is feasible, the configuration* q_1 *is vibrationally stabilizable, and there exists an* $\epsilon_0 > 0$ *such that stabilizing controls are computed according to* (5.5)*, with* $0 < \epsilon \leq \epsilon_0$ *and with* k_1, K_2, K_3 *solutions to the system of equations* (5.7)*.*

Proof. As a first step, we prove that $(q_1, 0)$ is a locally exponentially stable point for the averaged controlled system. We follow a well-known procedure (see [47]) and rely on Theorem 4.3 and Lemma 5.2. At $q = q_1$, the function V_{control} in (5.6) and its gradient vanish, while its Hessian is positive definite. The total energy $H_{\text{control}}(q, \dot{q}) \triangleq V_{\text{control}}(q) + \frac{1}{2}\dot{q}' M \dot{q}$ is therefore positive definite about $(q_1, 0)$. Because R is strictly quadratically dissipative, there exists a $\beta > 0$ such that, along the solutions of the averaged controlled system,

$$\dot{H}_{\text{control}} = -\beta \langle\!\langle \dot{q}\,,\, \dot{q} \rangle\!\rangle.$$

The function H_{control} is a Lyapunov function for the averaged controlled system, and $(q_1, 0)$ is a stable equilibrium point. Asymptotic stability follows from an application of LaSalle's lemma; exponential stability follows from a linearization argument.

As a second step, we prove that the controlled system has a unique periodic exponentially stable solution $q(t)$ in a neighborhood of q_1. We follow a well-known procedure (see [10]) and rely on Theorem 4.1. Since the averaged system has an exponentially stable point, the curve $z(t)$ is a solution to a differential equation which possesses a unique periodic orbit, say, $z^\star(t)$, which is exponentially stable and belongs to an $O(\epsilon)$ neighborhood of $(q_1, 0)$. The same statement can be made for the first component of $z(t)$, that is, the curve $q(t)$. We call this periodic orbit $q^\star(t)$ and its average $\overline{q}^\star$, as defined in (5.3). Since $q^\star(t)$ lives in a $O(\epsilon)$ neighborhood of q_1, so does $\overline{q}^\star$. Therefore, there must exist ϵ_0 such that $\|\overline{q}^* - q_1\| \leq \delta$ for any $\delta > 0$. □

The stability result relies on the open loop system having full rank dissipation; i.e., the tensor R is required to be strictly quadratically dissipative. This requirement can be weakened by augmenting the control input with a "derivative action" (a term negatively proportional to the velocity). Asymptotic stability is then guaranteed under a linear-controllability–like condition; see [47, 15].

The location of the poles of the linearized model about q_1 affects the behavior of the controlled system. Given that a large oscillatory signal is superimposed, better performance is achieved when these poles are far to the left of the imaginary axis. This and related performance requirements can be addressed within the linear matrix equality and inequality formulation; see the surveys in [48, 14].

5.2. Vibrational stabilization of an underactuated two-link manipulator. We present a simple example of vibrational stabilization. We consider a planar two-link manipulator as depicted in Figure 5.1: no potential energy is present. We assume the manipulator is subject to damping forces at both angles.

The configuration of the system is described by the pair (θ_1, θ_2), where θ_1 is the angle between the first link and the horizontal axis and θ_2 is the relative angle between the two links. Both angles are measured counterclockwise. The links' physical parameters are length ℓ, mass m, and moment of inertia I. We let $\ell_1 = 3, \ell_2 = 4$, $m_1 = I_1 = \ell_1^2$, and $m_2 = I_2 = \ell_2^2$. A known procedure provides the inertia matrix:

$$M(q) = \begin{bmatrix} \frac{1013}{4} + 192\cos(\theta_2) & 16(5 + 6\cos(\theta_2)) \\ 16(5 + 6\cos(\theta_2)) & 80 \end{bmatrix}.$$

We assume the system is subject to the damping force $(-.2\dot{\theta}_1, -.2\dot{\theta}_2)$ and to a single control input, i.e., a torque τ applied at the first joint. Accordingly, the force can

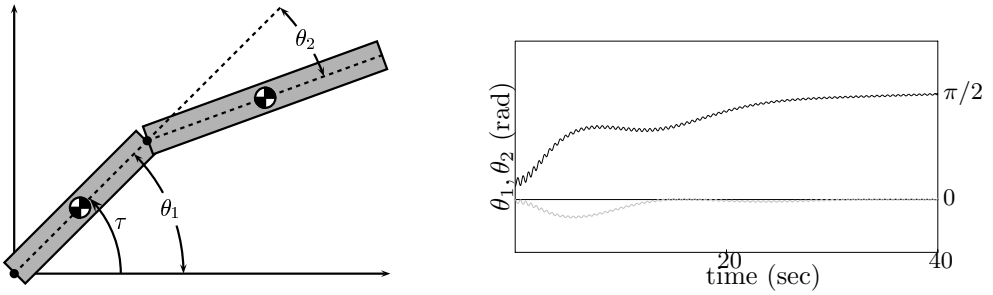


FIG. 5.1. *Two-link manipulator: θ_1 and θ_2 are measured counterclockwise. In the right figure, the gray line is θ_1, and the black line is θ_2. Despite the superimposed oscillatory behavior, the variables (θ_1, θ_2) converge to the global minimum of the averaged controlled potential energy.*

be described by the function $\varphi(q) = \theta_1$. The symmetric product is easily computed according to Lemma 4.2:

$$\langle \varphi : \varphi \rangle (q) = \frac{20}{2313 - 1152\cos(2\theta_2)}.$$

We adopt the control law in (5.5) and compute the averaged controlled potential according to (5.6):

$$V_{\text{control}}(q) = k_1\theta_1 + \frac{1}{2}k_2\theta_1^2 + k_3\frac{20}{2313 - 1152\cos(2\theta_2)}.$$

At $k_1 = 0$ and for any positive k_2 and k_3, the function V_{control} has two global minima at $(\theta_1, \theta_2) = (0, \pm\pi/2)$.

We run the simulation as follows. We design the control law parameters as $\epsilon = .5$, $T = 1$, $k_2 = 15$, and $k_3 = 150$. At initial time, the manipulator is at rest with angles $(\theta_1(0), \theta_2(0)) = (0, \pi/16)$. This initial condition is in the domain of attraction of the minimum $(\theta_1, \theta_2) = (0, \pi/2)$. The differential equation solver `NDSolve` within Mathematica generated the simulation results reported in Figure 5.1.

We conclude the example with a final remark. The stabilization result is not surprising, and it intuitively agrees with the classic example in [6], where the controlled variable is the speed of the joint connected to the second link and where the joint itself is constrained to move vertically.

6. Conclusions. This paper provides a systematic study of high magnitude high frequency averaging for mechanical systems. The averaging extends the results of earlier works in two directions. First, the analysis applies to the multi-input setting where controls are not necessarily applied to cyclic variables. Instead, forces are described as generic one-forms. Additionally, our analysis applies to the case of mechanical systems with nonholonomic constraints. From a control design viewpoint, the improved analysis leads to sufficient tests for an appropriate notion of vibrational stabilization.

At the heart of the proposed approach is a detailed analysis of the Lie algebraic structure of mechanical systems (with or without constraints, with or without non-integrable forces). It is this structure that enables closed form expressions for the averaging analysis. Furthermore, it is this same structure that underlies the controllability analysis in [32]. Our analysis provides a missing link between the notions of averaged potential [3] and symmetric product [32].

Numerous extensions appear promising. First, one could pursue generalizations to high order averaging and applications in the field of robotic motion planning; see [17, 16]. Second, the setting of distributed parameter systems with Lagrangian structure might provide a number of interesting applications and further theoretical challenges. Finally, the tools developed here might shed new light on the problem of existence and stability of limit cycles in the study of animal and robotics locomotion.

Appendix. The variations of constants formula in geometric terms.

LEMMA A.1. *Let f, g be smooth time-varying vector fields on $\mathbb{R}^n$. Let $x_0 \in \mathbb{R}^n$, and let $T \in \mathbb{R}^n$ be small enough so that the flow map $\Phi^g_{0,T}$ is a local diffeomorphism in a neighborhood of x_0. The final value $x(T) = \Phi^{f+g}_{0,T}(x_0)$ can be written as*

$$x(T) = \Phi^g_{0,T}(z(T)), \tag{A.1}$$

$$\dot{z}(t) = ((\Phi^g_{0,t})^* f)(z), \quad z(0) = x_0. \tag{A.2}$$

Additionally, we have the formal equality

$$((\Phi^g_{0,t})^* f)(t,x) = f(x) + \sum_{k=1}^{\infty} \int_0^t \dots \int_0^{s_{k-1}} \left(\mathrm{ad}_{g(s_k,x)} \dots \mathrm{ad}_{g(s_1,x)} f(x)\right) ds_k \dots ds_1. \tag{A.3}$$

Proof. Let $x(T) = \Phi^{f+g}_{0,T}(x_0)$, and let $y(T) = \Phi^g_{0,T}(z(T))$, where $z(t)$ is computed via (A.2). We compute

$$\begin{aligned} \dot{z} &= ((\Phi^g_{0,t})^* f)(z) = (T_z(\Phi^g_{0,t})^{-1} \circ f \circ \Phi^g_{0,t})(z) \\ &= \left(T_z \Phi^g_{0,t}\right)^{-1} \circ f(y(t),t) \end{aligned}$$

so that

$$\begin{aligned} \dot{y}(t) &= \frac{d}{dt}\left(\Phi^g_{0,t}(z(t))\right) = g\left(\Phi^g_{0,t}(z(t)),t\right) + \left(T_z \Phi^g_{0,t}(z(t))\right) \dot{z} \\ &= g\left(y(t),t\right) + \left(T_z \Phi^g_{0,t}(z(t))\right) \dot{z} = g\left(y(t),t\right) + f(y(t),t). \end{aligned}$$

Therefore, $y(t)$ obeys the same differential equation as $x(t)$. Since it is also clear that $x(0) = y(0)$, the curves x and y must be equal.

Next, we investigate the pull-back of f along the flow of g. We assume f to be time-invariant and g time-varying. The following statement is proved in [1, Theorem 4.2.31] and in [2, equation (3.3)]:

$$\frac{d}{dt}((\Phi^g_{0,t})^* f)(t,x) = (\Phi^g_{0,t})^*[g(t,x), f(x)],$$

where the Lie bracket between g and f is computed at t fixed. At fixed $x \in \mathbb{R}^n$, we integrate the previous equation from time 0 to t to obtain

$$((\Phi^g_{0,t})^* f)(t,x) = f(x) + \int_0^t (\Phi^g_{0,s})^*[g(s,x), f(x)] ds.$$

The formal expansion in (A.3) follows from iteratively applying the previous equality. ∎

REFERENCES

[1] R. Abraham, J. E. Marsden, and T. S. Ratiu, *Manifolds, Tensor Analysis, and Applications*, 2nd ed., Appl. Math. Sci. 75, Springer-Verlag, New York, 1988.

[2] A. A. Agrachev and R. V. Gamkrelidze, *The exponential representation of flows and the chronological calculus*, Sb. Math., 35 (1978), pp. 727–785.

[3] J. Baillieul, *Stable average motions of mechanical systems subject to periodic forcing*, in Dynamics and Control of Mechanical Systems: The Falling Cat and Related Problems, Fields Inst. Commun. 1, M. J. Enos, ed., AMS, Providence, RI, 1993, pp. 1–23.

[4] J. Baillieul, *Energy methods for stability of bilinear systems with oscillatory inputs*, Internat. J. Robust Nonlinear Control, 5 (1995), pp. 285–301.

[5] J. Baillieul, *The geometry of controlled mechanical systems*, in Mathematical Control Theory, J. Baillieul and J. C. Willems, eds., Springer-Verlag, New York, 1998, pp. 322–354.

[6] J. Baillieul and B. Lehman, *Open-loop control using oscillatory inputs*, in CRC Control Handbook, W. S. Levine, ed., CRC Press, Boca Raton, FL, 1996, pp. 967–980.

[7] J. Baillieul and S. Weibel, *Scale dependence in the oscillatory control of micromechanisms*, in Proceedings of the IEEE Conference on Decision and Control, Tampa, FL, 1998, pp. 3058–3063.

[8] N. S. Bedrossian and M. W. Spong, *Feedback linearization of robot manipulators and Riemannian curvature*, J. Robotic Systems, 12 (1995), pp. 541–552.

[9] R. E. Bellman, J. Bentsman, and S. M. Meerkov, *Vibrational control of nonlinear systems: Vibrational controllability and transient behavior*, IEEE Trans. Automat. Control, 31 (1986), pp. 717–724.

[10] R. E. Bellman, J. Bentsman, and S. M. Meerkov, *Vibrational control of nonlinear systems: Vibrational stabilization*, IEEE Trans. Automat. Control, 31 (1986), pp. 710–716.

[11] J. Bentsman, *Vibrational control of a class of nonlinear systems by nonlinear multiplicative vibrations*, IEEE Trans. Automat. Control, 32 (1987), pp. 711–716.

[12] A. M. Bloch and P. E. Crouch, *Nonholonomic control systems on Riemannian manifolds*, SIAM J. Control Optim., 33 (1995), pp. 126–148.

[13] N. N. Bogoliubov and Y. A. Mitropolsky, *Asymptotic Methods in the Theory of Non-Linear Oscillations*, Gordon and Breach, New York, 1961.

[14] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.

[15] F. Bullo, *Stabilization of relative equilibria for underactuated systems on Riemannian manifolds*, Automatica J. IFAC, 36 (2000), pp. 1819–1834.

[16] F. Bullo, *Series expansions for the evolution of mechanical control systems*, SIAM J. Control Optim., 40 (2001), pp. 166–190.

[17] F. Bullo, N. E. Leonard, and A. D. Lewis, *Controllability and motion algorithms for underactuated Lagrangian systems on Lie groups*, IEEE Trans. Automat. Control, 45 (2000), pp. 1437–1454.

[18] F. Bullo and R. M. Murray, *Tracking for fully actuated mechanical systems: A geometric framework*, Automatica J. IFAC, 35 (1999), pp. 17–34.

[19] S.-N. Chow, *Methods of Bifurcation Theory*, Springer-Verlag, New York, 1982.

[20] P. E. Crouch, *Geometric structures in systems theory*, Proc. IEE-D, 128 (1981), pp. 242–252.

[21] M. P. Do Carmo, *Riemannian Geometry*, Birkhäuser Boston, Boston, 1992.

[22] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Appl. Math. Sci. 42, Springer-Verlag, New York, 1990.

[23] H. Hermes, *Nilpotent and high-order approximations of vector field systems*, SIAM Rev., 33 (1991), pp. 238–264.

[24] S. Hirose, *Biologically Inspired Robots: Snake-Like Locomotors and Manipulators*, Oxford University Press, Oxford, UK, 1993.

[25] K.-S. Hong, K.-R. Lee, and K.-I. Lee, *Vibrational control of underactuated mechanical systems: control design through the averaging analysis*, in Proceedings of the IEEE American Control Conference, Philadelphia, 1998, pp. 3482–3486.

[26] A. Isidori, *Nonlinear Control Systems*, 3rd ed., Springer-Verlag, New York, 1995.

[27] M. Kawski, *Geometric homogeneity and applications to stabilization*, in Proceedings of the Nonlinear Control Systems Design Symposium, Tahoe City, CA, 1995, pp. 251–256.

[28] D. E. Koditschek, *The application of total energy as a Lyapunov function for mechanical control systems*, in Dynamics and Control of Multibody Systems, Contemp. Math. 97, J. E. Marsden, P. S. Krishnaprasad, and J. C. Simo, eds., AMS, Providence, RI, 1989, pp. 131–157.

[29] P. S. Krishnaprasad and D. P. Tsakiris, *Oscillations, SE(2)-snakes and motion control*, Dyn. Stab. Syst., 16 (2001), pp. 347–397.
[30] A. D. Lewis, *Affine connections and distributions with applications to nonholonomic mechanics*, Rep. Math. Phys., 42 (1998), pp. 135–164.
[31] A. D. Lewis, *Simple mechanical control systems with constraints*, IEEE Trans. Automat. Control, 45 (2000), pp. 1420–1436.
[32] A. D. Lewis and R. M. Murray, *Configuration controllability of simple mechanical control systems*, SIAM J. Control Optim., 35 (1997), pp. 766–790.
[33] A. D. Lewis and R. M. Murray, *Decompositions of control systems on manifolds with an affine connection*, Systems Control Lett., 31 (1997), pp. 199–205.
[34] P. Libermann and C.-M. Marle, *Symplectic Geometry and Analytical Mechanics*, Math. Appl. 35, Reidel, Dordrecht, The Netherlands, 1987.
[35] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*, 2nd ed., Springer-Verlag, New York, 1999.
[36] R. M. Murray, Z. X. Li, and S. S. Sastry, *A Mathematical Introduction to Robotic Manipulation*, CRC Press, Boca Raton, FL, 1994.
[37] Y. Nakamura, T. Suzuki, and M. Koinuma, *Nonlinear behavior and control of a nonholonomic free-joint manipulator*, IEEE Trans. Robotics and Automation, 13 (1997), pp. 853–862.
[38] H. Nijmeijer and A. J. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.
[39] L. Noakes, G. Heinzinger, and B. Paden, *Cubic splines on curved spaces*, IMA J. Math. Control Inform., 6 (1989), pp. 465–473.
[40] J. P. Ostrowski and J. W. Burdick, *The geometric mechanics of undulatory robotic locomotion*, Internat. J. Robotics Research, 17 (1998), pp. 683–701.
[41] M. Rathinam and R. M. Murray, *Configuration flatness of Lagrangian systems underactuated by one control*, SIAM J. Control Optim., 36 (1998), pp. 164–179.
[42] J. A. Sanders and F. Verhulst, *Averaging Methods in Nonlinear Dynamical Systems*, Springer-Verlag, New York, 1985.
[43] A. Seifert and L. G. Pack, *Oscillatory control of separation at high Reynolds numbers*, AIAA J., 37 (1999), pp. 1062–1071.
[44] E. D. Sontag, *Mathematical Control Theory: Deterministic Finite-Dimensional Systems*, 2nd ed., Texts Appl. Math. 6, Springer-Verlag, New York, 1998.
[45] E. D. Sontag and H. J. Sussmann, *Time-optimal control of manipulators*, in Proceedings of the IEEE International Conference on Robotics and Automation, San Francisco, CA, 1986, pp. 1692–1697.
[46] M. Takegaki and S. Arimoto, *A new feedback method for dynamic control of manipulators*, ASME J. Dynamic Systems, Measurement, and Control, 102 (1981), pp. 119–125.
[47] A. J. van der Schaft, *L2-Gain and Passivity Techniques in Nonlinear Control*, 2nd ed., Springer-Verlag, New York, 1999.
[48] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.
[49] V. S. Varadarajan, *Lie Groups, Lie Algebras, and Their Representations*, Grad. Texts in Math. 102, Springer-Verlag, New York, 1984.
[50] S. P. Weibel, *Applications of Qualitative Methods in the Nonlinear Control of Superarticolated Mechanical Systems*, Ph.D. thesis, Boston University, Boston, MA, 1997.
[51] S. P. Weibel and J. Baillieul, *Open-loop oscillatory stabilization of an n-pendulum*, Internat. J. Control, 71 (1998), pp. 931–57.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 2, pp. 563–582

EXTENSION OF THE PERRON–FROBENIUS THEOREM TO HOMOGENEOUS SYSTEMS*

DIRK AEYELS† AND PATRICK DE LEENHEER‡

Abstract. This paper deals with a particular class of positive systems. The state components of a positive system are positive or zero for all positive times. These systems are often encountered in applied areas such as chemical engineering or biology. It is shown that for this particular class the first orthant contains an invariant ray in its interior. An invariant ray generalizes the concept of an eigenvector of linear systems to nonlinear homogeneous systems. Then sufficient conditions for uniqueness of this ray are given. The main result states that the vector field on an invariant ray determines the stability properties of the zero solution with respect to initial conditions in the first orthant. The asymptotic behavior of the solutions is examined. Finally, we compare our results to the Perron–Frobenius theorem, which gives a detailed picture of the dynamical behavior of positive linear systems.

Key words. positive systems, cooperative systems, homogeneous systems, monotone flows, global asymptotic stability

AMS subject classifications. 37C10, 37C65, 34D23, 34D05

PII. S0363012900361178

1. Introduction.

A dynamical system is said to be *positive* if it leaves the first orthant of $\mathbb{R}^n$ invariant for future times when initiated in this orthant. Examples of these systems abound in a variety of applied areas such as biology, chemistry, economics, and sociology [6], [14], [9]. In a biological system, for example, a state component will typically be the number of individuals of a certain species in a population of interacting species. State components in a chemical system are typically concentrations or amounts of chemical substances. Important issues arising in the study of positive systems are the boundedness of solutions, permanence or persistence, and the (asymptotic) stability of equilibrium points. In this paper the stability properties of the zero solution of a class of positive systems is considered. At first glance, it might be surprising that we are interested in the zero solution, because in most applications this solution is not very interesting. For example, it corresponds to death of all species in a biological context, or to washout of all chemicals in chemical engineering. In the context of positive systems, nontrivial equilibrium points are of much more interest. However, these equilibrium points arise in models where some type of control action is already present in the model, although implicitly.

As an illustration, consider the simplest example of the well-known predator-prey model proposed by Volterra (see [6]) to explain the observed oscillations in the biomass of prey species (denoted by x) and predator species (denoted by y):

$$\dot{x} = -axy + bx,$$
$$\dot{y} = cxy - dy,$$

*Received by the editors September 8, 2000; accepted for publication (in revised form) November 21, 2001; published electronically June 26, 2002. This paper presents research results of the Belgian Programme on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology, and Culture. The scientific responsibility rests with its authors.
http://www.siam.org/journals/sicon/41-2/36117.html

†Universiteit Gent, SYSTeMS, Technologiepark-Zwijnaarde 9, 9052 Gent, Belgium (dirk.aeyels@rug.ac.be).

‡Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287 (leenheer@math.la.asu.edu).

where a, b, c, and d are positive constants. The term bx is the growth rate of the prey species and suggests the availability of a feeding source. The abundance of the food is represented by the parameter b, which can be interpreted as the implicit control action hinted at before. If there is no food ($b = 0$), then the only equilibrium point of the system is the trivial one. If there is food available ($b > 0$), then there is a nontrivial equilibrium point. It might not come as a surprise that the stability behavior of the zero solution of the simpler system with $b = 0$ determines to some extent the behavior of the system with $b > 0$. (Bifurcation theory might serve as a tool here.) This motivates the study of the zero solution for the case in which $b = 0$.

We provide another example of a chemical reactor at the end of section 5.

Therefore, the study of the stability behavior of the zero solution is not only interesting in its own right, but also important for control purposes.

Homogeneous systems on $\mathbb{R}^n$ (see, e.g., [3]) are a particular class of nonlinear systems. *Invariant rays*—when they exist—play an important role in the study of the stability behavior of the zero solution of a homogeneous system. An invariant ray is a particular curve that is invariant for the flow of the system. It can be interpreted as the generalization of the concept of the linear space spanned by an eigenvector of a real eigenvalue, in the context of *linear* systems, to the context of *homogeneous*—and thus generally nonlinear—systems. Suppose that a homogeneous system possesses a number of invariant rays. A necessary condition for global asymptotic stability (GAS) of the zero solution is that the vector field on all invariant rays points towards the origin. Although this condition is not sufficient in general for GAS, we introduce a particular class of homogeneous systems for which invariant rays determine the stability behavior.

Within the class of positive systems, cooperative and irreducible systems are well examined [5], [13]. In this paper a class of positive systems is introduced, characterized by *homogeneous* cooperative and irreducible vector fields. It is shown that these systems enjoy a fairly simple dynamical behavior. This may come as a surprise, since it is well known that the behavior of homogeneous systems is in principle as involved as the behavior of general nonlinear systems.

Next, to describe our results in some detail, we digress to discuss positive *linear* systems, known to model a number of important physical systems; see, e.g., [9]. A necessary and sufficient condition for a linear system $\dot{x} = Ax$ to be positive is that A be a Metzler matrix (i.e., have nonnegative off-diagonal entries) or equivalently that the system be cooperative. The principal tool for the analysis of the (stability) behavior of a positive linear system is the Perron–Frobenius theorem. It is natural to ask whether it is possible to generalize this to classes of positive *nonlinear* systems. The purpose of this paper is to show that this is indeed the case for homogeneous cooperative and irreducible systems.

First of all it is shown that these systems *always* possess an invariant ray in the interior of the first orthant. If the order of the homogeneous vector field is equal to zero, then this ray is unique in the first orthant; if the order of the vector field is strictly greater than zero, then the ray is unique in the first orthant if the vector field on this ray does not point away from the origin. In both cases the stability behavior of the zero solution of the system is determined by the behavior of the system on this unique invariant ray.

Several invariant rays may exist if the order of the homogeneous vector field is strictly greater than zero and if the vector field on every invariant ray points away from the origin. In this case the stability behavior of the zero solution of the system

is also determined by the behavior of the system on the invariant rays.

This paper is organized as follows. Basic definitions are given in section 2. In section 3, results on homogeneous systems are reviewed. A class of positive homogeneous systems is introduced in section 4, leading to a criterion for GAS of the zero solution with respect to the initial conditions in $\mathbb{R}^n_+$ (section 5). The asymptotic behavior of solutions of this class of systems is examined in section 6. The paper is concluded in section 7 with a discussion; in particular, the classical Perron–Frobenius theorem for linear differential equations (see [1] or [13]) is compared to the results of this paper.

2. Preliminaries. Let $\mathbb{R}$ be the set of real numbers, and $\mathbb{R}^n$ the set of n-tuples with all components belonging to $\mathbb{R}$. For $x \in \mathbb{R}^n$, $|x|$ is the Euclidean norm of x. $\mathbb{R}^+ := [0, +\infty)$, $\mathbb{R}^+_0 := (0, +\infty)$, and $\mathbb{R}^n_+$ $(\mathrm{int}(\mathbb{R}^n_+))$ is the set of n-tuples with all components belonging to $\mathbb{R}^+$ $(\mathbb{R}^+_0)$. Finally, $\mathrm{bd}(\mathbb{R}^n_+) := \mathbb{R}^n_+ \setminus \mathrm{int}(\mathbb{R}^n_+)$ is the boundary of $\mathbb{R}^n_+$.

Let $x, y \in \mathbb{R}^n_+$; then $x \leq y$ means $x_i \leq y_i$ $\forall i = 1, \ldots, n$. Furthermore, $x < y$ if and only if $x \leq y$ and $x \neq y$, while $x \ll y$ if and only if $x_i < y_i$ $\forall i = 1, \ldots, n$. For subsets U and V of $\mathbb{R}^n_+$, we denote $U \leq (<, \ll) V$ if $x \leq (<, \ll) y$ $\forall x \in U$ and $y \in V$.

For $x \in \mathbb{R}^n$, $\mathrm{diag}(x)$ stands for an $n \times n$ diagonal matrix, where the ith diagonal entry is equal to x_i, the ith component of the vector x. A real $n \times n$ matrix $A = (a_{ij})$ is *Metzler* if and only if its off-diagonal entries a_{ij}, $\forall i \neq j$, belong to $\mathbb{R}^+$.

A real $n \times n$ matrix $A = (a_{ij})$ is *reducible* if and only if the index set $N := \{1, 2, \ldots, n\}$ can be split into two sets J and K, with $J \cup K = N$ and $J \cap K = \emptyset$ such that $a_{jk} = 0$ $\forall j \in J$ and $k \in K$.

It is clear that A is reducible if and only if there exists a permutation matrix P such that

$$PAP^T = \begin{pmatrix} B & 0 \\ C & D \end{pmatrix},$$

where B and D are square matrices.

The *standard basis* of the vector space $\mathbb{R}^n$ is given by $\{e_i | i \in N\}$, where the ith entry of e_i is equal to 1, while the other entries are equal to 0. An m-dimensional *coordinate subspace* of $\mathbb{R}^n$ is a subspace of $\mathbb{R}^n$ with a basis $\{e_{k_1}, e_{k_2}, \ldots, e_{k_m}\}$, with $1 \leq k_1 < k_2 < \cdots < k_m$.

The matrix A is reducible if and only if the linear operator, associated to the matrix A and the standard basis, has an m-dimensional invariant coordinate subspace with $1 \leq m < n$.

When A is not reducible, it is *irreducible*.

Consider the system

$$\dot{x} = f(x), \tag{1}$$

where $x \in \mathbb{R}^n$ and $f(x)$ is a continuous vector field on $\mathbb{R}^n$, continuously differentiable (of class C^1) on $\mathbb{R}^n \setminus \{0\}$, and such that $f(0) = 0$. Later we give conditions such that the uniqueness of solutions for system (1) is guaranteed. The *forward solution* of system (1) with initial condition $x_0 \in \mathbb{R}^n$ at $t = 0$ is denoted as $x(t, x_0)$, $t \in \mathcal{I}_{x_0} := [0, T_{\max}(x_0))$, where $\mathcal{I}_{x_0}$ is the *maximal forward interval of existence*. A set $D \subset \mathbb{R}^n$ is *forward invariant* for system (1) if and only if $\forall x_0 \in D$, $x(t, x_0) \in D$ $\forall t \in \mathcal{I}_{x_0}$. System (1) is *positive* if and only if $\mathbb{R}^n_+$ is forward invariant.

Suppose that $D \subset \mathbb{R}^n$ is a forward invariant set for system (1). The flow of system (1) is *monotone in* D if and only if $\forall x_0$, $y_0 \in D$ with $x_0 \leq (<, \ll) y_0$ it holds that $x(t, x_0) \leq (<, \ll) x(t, y_0)$ $\forall t \in (\mathcal{I}_{x_0} \cap \mathcal{I}_{y_0})$.

The flow of system (1) is *strongly monotone in D* if and only if it is monotone in D and $\forall x_0, y_0 \in D$ with $x_0 < y_0$ it holds that $x(t, x_0) \ll x(t, y_0)\ \forall t \in (\mathcal{I}_{x_0} \cap \mathcal{I}_{y_0}) \setminus \{0\}$.

A point $p \in \mathbb{R}^n$ is an *omega limit point* of x_0 if there exists an increasing sequence of time instances $\{t_k\}$, with $t_k \to +\infty$ when $k \to +\infty$, such that $\lim_{t_k \to +\infty} x(t_k, x_0) = p$. The set of all omega limit points of x_0 is the *omega limit set* of x_0 and is denoted by $\omega(x_0)$. Notice that the *omega limit set* of x_0 may be the empty set, for instance if the solution starting in x_0 diverges. If $T_{\max(x_0)} = +\infty$, then the set $\mathcal{O}(x_0) := \{x(t, x_0) | t \in \mathbb{R}^+\}$ is the *forward orbit* of the forward solution $x(t, x_0)$. It follows from classical results on the theory of ordinary differential equations that if $\mathrm{cl}(\mathcal{O}(x_0))$, the closure of the forward orbit $\mathcal{O}(x_0)$, is compact, then $\omega(x_0)$ is nonempty and compact and $d(\omega(x_0), x(t, x_0)) \to 0$ when $t \to +\infty$ (where $d(A, z) := \inf_{y \in A} d(y, z)$ and $d(y, z)$ is the Euclidean distance between y and z).

3. Homogeneous systems. In this section we review the concept of a homogeneous system and discuss some of its properties. Many of these results are known, and no originality is claimed here. However, we have chosen to include the proofs to make the paper self-contained.

3.1. Definition and Euler's formula. We first introduce the concept of a homogeneous vector field [12].

DEFINITION 3.1. *A vector field $f(x)$, $x \in \mathbb{R}^n$, is* homogeneous *of order $\tau \in \mathbb{R}$ with respect to the dilation map $\delta^r_\lambda(x) : \mathbb{R}^n \to \mathbb{R}^n$ with $\delta^r_\lambda(x) = (\lambda^{r_1} x_1, \lambda^{r_2} x_2, \ldots, \lambda^{r_n} x_n)$, where $r := (r_1, r_2, \ldots, r_n)$ is a fixed n-tuple$(r_i \in \mathbb{R}^+_0 \forall i \in N)$, and $\forall \lambda \in \mathbb{R}^+_0$ if and only if*

$$\forall x \in \mathbb{R}^n, \lambda \in \mathbb{R}^+_0 \qquad f(\delta^r_\lambda(x)) = \lambda^\tau \delta^r_\lambda(f(x)). \tag{2}$$

To every n-tuple of positive real numbers, one can associate a dilation map $\delta^r_\lambda(x)$. When $r = (1, 1, \ldots, 1)$, then $\delta^r_\lambda(x)$ is the *standard dilation map*.

System (1) is *homogeneous* if $f(x)$ is homogeneous.

We introduce the following hypothesis:

($\mathcal{H}1$) $f(x)$ is a homogeneous vector field of order $\tau \in \mathbb{R}^+$ with respect to a dilation map $\delta^r_\lambda(x)$.

Notice that if ($\mathcal{H}1$) holds, then $f(0) = 0$ (by continuity of f on $\mathbb{R}^n$), and thus $x = 0$ is an equilibrium point of system (1). Since $f(x)$ is C^1 on $\mathbb{R}^n \setminus \{0\}$, solutions starting in $\mathbb{R}^n \setminus \{0\}$ exist and are unique. On the other hand, the vector field $f(x)$ is only continuous at $x = 0$. This implies that a solution starting in $x = 0$ exists (the zero solution satisfies the differential equation) but might not be unique. The additional hypothesis ($\mathcal{H}1$) excludes the possibility that there are multiple solutions starting in $x = 0$ as proved in [10].

Let U be an open subset of $\mathbb{R}^n$, and suppose that $f(x)$ is a homogeneous vector field of order τ with respect to the dilation map $\delta^r_\lambda(x)$ and of class C^1 on $\mathbb{R}^n$. For future reference we recall Euler's formula:

$$\forall x \in U \qquad \frac{\partial f}{\partial x}(x) \mathrm{diag}(r) x = \mathrm{diag}(r + \tau^*) f(x),$$

where $\tau^* := (\tau, \ldots, \tau)$. This formula is easily proved by first taking the derivative with respect to λ on both sides of (2) and then evaluating the resulting equation for $\lambda = 1$.

3.2. Invariant rays. For $x \in \mathbb{R}^n \setminus \{0\}$ and a fixed but arbitrary dilation map $\delta^r_\lambda(x)$, $R_x := \{\delta^r_\lambda(x) | \lambda \in \mathbb{R}^+_0\}$ is the *ray through x*.

The ω limit sets of points on a ray are related as follows.

LEMMA 3.2. *If system* (1) *satisfies* ($\mathcal{H}1$) *and if* $p \in \omega(x_0)$, *then* $\delta_\lambda^r(p) \in \omega(\delta_\lambda^r(x_0))$.

This follows immediately from the scaling property [7] of solutions of homogeneous differential equations. By the scaling property we mean the following: Suppose that $x(t, x_0)$, $t \in [0, T_{\max}(x_0))$, is a solution of system (1). Then $\forall \lambda \in \mathbb{R}_0^+$ the term $\delta_\lambda^r(x(\lambda^\tau t, x_0))$, $t \in [0, \frac{T_{\max}(x_0)}{\lambda^\tau})$, is also a solution of system (1).

LEMMA 3.3. *If system* (1) *satisfies* ($\mathcal{H}1$) *and if there exists a point* $\bar{x} \in \mathbb{R}^n \setminus \{0\}$ *such that*

$$f(\bar{x}) = \gamma_{\bar{x}} \mathrm{diag}(r) \bar{x} \tag{3}$$

for some $\gamma_{\bar{x}} \in \mathbb{R}$, *then the vector field* $f(x)$ *is tangent to* $R_{\bar{x}}$ *at each point of* $R_{\bar{x}}$.

Proof. Indeed, $\frac{d}{d\lambda}(\delta_\lambda^r(\bar{x}))|_{\lambda=1} = \mathrm{diag}(r)\bar{x}$. This and (3) imply that the vector field $f(x)$ is tangent to $R_{\bar{x}}$ at the point $\bar{x}$. Moreover, $\forall \lambda \in \mathbb{R}_0^+$

$$f(\delta_\lambda^r(\bar{x})) = (\gamma_{\bar{x}} \lambda^\tau) \mathrm{diag}(r) \delta_\lambda^r(\bar{x}), \tag{4}$$

and thus the vector field $f(x)$ is tangent to $R_{\bar{x}}$ in *every* point of $R_{\bar{x}}$, which proves the lemma. Notice that (4) implies that $\forall \lambda \in \mathbb{R}_0^+$

$$\gamma_{\delta_\lambda^r(\bar{x})} = \gamma_{\bar{x}} \lambda^\tau. \qquad \square \tag{5}$$

Suppose that there exists a point $\bar{x} \in \mathbb{R}^n \setminus \{0\}$ such that (3) holds. Then it follows from Lemma 3.3 that the forward (and backward) solution of system (1), starting in an arbitrary point of $R_{\bar{x}}$, stays on this ray for all future (and past) times for which this solution is defined. Such a ray is an *invariant ray* for system (1).

An invariant ray R_y is *asymptotically stable*, *stable*, or *unstable* if and only if $\gamma_x < 0$, $\gamma_x \leq 0$, respectively, $\gamma_x > 0$ for some $x \in R_y$ and by (5) for any $x \in R_y$.

An easy calculation shows that solutions starting on an invariant ray $R_{\bar{x}}$ satisfy the set of decoupled differential equations

$$\dot{x}_k = \left(\gamma_z r_k z_k^{-\frac{\tau}{r_k}}\right) x_k^{1+\frac{\tau}{r_k}} \qquad \forall k \in N \tag{6}$$

for some $z \in R_{\bar{x}}$, and thus $f(z) = \gamma_z \mathrm{diag}(r) z$ for some $\gamma_z \in \mathbb{R}$. Notice that invariant rays do not always exist for homogeneous systems. The linear harmonic oscillator, for example, is homogeneous of order zero with respect to the standard dilation map but does not possess an invariant ray.

3.3. Projection of a homogeneous system. We next introduce the concept of a homogeneous norm.

DEFINITION 3.4. *A* homogeneous norm *associated to the dilation map* $\delta_\lambda^r(x)$ *is a function* $\rho : \mathbb{R}^n \to \mathbb{R}^+$ *satisfying the following:*

1. $\rho(x)$ *is continuous on* $\mathbb{R}^n$ *and of class* C^1 *on* $\mathbb{R}^n \setminus \{0\}$.
2. $\rho(x) = 0$ *only if* $x = 0$.
3. $\forall x \in \mathbb{R}^n$ *and* $\forall \lambda \in \mathbb{R}_0^+$, $\rho(\delta_\lambda^r(x)) = \lambda \rho(x)$.

For example, the function

$$\left(\sum_{i=1}^n |x_i|^{\frac{p}{r_i}} \right)^{\frac{1}{p}}, \tag{7}$$

with $p > \max_{i=1,\ldots,n}(r_i)$, is a homogeneous norm.

Pick an arbitrary homogeneous norm $\rho(x)$. Then the *ρ-homogeneous unit $(n-1)$-sphere* is defined as $S_\rho := \{x \in \mathbb{R}^n | \rho(x) = 1\}$.

LEMMA 3.5. *Suppose that system* (1) *satisfies* ($\mathcal{H}1$). *Consider the system*

$$\dot{x} = g(x) := \begin{cases} \frac{1}{(\rho(x))^\tau} f(x) & \text{for } x \in \mathbb{R}^n_+ \setminus \{0\}, \\ 0 & \text{for } x = 0. \end{cases} \tag{8}$$

Then $g(x)$ is a continuous vector field on $\mathbb{R}^n$, of class C^1 on $\mathbb{R}^n \setminus \{0\}$, and homogeneous of order zero with respect to $\delta^r_\lambda(x)$.

Proof. Indeed, it is clear that $g(x)$ is continuous on $\mathbb{R}^n \setminus \{0\}$, so it only has to be shown that $g(x)$ is continuous at $x = 0$. Pick a sequence $\{x_k\} \to 0$ when $k \to +\infty$. Associated to this sequence is the sequence $\{x'_k\} \subset S_\rho$ with

$$x_k = \delta^r_{\lambda_k}(x'_k) \tag{9}$$

for suitable $\lambda_k \in \mathbb{R}^+_0$ and such that $\{\lambda_k\} \to 0$ when $k \to +\infty$. Then

$$\begin{aligned} \lim_{k \to +\infty} \frac{1}{(\rho(x_k))^\tau} f(x_k) &= \lim_{\lambda_k \to 0} \frac{1}{(\rho(\delta^r_{\lambda_k}(x'_k)))^\tau} f(\delta^r_{\lambda_k}(x'_k)) \\ &= \lim_{\lambda_k \to 0} \frac{1}{\lambda_k^\tau (\rho(x'_k))^\tau} \lambda_k^\tau \delta^r_{\lambda_k}(f(x'_k)) \text{ by homogeneity of } f(x) \text{ and of } \rho(x) \\ &= \lim_{\lambda_k \to 0} \delta^r_{\lambda_k}(f(x'_k)) \qquad \text{because } \rho(x'_k) = 1 \text{ as } \{x'_k\} \subset S_\rho \\ &= 0. \end{aligned}$$

Since $f(x)$ and $\rho(x)$ are of class C^1 on $\mathbb{R}^n \setminus \{0\}$, and since $\rho(x) > 0 \ \forall x \in \mathbb{R}^n \setminus \{0\}$, it follows that $g(x)$ is of class C^1 on $\mathbb{R}^n \setminus \{0\}$. It is also easily verified that $g(x)$ is homogeneous of order *zero* with respect to $\delta^r_\lambda(x)$. ∎

Uniqueness of the solutions for system (8) is guaranteed. A proof of this assertion proceeds along the same lines of the proof of the uniqueness of solutions of system (1). In addition, system (8) is topologically equivalent to system (1). Indeed, the direction of both the vectors $f(x)$ and $g(x)$ is the same $\forall x \in \mathbb{R}^n$. This implies that the solutions of system (1) are transformed to solutions of system (8) by a change in time scale. In particular, a ray is invariant for system (8) if and only if it is invariant for system (1), and system (8) is positive if and only if system (1) is positive.

With system (8) we now associate a system defined on S_ρ as follows. Consider the projection map $\pi : \mathbb{R}^n \setminus \{0\} \to S_\rho$ with $\pi(x) := \delta^r_{1/\rho(x)}(x)$. Notice that

$$\pi = \pi \circ \delta^r_\lambda \tag{10}$$

$\forall \lambda \in \mathbb{R}^+_0$. This means that the image of a ray under π is a unique point of S_ρ. Geometrically, $\pi(x)$ is the intersection of the ray through x and S_ρ.

Now pick a point $m \in \mathbb{R}^n \setminus \{0\}$ and consider R_m. It will now be shown that *for all points m'* of the ray R_m, the tangent mapping of π maps the vector $g(m')$ to the *same* vector at $\pi(m)$. More precisely, it will be shown that

$$T_m \pi(g(m)) = T_{m'} \pi(g(m')) \tag{11}$$

$\forall m' \in R_m$, where $T_m \pi$ is the derivative of π at m.

For $m' \in R_m$ there exists by the definition of R_m a $\tilde{\lambda} \in \mathbb{R}^+_0$ such that

$$m' = \delta^r_{\tilde{\lambda}}(m). \tag{12}$$

Then

$$\begin{aligned} T_m\pi(g(m)) &= T_m(\pi \circ \delta^r_{\tilde{\lambda}})(g(m)) && \text{by (10)} \\ &= T_{\delta^r_{\tilde{\lambda}}(m)}\pi \circ T_m\delta^r_{\tilde{\lambda}}(g(m)) && \text{by the chain rule} \\ &= T_{m'}\pi \circ T_m\delta^r_{\tilde{\lambda}}(g(m)) && \text{by (12)} \\ &= T_{m'}\pi \circ \delta^r_{\tilde{\lambda}}(g(m)) && \text{by linearity of the dilation map} \\ &= T_{m'}\pi(g(m')) && \text{by homogeneity of } g \text{ of order zero and (12).} \end{aligned}$$

Since the preimage of a point $y \in S_\rho$, $\pi^{-1}(y)$ is equal to the ray through y, R_y, and by (11), it follows that $\forall y \in S_\rho$ a unique tangent vector $h(y) \in T_yS_\rho$ can be defined as follows:

$$h(y) = T_m\pi(g(m)) \qquad \forall m \in R_y. \tag{13}$$

It remains to be shown that h defines a vector field of class C^1 on S_ρ. This is done by showing that for every C^∞ function $\tilde{f} : S_\rho \to \mathbb{R}$ the function $h(\tilde{f}) : S_\rho \to \mathbb{R}$ is of class C^1 on S_ρ.

For all $m \in \mathbb{R}^n \setminus \{0\}$ it holds that

$$\begin{aligned} h(\pi(m))(\tilde{f}) &= T_m\pi(g(m))(\tilde{f}) && \text{by (13)} \\ &= g(m)(\tilde{f} \circ \pi) && \text{by definition of the tangent mapping.} \end{aligned}$$

This implies that

$$h(\tilde{f}) \circ \pi|_m = g(\tilde{f} \circ \pi)|_m \tag{14}$$

$\forall m \in \mathbb{R}^n \setminus \{0\}$, and thus $h(\tilde{f}) \circ \pi = g(\tilde{f} \circ \pi)$. Denoting the canonical injection by $j : S_\rho \to \mathbb{R}^n \setminus \{0\}$, we obtain that

$$h(\tilde{f}) \circ \pi \circ j = g(\tilde{f} \circ \pi) \circ j. \tag{15}$$

Since $\pi \circ j$ is the identity mapping on S_ρ, it follows that $h(\tilde{f}) = g(\tilde{f} \circ \pi) \circ j$, which is clearly of class C^1 on S_ρ.

Thus the following system can be considered:

$$\dot{y} = h(y), \tag{16}$$

where $y \in S_\rho$ and h is the vector field of class C^1 defined above.

We conclude this section with the following claim.

PROPOSITION 3.6. *Suppose that $y_0 \in S_\rho$ is an equilibrium point for system* (16). *Then R_{y_0} is an invariant ray for system* (8).

Proof. Since $h(y_0) = 0$, it follows by (13) that

$$T_m\pi(g(m)) = 0 \qquad \forall m \in R_{y_0}. \tag{17}$$

By the chain rule, we have that the tangent mapping of the mapping $j \circ \pi : \mathbb{R}^n \setminus \{0\} \to \mathbb{R}^n \setminus \{0\}$ equals

$$T_m(j \circ \pi) = T_{\pi(m)}j \circ T_m\pi. \tag{18}$$

We obtain from (17) and (18) that $\forall m \in R_{y_0}$

$$(T_m(j \circ \pi))(g(m)) = 0. \tag{19}$$

In local coordinates we have that $j \circ \pi(x) = (\frac{1}{\rho^{r_1}(x)} x_1, \ldots, \frac{1}{\rho^{r_n}(x)} x_n)$, and this implies that

$$g(m) = \frac{1}{\rho(x)} \left(\sum_{i=1}^{n} \frac{\partial \rho}{\partial x_i}(m) g_i(m) \right) \operatorname{diag}(r) m \tag{20}$$

$\forall m \in R_{y_0}$, and therefore R_{y_0} is an invariant ray for system (8). □

4. A class of positive homogeneous systems. We call on the concept of a cooperative vector field, which has been widely studied [5], [13].

DEFINITION 4.1. *A vector field $f(x)$, $x \in \mathbb{R}^n$, is* cooperative in $W \subset \mathbb{R}^n$ *if the Jacobian matrix $\frac{\partial f}{\partial x}$ is Metzler $\forall x \in W$.*

System (1) is called *cooperative* if the following hypothesis holds.

($\mathcal{H}2$) $f(x)$ is cooperative in $\mathbb{R}^n_+ \setminus \{0\}$.

THEOREM 4.2. *If system* (1) *satisfies* ($\mathcal{H}1$) *and* ($\mathcal{H}2$), *then $\mathbb{R}^n_+$ is a forward invariant set for system* (1).

Proof. Since $\tau \geq 0$ and $f(x)$ is cooperative in $\mathbb{R}^n_+ \setminus \{0\}$, it follows from Euler's formula that $f_i(x) \geq 0 \; \forall x \in \mathbb{R}^n_+ \setminus \{0\} : x_i = 0$. Also $f(0) = 0$ and the uniqueness of solutions for system (1) are guaranteed, as has been shown in the previous section. Then forward solutions cannot leave $\mathbb{R}^n_+$, proving Theorem 4.2. □

For future reference we apply Kamke's theorem to obtain the following.

PROPOSITION 4.3. *If system* (1) *satisfies* ($\mathcal{H}1$) *and* ($\mathcal{H}2$), *then the flow of system* (1) *is monotone in $\mathbb{R}^n_+$.*

Proof. This follows from Kamke's theorem [13, Remark 1.4, p. 34] if the following conditions hold:

1. $f(x)$ is of *type K* on $\operatorname{int}(\mathbb{R}^n_+)$, i.e., $\forall x, y \in \operatorname{int}(\mathbb{R}^n_+)$ with $x \leq y$ it holds that $f_i(x) \leq f_i(y) \; \forall i : x_i = y_i$. It is easily checked that $f(x)$ is of type K on $\operatorname{int}(\mathbb{R}^n_+)$ since ($\mathcal{H}2$) holds (see [13, Remark 1.1, p. 33]).
2. $\mathbb{R}^n_+$ is a forward invariant set for system (1). This follows from Theorem 4.2.
3. $\forall x, y \in \mathbb{R}^n_+$ with $x < y$ there exist sequences $\{x_n\}, \{y_n\} \subset \operatorname{int}(\mathbb{R}^n_+)$ such that $x_n < y_n \; \forall n$ and $x_n \to x$, $y_n \to y$ when $n \to +\infty$. It is easily checked that this condition holds.
4. $f(x)$ is continuously differentiable on some neighborhood of $\mathbb{R}^n_+$. This condition is not necessarily satisfied here. Indeed, it may happen that $\frac{\partial f}{\partial x}$ is not defined at $x = 0$. A closer look at the proof of Kamke's theorem as stated in [13] indicates that only the continuity of solutions with respect to initial conditions is needed. Since we have shown that with ($\mathcal{H}1$) the solutions of system (1) are unique, it follows that solutions of system (1) are continuous with respect to initial conditions [4, Theorem 2.1, p. 94].

This concludes the proof. □

THEOREM 4.4. *If system* (1) *satisfies* ($\mathcal{H}1$) *and* ($\mathcal{H}2$), *then there exists at least one invariant ray in $\mathbb{R}^n_+$ for system* (1).

Proof. Since, by Theorem 4.2, $\mathbb{R}^n_+$ is a forward invariant set for system (1), $\mathbb{R}^n_+$ is also a forward invariant set for system (8). It follows that the set $S_{\rho,+} := \mathbb{R}^n_+ \cap S_\rho$ is a forward invariant set for system (16). If not, there exists a forward solution $y(t, y_0)$, $y_0 \in S_{\rho,+}$ and $t \in \mathcal{I}_{y_0}$, for system (16) and a $T \in \mathcal{I}_{y_0}$ such that $y(T, y_0) \in S_\rho \setminus S_{\rho,+}$. But then the forward solution of system (8) starting in y_0 at $t = 0$, $z(t, y_0)$, is such that $z(T, y_0) \in \mathbb{R}^n \setminus \mathbb{R}^n_+$, and thus we obtain a contradiction.

Since $S_{\rho,+}$ is compact, system (16), restricted to $S_{\rho,+}$, is forward complete; i.e., forward solutions are defined $\forall t \in \mathbb{R}^+$. For each $t \in \mathbb{R}^+$ consider the mapping

$\psi_t : S_{\rho,+} \to S_{\rho,+}$ of system (16), mapping $y_0 \to \psi_t(y_0) := y(t, y_0)$. Since $S_{\rho,+}$ is compact and since it is a retract of the $(n-1)$-dimensional unit disk $D^{n-1} := \{x \in \mathbb{R}^{n-1} | \, |x| \leq 1\}$, it follows from a generalization of Brouwer's fixed point theorem (see, e.g., [2, p. 171]) that each (continuous) mapping ψ_t, $t \in \mathbb{R}^+$, has a fixed point x_t^*.

Pick an arbitrary $T^* \in \mathbb{R}_0^+$ and consider the sequence of times $\{\frac{T^*}{n}\}$ with $n \geq 1$ integer and $n \to +\infty$. Then $\forall n$ the map $\psi_{\frac{T^*}{n}}$ has at least one fixed point x_n^*. Since the elements of the sequence $\{x_n^*\}$, $n \to +\infty$, belong to the compact set $S_{\rho,+}$, there exists a convergent subsequence $\{x_{n_k}^*\}$ with $n_k \to +\infty$ when $k \to +\infty$. The limit of this subsequence is denoted as x^*. We will prove that $\psi_t(x^*) = x^* \ \forall t \in \mathbb{R}_0^+$ and thus that x^* is an equilibrium point for system (16).

Pick an arbitrary $t \in \mathbb{R}_0^+$. We can find a sequence of nonnegative integers $\{l_k\}$ and a sequence of real numbers $\{d_k\}$ with $0 \leq d_k < \frac{T^*}{n_k}$ and $d_k \to 0$ when $k \to +\infty$ and such that $t = l_k \frac{T^*}{n_k} + d_k$. It follows that

$$\begin{aligned}
\psi_t(x^*) &= \lim_{k \to +\infty} \psi_t(x_{n_k}^*) \\
&= \lim_{k \to +\infty} \psi_{d_k}(\psi_{l_k \frac{T^*}{n_k}}(x_{n_k}^*)) && \text{because } t = l_k \tfrac{T^*}{n_k} + d_k \\
&= \lim_{k \to +\infty} \psi_{d_k}(x_{n_k}^*) && \text{because } x_{n_k}^* \text{ is a fixed point of } \psi_{l_k \frac{T^*}{n_k}} \\
&= \lim_{k \to +\infty} y(d_k, x_{n_k}^*) && \text{by definition of } \psi_{d_k} \\
&= x^*.
\end{aligned}$$

In the above, the third equality is valid since if x_n^* is a fixed point of $\psi_{\frac{T^*}{n}}$, then x^* is also a fixed point of $\psi_{k\frac{T^*}{n}}$ $\forall$ nonnegative integers k; the fifth equality is valid since $y(t, x_0)$ is continuous on $\mathbb{R}^+ \times S_{\rho,+}$. Since $t \in \mathbb{R}_0^+$ was arbitrary, we have proved that $\psi_t(x^*) = x^* \ \forall t \in \mathbb{R}_0^+$, and thus x^* is an equilibrium point of system (16).

Then by Proposition 3.6, R_{x^*} is an invariant ray for system (8) and thus also for system (1). □

We introduce the following hypothesis:

($\mathcal{H}3$) For $x \in \text{int}(\mathbb{R}_+^n)$, $\frac{\partial f}{\partial x}$ is irreducible. For $x \in \text{bd}(\mathbb{R}_+^n) \setminus \{0\}$, either $\frac{\partial f}{\partial x}(x)$ is irreducible or $f_i(x) > 0 \ \forall i : x_i = 0$.

System (1) is called *irreducible* if ($\mathcal{H}3$) holds.

PROPOSITION 4.5. *If system* (1) *satisfies* ($\mathcal{H}1$), ($\mathcal{H}2$), *and* ($\mathcal{H}3$), *then the flow of system* (1) *is strongly monotone in* $\mathbb{R}_+^n$.

Proof. The flow of system (1) is monotone by Proposition 4.3.

It will now be shown that $\forall x_0, y_0 \in \mathbb{R}_+^n$ with $x_0 < y_0$ it holds that

$$x(t, x_0) \ll x(t, y_0) \qquad \forall t \in (\mathcal{I}_{x_0} \cap \mathcal{I}_{y_0}) \setminus \{0\}. \tag{21}$$

Case 1. $y_0 \in \text{int}(\mathbb{R}_+^n)$.

1. If $x_0 = 0$, then $x_0 \ll y_0$, and thus (21) follows from Proposition 4.3.
2. If $x_0 \neq 0$, then (21) follows from the generalized Kamke theorem in [13, Remark 1.1, p. 58].

Case 2. $y_0 \in \text{bd}(\mathbb{R}_+^n) \setminus \{0\}$. Notice that in this case x_0 belongs to $\text{bd}(\mathbb{R}_+^n)$ since $x_0 < y_0$. We distinguish two cases:

1. $x_0 = 0$. We distinguish two subcases:
   (a) $f_i(y_0) > 0 \ \forall i : (y_0)_i = 0$. It is clear that for t small enough and strictly positive,

   $$0 = x(t, x_0) \ll x(t, y_0). \tag{22}$$

   Then it follows by Proposition 4.3 that (22) holds $\forall t \in \mathcal{I}_{y_0} \setminus \{0\}$, which proves (21).

(b) $\frac{\partial f}{\partial x}(y_0)$ is irreducible. It follows from Proposition 4.3 that $x(t, y_0) > 0$ $\forall t \in \mathcal{I}_{y_0} \setminus \{0\}$. Suppose that (21) does not hold; then there exists some $i \in N$ such that $x_i(t, y_0) = 0$ $\forall t \in [0, t']$, where t' is some strictly positive real number.
On the other hand, since $\frac{\partial f}{\partial x}(y_0)$ is irreducible, it follows from Euler's formula that there exists some $j \in N$ such that $f_j(y_0) > 0$, implying that for small and strictly positive t, $x_j(t, y_0) > 0$.
If $j = i$, then we have reached a contradiction. If $j \neq i$, then we can pick $0 < t_1 < t'$ and consider $x(t_1, y_0) \in \mathrm{bd}(\mathbb{R}^n_+) \setminus \{0\}$.
If $\frac{\partial f}{\partial x}(x(t_1, y_0))$ is reducible, then it follows from $(\mathcal{H}3)$ that $f_k(x(t_1, y_0)) > 0$ $\forall k : x_k(t_1, y_0) = 0$ (and thus, in particular, for $k = i$), yielding a contradiction.
If $\frac{\partial f}{\partial x}(x(t_1, y_0))$ is irreducible, then there exists a $j' \in N$ with $j' \neq j$ such that $x_{j'}(t, y_0) > 0$ for small $t > t_1$. This argument is repeated and ends in a finite number of steps, leading to a contradiction.

2. $x_0 \neq 0$. It follows from Case 2.1 that both solution $x(t, x_0)$ and $y(t, y_0)$ belong to $\mathrm{int}(\mathbb{R}^n_+)$ for $t \in \mathcal{I}_{x_0} \setminus \{0\}$, respectively, $t \in \mathcal{I}_{y_0} \setminus \{0\}$.
On the other hand, it follows from Proposition 4.3 that $x(t, x_0) < y(t, y_0)$ $\forall t \in (\mathcal{I}_{x_0} \cap \mathcal{I}_{y_0}) \setminus \{0\}$. Suppose that (21) does not hold; then there exists some $l \in N$ such that $x_l(t, x_0) = x_l(t, y_0)$ $\forall t \in [0, t'']$, where t'' is some strictly positive real number. This contradicts that the flow on $\mathrm{int}(\mathbb{R}^n_+)$ is strongly monotone, which follows from the generalized Kamke theorem in [13]. ∎

It follows from Theorem 4.4 that if system (1) satisfies $(\mathcal{H}1)$ and $(\mathcal{H}2)$, there exists at least one invariant ray in $\mathbb{R}^n_+$ for system (1). Adding hypothesis $(\mathcal{H}3)$ allows us to draw more conclusions regarding the location of these invariant rays in $\mathbb{R}^n_+$ and their possible uniqueness, by means of the strong monotonicity property of the flow of system (1) as expressed in Proposition 4.5.

THEOREM 4.6. *If system* (1) *satisfies* $(\mathcal{H}1)$, $(\mathcal{H}2)$, *and* $(\mathcal{H}3)$, *then the invariant rays for system* (1) *in* $\mathbb{R}^n_+$ *belong to* $\mathrm{int}(\mathbb{R}^n_+)$.

If the order τ *of the homogeneous vector field* $f(x)$ *is equal to zero, then there exists a* unique *invariant ray for system* (1) *in* $\mathrm{int}(\mathbb{R}^n_+)$.

If the order τ *of the homogeneous vector field* $f(x)$ *is greater than zero and if there exists a stable invariant ray for system* (1) *in* $\mathrm{int}(\mathbb{R}^n_+)$, *then this invariant ray is unique in* $\mathrm{int}(\mathbb{R}^n_+)$.

If the order τ *of the homogeneous vector field* $f(x)$ *is greater than zero and if there exists an unstable invariant ray for system* (1) *in* $\mathrm{int}(\mathbb{R}^n_+)$, *then this invariant ray is not necessarily unique in* $\mathrm{int}(\mathbb{R}^n_+)$. *There may be multiple invariant rays for system* (1) *in* $\mathrm{int}(\mathbb{R}^n_+)$, *all of them unstable.*

Proof. Let $R_{x^*} \subset \mathbb{R}^n_+$ be an invariant ray for system (1); then

$$f(x^*) = \gamma_{x^*} \mathrm{diag}(r) x^* \tag{23}$$

for some $\gamma_{x^*} \in \mathbb{R}$.

First it is shown that $R_{x^*} \subset \mathrm{int}(\mathbb{R}^n_+)$. Suppose not; then $R_{x^*} \subset \mathrm{bd}(\mathbb{R}^n_+)$. According to $(\mathcal{H}3)$, two cases can be distinguished: $\frac{\partial f}{\partial x}(x^*)$ is irreducible or $f_i(x^*) > 0$ $\forall i : x^*_i = 0$.

Case 1. $\frac{\partial f}{\partial x}(x^*)$ is irreducible. From Euler's formula and (23),

$$(\mathrm{diag}(r + \tau^*))^{-1} \frac{\partial f}{\partial x}(x^*) \mathrm{diag}(r) x^* = \gamma_{x^*} \mathrm{diag}(r) x^*. \tag{24}$$

Since $r \in \operatorname{int}(\mathbb{R}^n_+)$, $\tau \in \mathbb{R}^n_+$, and $\frac{\partial f}{\partial x}(x^*)$ is Metzler and irreducible, $(\operatorname{diag}(r+\tau^*))^{-1}\frac{\partial f}{\partial x}(x^*)$ is also Metzler and irreducible. In addition, $\operatorname{diag}(r)x^* \in \operatorname{bd}(\mathbb{R}^n_+) \setminus \{0\}$. However, an irreducible Metzler matrix has no nonzero eigenvector belonging to $\operatorname{bd}(\mathbb{R}^n_+)$ [1]. Thus we obtain a contradiction.

Case 2. $f_i(x^*) > 0\ \forall i : x_i^* = 0$. Since $x^* \in \operatorname{bd}(\mathbb{R}^n_+)$ and with (23), there exists $i \in N$ such that $x_i^* = 0$ and $f_i(x^*) = 0$, yielding a contradiction.

Next it is shown that an invariant ray for system (1) is unique.

Suppose that there are two invariant rays $R_1, R_2 \subset \operatorname{int}(\mathbb{R}^n_+)$ $(R_1 \neq R_2)$ for system (1). Pick an arbitrary point $\bar{x} \in R_1$. There exist two points $p, q \in R_2$ such that $p < \bar{x} < q$ and $p_i = \bar{x}_i$, $q_j = \bar{x}_j$ for some $i \neq j$, $i, j \in N$. Indeed, pick an arbitrary $y \in R_2$. Since $R_2 \subset \operatorname{int}(\mathbb{R}^n_+)$, it follows that $y \gg 0$, and thus the following positive real numbers can be defined:

$$\lambda_1 = \min_{k \in N} \left(\left(\frac{\bar{x}_k}{y_k} \right)^{\frac{1}{r_k}} \right), \tag{25}$$

$$\lambda_2 = \max_{k \in N} \left(\left(\frac{\bar{x}_k}{y_k} \right)^{\frac{1}{r_k}} \right). \tag{26}$$

Since $R_1 \neq R_2$, it follows that $\lambda_1 \neq \lambda_2$. Then there exist $i, j \in N$ with $i \neq j$ such that $\lambda_1 = (\frac{\bar{x}_i}{y_i})^{\frac{1}{r_i}}$ and $\lambda_2 = (\frac{\bar{x}_j}{y_j})^{\frac{1}{r_j}}$. This implies that $p := \delta^r_{\lambda_1}(y)$ and $q := \delta^r_{\lambda_2}(y)$ satisfy the desired properties.

Solutions of system (1) starting on R_1, respectively on R_2, satisfy (6). Since p, $q \in R_2$, there exists a $\tilde{\lambda} \in \mathbb{R}^+_0$ such that $q = \delta^r_{\tilde{\lambda}}(p)$ and $\gamma_q = \tilde{\lambda}^\tau \gamma_p$ (see (5)). It follows from Proposition 4.5 that

$$x(t,p) \ll x(t,\bar{x}) \ll x(t,q), \tag{27}$$

where the first inequality holds $\forall t \in (\mathcal{I}_{\bar{x}} \cap \mathcal{I}_p) \setminus \{0\}$, and the second inequality $\forall t \in (\mathcal{I}_{\bar{x}} \cap \mathcal{I}_q) \setminus \{0\}$. We obtain from (6), (27) and since $p_i = \bar{x}_i$ and $\bar{x}_j = q_j$ that

$$\left(\gamma_p r_i p_i^{-\frac{\tau}{r_i}} \right) p_i^{1+\frac{\tau}{r_i}} < \left(\gamma_{\bar{x}} r_i \bar{x}_i^{-\frac{\tau}{r_i}} \right) \bar{x}_i^{1+\frac{\tau}{r_i}}, \tag{28}$$

$$\left(\gamma_{\bar{x}} r_j \bar{x}_j^{-\frac{\tau}{r_j}} \right) \bar{x}_j^{1+\frac{\tau}{r_j}} < \left(\gamma_q r_j q_j^{-\frac{\tau}{r_j}} \right) q_j^{1+\frac{\tau}{r_j}}, \tag{29}$$

or

$$\gamma_p < \gamma_{\bar{x}} < \gamma_q. \tag{30}$$

Two cases can be distinguished: $\tau = 0$ and $\tau > 0$.

Case 1. $\tau = 0$. If $\tau = 0$, then $\gamma_p = \tilde{\lambda}^0 \gamma_q = \gamma_q$, contradicting (30).

Case 2. $\tau > 0$. We introduce the sign function $\operatorname{sign} : \mathbb{R} \to \mathbb{R}$ as

$$\operatorname{sign}(x) = \begin{cases} -1 & \text{for } x < 0, \\ 0 & \text{for } x = 0, \\ +1 & \text{for } x > 0. \end{cases} \tag{31}$$

From $\gamma_q = \tilde{\lambda}^\tau \gamma_p$ it follows that $\operatorname{sign}(\gamma_p) = \operatorname{sign}(\gamma_q)$.

Case 2(a). $\operatorname{sign}(\gamma_{\bar{x}}) \neq \operatorname{sign}(\gamma_p)$. Since $\operatorname{sign}(\gamma_p) = \operatorname{sign}(\gamma_q)$, it follows from (30) that $\operatorname{sign}(\gamma_{\bar{x}}) = \operatorname{sign}(\gamma_p) = \operatorname{sign}(\gamma_q)$, which is impossible because $\operatorname{sign}(\gamma_{\bar{x}}) \neq \operatorname{sign}(\gamma_p)$.

Case 2(b). $\operatorname{sign}(\gamma_{\bar{x}}) = \operatorname{sign}(\gamma_p)$. Two cases can be distinguished: $\operatorname{sign}(\gamma_{\bar{x}}) = 0$ and $\operatorname{sign}(\gamma_{\bar{x}}) = -1$.

1. $\text{sign}(\gamma_{\bar{x}}) = \text{sign}(\gamma_p) = 0$. This is impossible since, from (30), $\gamma_p < \gamma_{\bar{x}}$.
2. $\text{sign}(\gamma_{\bar{x}}) = \text{sign}(\gamma_p) = -1$. From (30),

$$\gamma_p < \gamma_q. \tag{32}$$

On the other hand, since $p < \bar{x} < q$ and $q = \delta^r_{\tilde{\lambda}}(p)$, it follows that $\tilde{\lambda} > 1$. Furthermore, $\gamma_q = \tilde{\lambda}^\tau \gamma_p$. But since $\text{sign}(\gamma_p) = \text{sign}(\gamma_q) = -1$, $\tau > 0$, and $\tilde{\lambda} > 1$, it follows that $\gamma_q \leq \gamma_p$, contradicting (32). ∎

Notice that Theorem 4.6 does not exclude the possibility of several invariant rays in $\text{int}(\mathbb{R}^n_+)$ for system (1). This can happen only if $\tau > 0$ and if all invariant rays are unstable. This situation can indeed occur; in particular, we will give an example of a planar cooperative irreducible and homogeneous system of order $\tau = 1$ possessing infinitely many unstable invariant rays in $\text{int}(\mathbb{R}^n_+)$.

Example. Consider the following system:

$$\dot{x} = f_1(x) := (x_1 + x_2)x, \tag{33}$$

where $x := (x_1, x_2)^T \in \mathbb{R}^2$. This system is homogeneous of order $\tau = 1$ with respect to the standard dilation map. In addition, $f_1(x)$ is cooperative in $\mathbb{R}^2_+$, and $\frac{\partial f_1}{\partial x}$ is irreducible $\forall x \in \text{int}(\mathbb{R}^2_+)$. Notice that $(\mathcal{H}3)$ is not satisfied.

Next consider the following system:

$$\dot{x} = f_2(x) := \begin{pmatrix} x_1^2 + x_1x_2 + x_2^2 \\ x_1^2 + x_1x_2 + x_2^2 \end{pmatrix}. \tag{34}$$

This system is also homogeneous of order $\tau = 1$ with respect to the standard dilation map. In addition, $f_2(x)$ is cooperative in $\mathbb{R}^2_+$, and $\frac{\partial f_2}{\partial x}$ is irreducible $\forall x \in \mathbb{R}^2_+ \setminus \{0\}$. In particular, $\frac{\partial f_2}{\partial x}$ is irreducible when $x \in \text{bd}(\mathbb{R}^2_+)\setminus\{0\}$.

Based on systems (33) and (34), we would like to construct a system with infinitely many unstable invariant rays in $\text{int}(\mathbb{R}^2_+)$. Before doing so, partition $\mathbb{R}^2_+$ into five conic parts:

$$\mathbb{R}^2_+ := \bigcup_{i=1}^{5} C_i, \tag{35}$$

where $C_1 := \{x \in \mathbb{R}^2_+ | x_2 - \frac{1}{2}x_1 \geq 0, x_1 - \frac{1}{2}x_2 \geq 0\}$, $C_2 := \{x \in \mathbb{R}^2_+ | x_2 - \frac{1}{4}x_1 \leq 0\}$, $C_3 := \{x \in \mathbb{R}^2_+ | x_2 - \frac{1}{2}x_1 < 0, x_2 - \frac{1}{4}x_1 > 0\}$, $C_4 := \{x \in \mathbb{R}^2_+ | x_1 - \frac{1}{4}x_2 \leq 0\}$, and $C_5 := \{x \in \mathbb{R}^2_+ | x_1 - \frac{1}{4}x_2 > 0, x_1 - \frac{1}{2}x_2 < 0\}$.

Now define the system

$$\dot{x} = f(x) := \begin{cases} f_1(x) & \text{for } x \in C_1, \\ f_2(x) & \text{for } x \in C_2, C_4, \\ f_1(x) + g_1(\theta)p(x) & \text{for } x \in C_3, \\ f_1(x) + g_2(\theta)p(x) & \text{for } x \in C_5, \end{cases} \tag{36}$$

where $\theta := \frac{x_2}{x_1}$, $p(x) := (x_2^2, x_1^2)^T$, and $g_1(\theta)$ ($g_2(\theta)$) is a continuously differentiable function defined on $[\theta_2, \theta_1] := [\frac{1}{4}, \frac{1}{2}]$ ($[\theta_3, \theta_4] := [2, 4]$) to be specified hereafter. The aim is to construct $g_1(\theta)$ and $g_2(\theta)$ such that $f(x)$ is continuously differentiable on $\mathbb{R}^2_+$, homogeneous of order $\tau = 1$ with respect to the standard dilation map, cooperative

on $\mathbb{R}^2_+$, and such that $\frac{\partial f}{\partial x}$ is irreducible when $x \in \mathbb{R}^2_+ \setminus \{0\}$. First $g_1(\theta)$ is constructed. Notice that $f_1(x) + g_1(\theta)p(x)$ is homogeneous of order $\tau = 1$ with respect to the standard dilation map since $p(x)$ is homogeneous of order $\tau = 1$ with respect to the standard dilation map and since g_1 is constant along any ray in C_3. For $f(x)$ to be continuous on $\cup_{i=1}^3 C_i$, it suffices that the following hold: $g_1(\theta)$ is continuous on $[\theta_2, \theta_1]$, $g_1(\theta_1) = 0$, and $g_1(\theta_2) = 1$. Consider the Jacobian of $f_1(x) + g_1(\theta)p(x)$ on C_3:

$$\frac{\partial f_1}{\partial x} + g_1(\theta) \begin{pmatrix} 0 & 2x_2 \\ 2x_1 & 0 \end{pmatrix} + \frac{\partial g_1}{\partial \theta} \begin{pmatrix} -\frac{x_2^3}{x_1^2} & \frac{x_2^2}{x_1} \\ -x_2 & x_1 \end{pmatrix}. \tag{37}$$

For $f(x)$ to be continuously differentiable on $\cup_{i=1}^3 C_i$, it suffices that the following hold: $g_1(\theta)$ is continuously differentiable on $[\theta_2, \theta_1]$ and $\frac{\partial g_1}{\partial \theta}(\theta_1) = \frac{\partial g_1}{\partial \theta}(\theta_2) = 0$. Finally, $f(x)$ is cooperative in $\cup_{i=1}^3 C_i$ and irreducible in $(\cup_{i=1}^3 C_i) \setminus \{0\}$ if and only if the off-diagonal elements of the Jacobian (37) are strictly positive when $x \in C_3$. This is the case when the following two conditions are satisfied:

I. $x_1(1 + \theta(2g_1(\theta) + \theta\frac{\partial g_1}{\partial \theta})) > 0$ when $\theta \in [\theta_2, \theta_1]$ and $x \in C_3$.

II. $x_2 + 2g_1(\theta)x_1 - \frac{\partial g_1}{\partial \theta}x_2 > 0$ when $\theta \in [\theta_2, \theta_1]$ and $x \in C_3$.

Condition II is satisfied if $g_1(\theta)$ is chosen such that $g_1(\theta) \geq 0$ and $\frac{\partial g_1}{\partial \theta} \leq 0$ when $\theta \in [\theta_2, \theta_1]$. Summarizing, we are looking for $g_1(\theta) : [\theta_2, \theta_1] \to \mathbb{R}$ such that the following conditions are satisfied:

($\mathcal{C}1$) $g_1(\theta)$ is continuously differentiable in $[\theta_2, \theta_1]$.

($\mathcal{C}2$) $g_1(\theta) \geq 0$ in $[\theta_2, \theta_1]$ and $g_1(\theta_1) = 0$, $g_1(\theta_2) = 1$.

($\mathcal{C}3$) $\frac{\partial g_1}{\partial \theta} \leq 0$ in $[\theta_2, \theta_1]$ and $\frac{\partial g_1}{\partial \theta}(\theta_1) = \frac{\partial g_1}{\partial \theta}(\theta_2) = 0$.

($\mathcal{C}4$) $x_1(1 + \theta(2g_1(\theta) + \theta\frac{\partial g_1}{\partial \theta})) > 0$ when $\theta \in [\theta_2, \theta_1]$ and $x \in C_3$.

From this we propose that $g_1(\theta)$ is a third order polynomial in θ:

$$g_1(\theta) = a\left(\frac{\theta^3}{3} - \frac{\theta_1 + \theta_2}{2}\theta^2 + \theta_1\theta_2\theta + c\right), \tag{38}$$

where a and c are real numbers that we will determine hereafter. It is clear that ($\mathcal{C}1$) and ($\mathcal{C}3$) are satisfied if $a > 0$. From ($\mathcal{C}2$) we find that $c = -\frac{1}{96}$ and $a = 384$, fixing the function $g_1(\theta)$. We have to check whether ($\mathcal{C}4$) holds. This amounts to the following question:

$$h(\theta) := 640\theta^4 - 576\theta^3 + 144\theta^2 - 8\theta + 1 > 0 \qquad \forall\theta \in [\theta_2, \theta_1]? \tag{39}$$

It is easily verified by means of Cardano's formula for finding the roots of a third order polynomial that $\frac{\partial h}{\partial \theta} = 2560(\theta - \theta')(\theta - \theta'')(\theta - \theta^*)$, where $\theta', \theta'' < \theta_2$ and $\theta^* \in (\theta_2, \theta_1)$. In addition, $\frac{\partial^2 h}{\partial \theta^2}(\theta^*) > 0$, and thus $h(\theta)$ reaches a global minimum in $[\theta_2, \theta_1]$. Finally, $h(\theta^*) > 0$, and this implies that $h(\theta) > 0$ $\forall\theta \in [\theta_2, \theta_1]$.

We are left with finding an appropriate $g_2(\theta) : [\theta_3, \theta_4] \to \mathbb{R}$. Because of the symmetry of both $f_1(x)$ and $f_2(x)$ and because $\frac{1}{\theta_3} = \theta_1$ and $\frac{1}{\theta_4} = \theta_2$, we can set $g_2(\theta) := g_1(\frac{1}{\theta})$ $\forall\theta \in [\theta_3, \theta_4]$.

In conclusion, system (36) is homogeneous of order $\tau = 1$ with respect to the dilation map, $f(x)$ is cooperative in $\mathbb{R}^2_+$, and $\frac{\partial f}{\partial x}$ is irreducible for $x \in \mathbb{R}^2_+ \setminus \{0\}$. There are an infinite number of unstable invariant rays for system (36) in $\text{int}(\mathbb{R}^2_+)$. Indeed, *every* ray in C_1 is invariant and unstable for system (36). □

5. Main result. Suppose that system (1) satisfies ($\mathcal{H}1$), ($\mathcal{H}2$), and ($\mathcal{H}3$), and assume that initial conditions for system (1) belong to $\mathbb{R}^n_+$. From Theorem 4.6 it

follows that there exists at least one invariant ray in $\mathbb{R}^n_+$ for system (1) and that this invariant ray belongs to $\text{int}(\mathbb{R}^n_+)$.

If $\tau = 0$, then there exists a unique invariant ray in $\text{int}(\mathbb{R}^n_+)$.

If $\tau > 0$ and if there exists a stable invariant ray in $\text{int}(\mathbb{R}^n_+)$, then this invariant ray is unique.

Consider the flow $\phi_t : \mathbb{R}^n_+ \to \mathbb{R}^n$ of system (1) mapping x_0 to $\phi_t(x_0) := x(t, x_0)$. $\forall x_0 \in \mathbb{R}^n_+$, $\phi_t(x_0)$ exists when $t \in \mathcal{I}_{x_0}$. In the following lemma, we provide sufficient conditions guaranteeing that ϕ_t is defined $\forall t \in \mathbb{R}^+$ when restricting initial conditions to $\mathbb{R}^n_+$.

LEMMA 5.1. *If system* (1) *satisfies* ($\mathcal{H}1$) *and if* $\tau = 0$, *then* $\phi_t : \mathbb{R}^n_+ \to \mathbb{R}^n_+$ *is defined* $\forall t \in \mathbb{R}^+$.

If system (1) *satisfies* ($\mathcal{H}1$), ($\mathcal{H}2$), *and* ($\mathcal{H}3$); *if* $\tau > 0$; *and if there exists a stable (and thus unique) invariant ray for system* (1) *in* $\text{int}(\mathbb{R}^n_+)$, *then* $\phi_t : \mathbb{R}^n_+ \to \mathbb{R}^n_+$ *is defined for all* $t \in \mathbb{R}^+$.

Proof. Case 1. $\tau = 0$. First it will be shown that every forward solution of system (1) remains in a compact set in finite time intervals. Therefore we consider the dynamics of a homogeneous norm:

$$\begin{aligned} \dot{\rho} &= \left(\frac{1}{\rho} \frac{\partial \rho}{\partial x} f \right) \rho \\ &= k(x)\rho. \end{aligned} \tag{40}$$

It is easily verified that

$$k(\delta^r_\lambda(x)) = k(x) \tag{41}$$

$\forall x \in \mathbb{R}^n \setminus \{0\}$ and $\lambda \in \mathbb{R}^+_0$. The function $k(x)$ takes a maximal value M on the compact set $\{z \in \mathbb{R}^n | \rho(z) = 1\}$. In fact, by (41), M is the maximum of $k(x)$ in $\mathbb{R}^n \setminus \{0\}$. Then $\dot{\rho} \leq M\rho$, and thus

$$\rho(x(t, x_0)) \leq e^{Mt}\rho(x_0). \tag{42}$$

This implies that $x(t, x_0)$, $t \in \mathcal{I}_{x_0}$, belongs to the compact set $K := \{z \in \mathbb{R}^n | \rho(z) \leq e^{MT_{\max}(x_0)\rho(x_0)}\}$.

Now suppose that $\mathcal{I}_{x_0} = [0, T_{\max}(x_0))$, with $T_{\max}(x_0) < +\infty$. Pick a sequence $\{x_{t_n}\} \subset x(t, x_0)$ with $t_n \to T_{\max}(x_0)$ when $n \to +\infty$. Since $|f(x)|$ is continuous, it attains a maximum M' on K. Then $\forall t_n$ and t_m

$$\begin{aligned} |x_{t_n} - x_{t_m}| &\leq \int_{t_m}^{t_n} |f(x(t, x_0))| dt \\ &\leq M' |t_n - t_m|. \end{aligned}$$

This implies that $|x_{t_n} - x_{t_m}| \to 0$ when n and $m \to +\infty$. Therefore $\{x_{t_n}\}$ is a Cauchy sequence and thus converges.

Every sequence on $x(t, x_0)$ converges to the same point $p \in K$. Indeed, if this were not the case, then there would exist two Cauchy sequences on $x(t, x_0)$, converging to different points p_1 and $p_2 \in K$. Then there would exist a sequence on $x(t, x_0)$ with *two* limit points p_1 and p_2. This is impossible since, as we have shown, every sequence on $x(t, x_0)$ is a Cauchy sequence and therefore converges.

Since every sequence on $x(t, x_0)$ converges to p, it follows that $\lim_{t \to T_{\max}(x_0)} x(t, x_0) = p$. The solution $x(t, x_0)$ can then be extended by concatenating it with the solution starting in p. This implies that the maximal forward interval of existence of the

solution starting at $t = 0$ in x_0 contains $\mathcal{I}_{x_0}$ as a proper subset, contradicting the assumption that $\mathcal{I}_{x_0}$ is the maximal forward interval of existence. Thus $\forall x_0 \in \mathbb{R}^n_+$, $T_{\max}(x_0) = +\infty$, implying that ϕ_t is defined $\forall t \in \mathbb{R}^+$.

Case 2. $\tau > 0$. Let R_{x^*} be the unique stable invariant ray for system (1) in $\operatorname{int}(\mathbb{R}^n_+)$. For each $x_0 \in \mathbb{R}^n_+$ we can find $y_0 \in R_{x^*}$ such that $x_0 < y_0$. Since $x(t, y_0)$ satisfies (6) with $z = x^*$ and $\gamma_{x^*} \le 0$, we obtain that $\mathcal{I}_{y_0} = [0, +\infty)$. This implies that $x(t, x_0)$ belongs to the compact hypercube $C := \{z \in \mathbb{R}^n_+ | 0 \le z \le x(T_{\max}(x_0), y_0)\}$.

The rest of the proof follows the same lines as the proof of Case 1. The role of the compact set K is now played by the compact set C. □

We are ready to state the main theorem.

THEOREM 5.2 (Main Theorem). *Assume that system* (1) *satisfies* ($\mathcal{H}1$), ($\mathcal{H}2$), *and* ($\mathcal{H}3$), *and assume that initial conditions for system* (1) *belong to* $\mathbb{R}^n_+$. *Then there exists at least one invariant ray* $R_{x^*} \subset \operatorname{int}(\mathbb{R}^n_+)$.

If $\tau = 0$, *then* R_{x^*} *is unique in* $\operatorname{int}(\mathbb{R}^n_+)$. *If* $\tau > 0$ *and if* R_{x^*} *is stable, then* R_{x^*} *is unique. If* $\tau > 0$ *and if* R_{x^*} *is unstable, then* R_{x^*} *is not necessarily unique. If there are several invariant rays, all of them are unstable.*

The zero solution of system (1) *is*

- *unstable if and only if* R_{x^*} *is unstable,*
- *stable if and only if* R_{x^*} *is stable,*
- *globally asymptotically stable if and only if* R_{x^*} *is asymptotically stable.*

Proof. Sufficiency.

1. R_{x^*} is unstable. $\forall x_0 \in R_{x^*}$, the forward solution of system (1) starting at $t = 0$ in x_0 satisfies (6) with $z = x^*$. Since R_{x^*} is unstable, $\gamma_{x^*} > 0$, and thus $x(t, x_0)$ diverges when $t \to T_{\max}(x_0)$. Then the zero solution of system (1) is unstable.
2. R_{x^*} is stable. $\forall x_0 \in R_{x^*}$, the forward solution of system (1) starting at $t = 0$ in x_0 satisfies (6) with $z = x^*$. Since R_{x^*} is stable, $\gamma_{x^*} \le 0$, and thus $x(t, x_0) \le x_0 \ \forall t \in \mathbb{R}^+$.
   It follows that $\forall x_0 \in R_{x^*}$ the hypercube $C_{x_0} := \{y_0 \in \mathbb{R}^n_+ | 0 \le y_0 \le x_0\}$ is a forward invariant set for system (1). Indeed, if $y_0 = x_0$, then $x(t, y_0) \in C_{x_0}$ $\forall t \in \mathbb{R}^+$, since $x(t, x_0) \le x_0 \ \forall t \in \mathbb{R}^+$. If $y_0 < x_0$ with $y_0 \in C_{x_0}$, then from Proposition 4.5 and Lemma 5.1, $x(t, y_0) \ll x(t, x_0) \le x_0 \ \forall t \in \mathbb{R}^+$, and thus $x(t, x_0) \in C_{x_0} \ \forall t \in \mathbb{R}^+$.
   Since x_0 can be chosen arbitrarily close to the origin, the zero solution of system (1) is stable.
3. R_{x^*} is asymptotically stable. $\forall x_0 \in R_{x^*}$, the forward solution of system (1) starting at $t = 0$ in x_0 satisfies (6) with $z = x^*$. Since R_{x^*} is asymptotically stable, $\gamma_{x^*} < 0$, and thus $\lim_{t \to +\infty} x(t, x_0) = 0$. Also all forward solutions starting outside R_{x^*} converge to the origin. Indeed, $\forall y_0 \notin R_{x^*}$ there exists $p \in R_{x^*}$ such that $y_0 < p$ since $R_{x^*} \subset \operatorname{int}(\mathbb{R}^n_+)$. Then it follows from Proposition 4.5 and Lemma 5.1 that $x(t, y_0) \ll x(t, p) \ \forall t \in \mathbb{R}^+$. From $\lim_{t \to +\infty} x(t, p) = 0$ it follows that $\lim_{t \to +\infty} x(t, y_0) = 0$.
   It has been shown that all forward solutions converge to the origin. Stability of the zero solution follows from the proof of item 2 above. Thus the zero solution of system (1) is globally asymptotically stable.

Necessity.

1. System (1) is unstable. This follows from the contrapositive statement of the statement proved in item 2 of the sufficiency part of this theorem.
2. System (1) is stable. This follows from the contrapositive statement of the statement proved in item 1 of the sufficiency part of this theorem.

3. System (1) is asymptotically stable. Suppose that R_{x^*} is not asymptotically stable. Then $\gamma_{x^*} \geq 0$, implying that no forward solution starting on R_{x^*} converges to the origin and contradicting the assumption that system (1) is asymptotically stable. □

Example. Consider a reversible chemical reaction at a given, constant temperature:

$$A \leftrightarrows B, \tag{43}$$

where A and B are chemical components. Denote the concentrations of A and B, respectively, by x_1 and x_2. We assume that this reaction takes place in a *closed* chemical reactor and thus there is no exchange of material with the environment. The rate constant of reaction $A \to B$ is denoted by $k_1 \in \mathbb{R}^+$, and the rate constant of reaction $B \to A$ by $k_2 \in \mathbb{R}_+$. Supposing that the dynamics of both reactions is dictated by the *mass action principle* [11] and that both reactions are of second order, we obtain that the concentrations satisfy the following differential equations:

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} -k_1 x_1^2 + k_2 x_2^2 \\ k_1 x_1^2 - k_2 x_2^2 \end{pmatrix}. \tag{44}$$

System (44) is homogeneous of order $\tau = 1$ with respect to the standard dilation map, cooperative in $\mathbb{R}^2_+$, and $(\mathcal{H}3)$ holds. Therefore Theorem 5.2 can be applied. It is easily verified that R_{x^*} with $x^* = (\sqrt{k}_2 \ \sqrt{k}_1)$ is the unique invariant ray in $\mathbb{R}^2_+$ that belongs to $\text{int}(\mathbb{R}^2_+)$ and that this ray is stable. We conclude that the zero solution of system (44) is also stable but not asymptotically stable.

One of the basic assumptions in model (44) is that the chemical reactor is closed, which is usually not satisfied. Indeed, in most models of chemical reactors there is exchange of chemicals with the environment, and a more realistic model would be (see [14])

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} -k_1 x_1^2 + k_2 x_2^2 + (p_1(x_1,\, x_2) - q_1(x_1,\, x_2)) \\ k_1 x_1^2 - k_2 x_2^2 + (p_2(x_1,\, x_2) - q_2(x_1,\, x_2)) \end{pmatrix}, \tag{45}$$

where the functions $p_i \geq 0$, respectively $q_i \geq 0$, model the inflow, respectively outflow, of the chemicals. We show in [8] that the stability behavior of the trivial solution of the (simpler) system (44) plays an important role in determining the behavior of system (45) and that this can be extended to more general chemical reactors (in particular, to reactors containing more than two chemicals, in which several reactions take place).

6. Asymptotic behavior. Assume that system (1) satisfies $(\mathcal{H}1)$, $(\mathcal{H}2)$, and $(\mathcal{H}3)$, that the initial conditions of (1) belong to $\mathbb{R}^n_+$, and that $\tau = 0$. We recall from Theorem 4.6 that system (1) possesses a unique invariant ray R_{x^*} in $\mathbb{R}^n_+$, which belongs to $\text{int}(\mathbb{R}^n_+)$, such that

$$f(x^*) = \gamma \text{diag}(r) x^* \tag{46}$$

for some $\gamma \in \mathbb{R}$. The sign of γ then determines the stability properties of the zero solution of system (1). In this section the limiting behavior of solutions of systems satisfying these conditions will be described in more detail.

Introduce the following variable:

$$z(t, x_0) := (\text{diag}(e^{\gamma r t}))^{-1} \phi_t(x_0), \tag{47}$$

where $e^{\gamma r t} := (e^{\gamma r_1 t}, e^{\gamma r_2 t}, \dots, e^{\gamma r_n t})$.

Since $f(x)$ is homogeneous of order $\tau = 0$ with respect to $\delta_\lambda^r(x)$, it is easily verified that $z(t, x_0)$ satisfies

$$\dot{z} = -\gamma \mathrm{diag}(r) z + f(z), \tag{48}$$

where $z \in \mathbb{R}_+^n$ and $z(0, x_0) = x_0$.

System (48) satisfies $(\mathcal{H}1)$ with $\tau = 0$, $(\mathcal{H}2)$, and $(\mathcal{H}3)$, and therefore Theorem 4.6 can be applied to system (48). In particular, there exists a unique invariant ray for system (48) in $\mathrm{int}(\mathbb{R}_+^n)$. It is easy to see that this invariant ray is R_{x^*}, the unique invariant ray of system (1). For system (48) this ray consists of equilibrium points. It follows from Theorem 4.2 that system (48) is positive. Restricting initial conditions for system (48) to $\mathbb{R}_+^n$, it is possible to define the flow $\Phi_t : \mathbb{R}_+^n \to \mathbb{R}^n$ of system (48) (which is defined $\forall t \in \mathbb{R}^+$ by Lemma 5.1), mapping $\mathbb{R}_+^n$ *into* $\mathbb{R}_+^n$.

We recall the following results (see [13, Theorem 2.3, p. 5, and Theorem 3.7, p. 8]).

LEMMA 6.1. *If the forward flow of a system is strongly monotone, then a limit set cannot contain two limit points x and y with $x < y$.*

LEMMA 6.2. *If the forward flow of a system is strongly monotone, if $z_1 < z_2$, and if $\omega(z_1)$ and $\omega(z_2)$ are nonempty, then $\omega(z_1) \le \omega(z_2)$.*

It follows from Proposition 4.5 that the forward flow of system (48) is strongly monotone since system (48) satisfies $(\mathcal{H}1)$ with $\tau = 0$, $(\mathcal{H}2)$, and $(\mathcal{H}3)$. This implies that Lemmas 6.1 and 6.2 apply to system (48).

THEOREM 6.3. *If system* (1) *satisfies* $(\mathcal{H}1)$ *with* $\tau = 0$, $(\mathcal{H}2)$, *and* $(\mathcal{H}3)$, *then* $\forall x_0 \in \mathbb{R}_+^n$ *there exists a* $p_{x_0} \in R_{x^*} \cup \{0\}$ *such that* $\lim_{t \to +\infty} (\mathrm{diag}(e^{\gamma r t}))^{-1} \phi_t(x_0) = p_{x_0}$.

Proof. The forward solutions of system (48) are bounded. Indeed, for each $x_0 \in \mathbb{R}_+^n$ there exists a $y \in R_{x^*}$ such that $x_0 < y$. From Proposition 4.5 it follows that $z(t, x_0) \ll z(t, y) \equiv y \ \forall t \in \mathbb{R}^+$, since y is an equilibrium point of system (48). Thus all forward solutions of system (48) are bounded, implying that the omega limit set of every forward solution is nonempty.

The proof of the theorem proceeds in two steps:

1. The omega limit set of every forward solution of system (48) is a subset of $R_{x^*} \cup \{0\}$.
2. The omega limit set of every forward solution of system (48) consists of a single equilibrium point.

First a proof of item 1 is given. Suppose that there exists a $x_0 \in \mathbb{R}_+^n$ such that $\omega(x_0) \not\subset R_{x^*} \cup \{0\}$. Then there exist at least two points p and $q \in \omega(x_0)$ with $p \neq q$ and $p \notin R_{x^*} \cup \{0\}$. Indeed, the existence of $p \in \omega(x_0)$ with $p \notin R_{x^*} \cup \{0\}$ follows immediately from the assumption that $\omega(x_0) \not\subset R_{x^*} \cup \{0\}$. Now suppose that p is the only element in $\omega(x_0)$. Then p is an equilibrium point of system (48) and thus belongs to $R_{x^*} \cup \{0\}$, since omega limit sets are (forward) invariant sets. This contradicts the fact that $p \notin R_{x^*} \cup \{0\}$. Therefore there exists a second element $q \in \omega(x_0)$ with $p \neq q$.

Three cases can occur: $p \le q$, $q \le p$, or p and q are not related by $\le$.

Case 1. $p \le q$. Now $\delta_\lambda^r(x_0) < x_0 \ \forall \lambda \in (0, 1)$, and using Lemmas 3.2 and 6.2, this implies in particular that $\forall \lambda \in (0, 1)$

$$\delta_\lambda^r(q) \le p. \tag{49}$$

On the other hand, it follows from $p \le q$ and $p \neq q$ that $p < q$. This means that there exist two subsets J and K of N, where J is a subset of N and K can be empty such

that $N = J \cup K$ and

$$\begin{aligned} p_j &< q_j \qquad \forall j \in J, \\ p_k &= q_k \qquad \forall k \in K. \end{aligned}$$

This implies that there exists a $\lambda^* \in (0,1)$ close to 1 such that

$$\begin{aligned} p_j &< (\delta^r_{\lambda^*}(q))_j \qquad \forall j \in J, \\ (\delta^r_{\lambda^*}(q))_k &\leq p_k \qquad \forall k \in K. \end{aligned}$$

This implies that $p < \delta^r_{\lambda^*}(q)$ or that p and $\delta^r_{\lambda^*}(q)$ are not related by $\leq$, contradicting (49).

Case 2. $q \leq p$. If $q < p$, then a contradiction is obtained using an argument similar to that of Case 1.

Case 3. p and q are not related by $\leq$. We have $\delta^r_\lambda(x_0) < x_0$ $\forall \lambda \in (0,1)$, and using Lemmas 3.2 and 6.2, this implies in particular that $\forall \lambda \in (0,1)$

$$\delta^r_\lambda(p) \leq q. \tag{50}$$

On the other hand, since p and q, $p \neq q$, are not related by $\leq$, there exist $i, j \in N$ with $i \neq j$ such that

$$\begin{aligned} p_i &< q_i, \\ p_j &> q_j. \end{aligned}$$

This implies that there exists a $\tilde{\lambda} \in (0,1)$ close to 1 such that

$$\begin{aligned} (\delta^r_{\tilde{\lambda}}(p))_i &< q_i, \\ (\delta^r_{\tilde{\lambda}}(p))_j &> q_j. \end{aligned}$$

This implies that $\delta^r_{\tilde{\lambda}}(p)$ and q are not related by $\leq$, contradicting (50). This concludes the proof of item 1.

Next a proof of item 2 is given. Suppose that there exists a $x_0 \in \mathbb{R}^n_+$ such that $\omega(x_0) \subset R_{x^*} \cup \{0\}$ contains two points p and q with $p \neq q$. Since both p and q belong to $R_{x^*} \cup \{0\}$ and since $R_{x^*} \subset \operatorname{int}(\mathbb{R}^n_+)$, we may assume that $p \ll q$. However, it follows from Lemma 6.1 that p and q cannot be related by $<$. Thus a contradiction is obtained, and this proves item 2. □

7. Discussion of the results. In this paper a particular class of *positive* homogeneous systems has been introduced for which the stability behavior with respect to initial conditions in $\mathbb{R}^n_+$ can be characterized by means of a simple criterion expressed in Theorem 5.2. This contrasts with the case of homogeneous systems on $\mathbb{R}^n$, where in general no criteria for (asymptotic) stability are available.

In the following we will review the Perron–Frobenius theorem. Although this theorem normally refers to *discrete*-time systems, we consider its linear continuous-time version.

Consider the linear system

$$\dot{x} = Ax, \tag{51}$$

where A is an irreducible Metzler matrix and $x \in \mathbb{R}^n$. This system is cooperative and irreducible in $\mathbb{R}^n_+$ (in fact, also in $\mathbb{R}^n$) and homogeneous of order $\tau = 0$ with respect

to the standard dilation map. Since $(Ax)_i \geq 0$ when $x_i = 0$, $\mathbb{R}^n_+$ is a forward invariant set for system (51). Thus system (51) is a positive system.

For this class of systems the Perron–Frobenius theorem states that there exists a unique eigenvector z in $\mathbb{R}^n_+$ (up to multiplication with positive scalars) and such that $z \in \text{int}(\mathbb{R}^n_+)$. Also, the eigenvalue γ associated with z is real and simple and has the property that if γ' is an eigenvalue of A and $\gamma' \neq \gamma$, then $\text{Re}(\gamma') < \gamma$, where $\text{Re}(\gamma')$ stands for the real part of γ'. The sign of γ then determines the stability behavior of the zero solution of system (51): It is unstable if $\gamma > 0$, stable if $\gamma \leq 0$, and GAS if $\gamma < 0$. Furthermore, $\forall x_0 \in \mathbb{R}^n_+$ there exists a c_{x_0} such that $\lim_{t\to+\infty} e^{At}x_0/e^{\gamma t} = c_{x_0}z$.

Theorems 4.6 and 5.2 generalize the Perron–Frobenius theorem for *linear* cooperative and irreducible systems to the class of *homogeneous* cooperative and irreducible systems and therefore to a nonlinear context. We distinguish two cases.

Case 1. $\tau = 0$. If the order of the homogeneous vector field equals zero, then according to Theorem 4.6 there exists a unique invariant ray in $\mathbb{R}^n_+$, and it belongs to $\text{int}(\mathbb{R}^n_+)$. This ray plays the role of the unique eigenvector associated to the dominating eigenvalue featured in the Perron–Frobenius theorem for linear cooperative and irreducible systems.

According to Theorem 5.2, the stability behavior of the zero solution with respect to the initial conditions in $\mathbb{R}^n_+$ is completely determined by the flow on the unique invariant ray. This is reminiscent of the Perron–Frobenius theorem for linear cooperative and irreducible systems, where the sign of the eigenvalue associated to the unique eigenvector determines the stability behavior of the system. The only difference is that the stability behavior in the linear case holds with respect to initial conditions in $\mathbb{R}^n$ and not just in $\mathbb{R}^n_+$.

It follows from Theorem 6.3 that the properties of the asymptotic behavior of solutions of homogeneous order zero, systems which are cooperative and irreducible, are similar to those of solutions of system (51).

Therefore the conclusions of the Perron–Frobenius theorem for linear cooperative and irreducible systems remain valid for the class of homogeneous *order zero* cooperative and irreducible systems, provided one restricts initial conditions to $\mathbb{R}^n_+$.

Case 2. $\tau > 0$. If the order of the homogeneous vector field is strictly positive, then according to Theorem 4.4 there is at least one invariant ray in $\mathbb{R}^n_+$. It follows from Theorem 4.6 that every invariant ray in $\mathbb{R}^n_+$ belongs to $\text{int}(\mathbb{R}^n_+)$.

1. An invariant ray is unique in $\mathbb{R}^n_+$ if it is stable or asymptotically stable. The stability behavior of the zero solution with respect to the initial conditions in $\mathbb{R}^n_+$ is completely determined by the flow on this unique invariant ray. Therefore the conclusions of the Perron–Frobenius theorem remain valid for the class of homogeneous systems of *positive order*, which are cooperative and irreducible, if the invariant ray is stable or asymptotically stable, provided one restricts initial conditions to $\mathbb{R}^n_+$.
2. If an invariant ray is unstable, then it is not necessarily unique. In case there are several invariant rays in $\mathbb{R}^n_+$, all of them belong to $\text{int}(\mathbb{R}^n_+)$, and they are all unstable. The zero solution is then unstable. This is in contrast with the Perron–Frobenius theorem for linear cooperative and irreducible systems, where the eigenvector associated with the dominating eigenvalue is always unique.

REFERENCES

[1] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Classics in Appl. Math. 9, SIAM, Philadelphia, 1994.

[2] R. E. Edwards, *Functional Analysis*, Dover, New York, 1995.
[3] W. Hahn, *Stability of Motion*, Springer-Verlag, Berlin, 1967.
[4] P. Hartman, *Ordinary Differential Equations*, Birkhäuser Boston, Cambridge, MA, 1982.
[5] M. W. Hirsch, *Systems of differential equations that are competitive or cooperative* II: *Convergence almost everywhere*, SIAM J. Math. Anal., 16 (1985), pp. 423–439.
[6] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press, Cambridge, UK, 1998.
[7] P. De Leenheer and D. Aeyels, *A note on uniform boundedness of a class of positive systems*, in Proceedings of the 38th Conference on Decision and Control, Phoenix, AZ, 1999, pp. 2575–2579.
[8] P. De Leenheer and D. Aeyels, *Stability properties of equilibria of classes of cooperative systems*, IEEE Trans. Automatic Control, 46 (2001), pp. 1996–2001.
[9] D. G. Luenberger, *Introduction to Dynamic Systems*, John Wiley, New York, 1979.
[10] L. Moreau and D. Aeyels, *Approximation of Systems and Stability*, submitted.
[11] F. Jadot, *Dynamics and Robust Nonlinear PI Control of Stirred Tank Reactors*, Ph.D. thesis, Université Catholique de Louvain, Louvain, Belgium, 1996.
[12] J. Peuteman and D. Aeyels, *Averaging results and the study of uniform asymptotic stability of homogeneous differential equations that are not fast time-varying*, SIAM J. Control Optim., 37 (1999), pp. 997–1010.
[13] H. L. Smith, *Monotone Dynamical Systems*, AMS, Providence, RI, 1995.
[14] F. Viel, F. Jadot, and G. Bastin, *Global stabilization of exothermic chemical reactors under input constraints*, Automatica, 33 (1997), pp. 1437–1448.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 2, pp. 583–606

OPTIMAL CONTROLS OF 3-DIMENSIONAL NAVIER–STOKES EQUATIONS WITH STATE CONSTRAINTS*

GENGSHENG WANG†

Abstract. This work is concerned with the maximum principles for optimal control problems governed by 3-dimensional Navier–Stokes equations. Some types of state constraints (time variables) are considered.

Key words. optimal control, Navier–Stokes equation, state constraint, maximum principle

AMS subject classifications. 93C05, 93B50, 93C35

PII. S0363012901385769

1. Introduction.

In this paper, we shall study the optimal control problems governed by 3-dimensional Navier–Stokes equations

$$\begin{cases} \frac{\partial y}{\partial t} - \gamma \Delta y + y \cdot \nabla y + \nabla p = D_0 u + f_0 \ \text{ in } \Omega \times (0,T), \\ \nabla \cdot y = \operatorname{div} y = 0 \ \text{ in } \Omega \times (0,T); \ \ y = 0 \ \text{ in } \partial\Omega \times (0,T) \end{cases} \tag{1.1}$$

with some types of state constraints, which we shall state later. Here and throughout this paper, we shall omit (x,t) in all functions of (x,t) if there is no ambiguity. In (1.1), Ω is a bounded and open subset of R^3 with smooth boundary $\partial\Omega$, $T > 0$ is a given constant, $\gamma > 0$ is the viscosity constant, $f_0 \in L^2(0,T;L^2(\Omega))$ is a source field, $y(x,t) = (y_1(x,t), y_2(x,t), y_3(x,t))$ is the velocity vector, $\nabla \cdot y$ is the divergence of y, p stands for the pressure, and $u \in L^2(0,T;U)$ is input; here we have denoted by U a real Hilbert space and by D_0 a linear bounded operator from U to $(L^2(\Omega))^3$.

Let us introduce some functional spaces to represent the Navier–Stokes equation (1.1) as infinite dimensional differential equations. For the details, we refer the reader to [7] and [18].

Let V be the divergence free subspace of $(H_0^1(\Omega))^3$; i.e.,

$$V = \{y \in (H_0^1(\Omega))^3 : \nabla \cdot y = 0 \ \text{ in } \Omega\}$$

and

$$H = \{y \in (L^2(\Omega))^3 : \nabla \cdot y = 0 \ \text{ in } \Omega; \ \ n \cdot y = 0 \ \text{ on } \partial\Omega\}.$$

The space H is endowed with the usual $(L^2(\Omega))^3$-norm denoted by $|\cdot|$ and V with the norm $\|\cdot\|$ defined by

$$\|y\|^2 = \sum_{1 \le i \le 3} \int_\Omega |\nabla y_i|^2 dx, \quad y = (y_1, y_2, y_3).$$

*Received by the editors March 5, 2001; accepted for publication (in revised form) January 13, 2002; published electronically July 1, 2002. The author's work was supported by the National Natural Science Foundation of China under grants 10071028 and 60174043 and by the Key Program Foundation of the Ministry of National Education of China.

http://www.siam.org/journals/sicon/41-2/38576.html

†Department of Mathematics, Huazhong Normal University, Wuhan 430079, People's Republic of China (wanggs@ccnu.edu.cn).

We shall denote by $\langle\cdot,\cdot\rangle$ the scalar product of H and the pairing between V and its dual V^* with the norm $\|\cdot\|_*$. Let $A \in L(V,V^*)$ and $b : V\times V\times V \to R$ be defined by

$$\langle Ay, z\rangle = \sum_{1\le i,j\le 3}\int_\Omega \nabla y_i\cdot\nabla z_i dx \quad \forall\ y,z\in V \tag{1.2}$$

and

$$b(y,z,w) = \sum_{1\le i,j\le 3}\int_\Omega y_i D_i z_j w_j dx \quad \forall\ y,z,w\in V, \tag{1.3}$$

respectively, where $D_i = \frac{\partial}{\partial x_i}$. We define $B : V\to V^*$ by

$$\langle B(y), w\rangle = b(y,y,w) \quad \forall\ y,w\in V. \tag{1.4}$$

Let $f(t) = Pf_0(t)$ and $D\in L(U,H)$ be given by $D = PD_0$, where $P : (L^2(\Omega))^3 \to H$ is the projection on H. Then we may rewrite the state system (1.1) as

$$y'(t) + \gamma Ay(t) + By(t) = Du(t) + f(t) \quad \text{a.e. in } (0,T). \tag{1.5}$$

The general functional framework which will be in effect throughout this paper is explained in the following hypotheses.

(H_1) V and H are two real Hilbert spaces with the norms $\|\cdot\|$, $|\cdot|$ and the scalar products $\langle\cdot,\cdot\rangle_V$ and $\langle\cdot,\cdot\rangle$. Moreover, $V\subset H\subset V^*$ algebraically and topologically with compact injection, where we identify H with its own dual and denote by V^* the dual space of V, with norm denoted by $\|\cdot\|_*$. Denote again by $\langle\cdot,\cdot\rangle$ the pairing between V and V^*.

(H_2) $\gamma > 0$ is a constant. $A\in L(V,V^*)$ is symmetric and coercive, satisfying that $\langle Ay, y\rangle = \|y\|^2$ for all $y\in V$. We set $D(A) = \{y\in V : Ay\in H\}$ and denote by A again the restriction of A to H. Then, as we know (cf. [3]), A is a positive, self-adjoint operator, and $D(A^{\frac12}) = V$.

(H_3) The operator $B : V\to V^*$ is defined by (1.4), where $b : V\times V\times V\to R$ is a trilinear continuous functional satisfying

$$b(y,z,w) = -b(y,w,z) \quad \forall\ y,z,w\in V, \tag{1.6}$$

$$|b(y,z,w)| \le C\cdot\begin{cases} |y|^{\frac12}\|y\|^{\frac12}\|z\|\|w\|, \\ \|y\|\|z\|\|w\|^{\frac12}|w|^{\frac12} \end{cases} \quad \forall\ y,z,w\in V, \tag{1.7}$$

$$|b(y,z,w)| \le C\cdot\begin{cases} \|y\|^{\frac12}|Ay|^{\frac12}\|z\||w| & \forall\ y\in D(A), z\in V, w\in H, \\ \|y\|\|z\|^{\frac12}|Az|^{\frac12}|w| & \forall\ y\in V, z\in D(A), w\in H, \\ |y|\|z\|\|w\|^{\frac12}|Aw|^{\frac12} & \forall\ y\in H, v\in V, w\in D(A), \end{cases} \tag{1.8}$$

where C denotes several positive constants.

(H_4) $D\in L(U,H)$, $f\in L^2(0,T;H)$, where U is a real Hilbert space with the norm denoted by $|\cdot|_U$ and the scalar product $\langle\cdot,\cdot\rangle_U$.

Recall that assumption (H_3) is satisfied for b given by (1.3) (see [18]). We shall denote by Y the space $W^{1,2}([0,T];H)\cap L^2(0,T;D(A))$, where $W^{1,2}([0,T];H)$ is the space of all absolutely continuous functionals $y : [0,T]\to H$ such that $y' = \frac{dy}{dt}\in L^2(0,T;H)$. We have (cf. [3]) that $Y\subset C([0,T];V)$. We shall denote by $W(0,T)$ the space $\{y : y\in L^2(0,T;V), y'\in L^2(0,T;V^*)\}$ endowed with the norm $\|y\|_{W(0,T)} = [\int_0^T\|y\|^2dt + \int_0^T\|y'\|_*dt]^{\frac12}$. We have (cf. [12, Chapter 4]) that $W(0,T)\subset C([0,T];H)$.

The cost functional we shall study in this paper is as follows:

$$L(y,u)=\int_0^T [g(t,y(t))+h(u(t))]dt, \tag{1.9}$$

where we assume the following.

(H_5) $g:[0,T]\times V\to R^+$ is measurable in the first variable, $g(t,0)\in L^\infty(0,T)$, and for every $r>0$, there exists an $L_r>0$ independent of t such that

$$|g(t,y_1)-g(t,y_2)|\le L_r\|y_1-y_2\| \quad \forall\, t\in[0,T], \|y_1\|+\|y_2\|\le r. \tag{1.10}$$

$h:U\to \bar{R}\equiv(-\infty,\infty]$ is convex and lower semicontinuous. Moreover, there exist $c_1>0$ and $c_2\in R$ such that

$$h(u)\ge c_1|u|_U^2-c_2 \quad \forall\, u\in U. \tag{1.11}$$

In section 4, we shall give some explicit forms of the integrand $g(t,y)$, such as $|\operatorname{curl} y|^2$, which physically motivate the results.

In this paper, we shall derive the Pontryagin maximum principle for optimal control governed by system (1.5) with three types of state constraints, including a type of integral, a type of two point boundary (time variable), and a periodic type. The periodic type is a special case of a type of two point boundary. However, in the maximum principle for optimality of the two point boundary case, obtained in this paper (Theorem 2.4), we do not know if the multiplier is nonzero, i.e., if the maximum principle is qualified for the two point boundary case. Even for the periodic case, we cannot get a qualified maximum principle by the same methods as those used in getting the maximum principle for the two point boundary case. Such difficulty exists also for the optimal control governed by a semilinear parabolic equation with a two point boundary state constraint (cf. [16], [17], and [20]). This stimulates us to investigate, in particular, the periodic case to derive a qualified maximum principle. Now we formulate our problems as follows.

The first optimal control problem (P_1) we shall study in this paper is as follows.

(P_1) $\inf L(y,u)$ over all $(y,u)\in Y\times L^2(0,T;U)$ subject to

$$\begin{cases} y'+\gamma Ay+By=Du+f \quad \text{a.e. in } (0,T),\\ y(0)=y_0, \end{cases} \tag{1.12}$$

with

$$F(y)\in W, \tag{1.13}$$

where $y_0\in V$ and we assume the following.

(H_6) $F:L^2(0,T;V)\to X$ is continuously Frechet differentiable, where X is a Banach space with the dual X^* strictly convex, and $W\subset X$ is a closed and convex subset.

The second problem (P_2) we shall study is as follows.

(P_2) $\inf L(y,u)$ over all $(y,u)\in Y\times L^2(0,T;U)$ subject to (1.5) and

$$(y(0),y(T))\in S, \tag{1.14}$$

where we assume the following.

(H_7) $S\subset H\times H$ is a closed and convex subset.

The third problem (P_3) we shall study is as follows.

(P_3) $\inf L(y,u) = \int_0^T [g(t,y) + h(u)]dt$ over all $(y,u) \in Y \times L^2(0,T;U)$ subject to (1.5) and

$$y(0) = y(T). \tag{1.15}$$

In (P_3), we modify the assumption on the functional g as follows.

(H_8) $g : [0,T] \times H \to R^+$ is measurable in the first variable, $g(t,0) \in L^\infty(0,T)$, and for each $r > 0$, there exists an $L_r > 0$ independent of t such that

$$|g(t,y_1) - g(t,y_2)| \le L_r|y_1 - y_2| \quad \forall\, t \in [0,T]; |y_1| + |y_2| \le r.$$

For both physical and mathematical reasons, we consider the space Y as the state space in our problems. Physically, one needs the optimal state to have some smoothness, and, mathematically, because of the complexity of operator B, it is very difficult to analyze the linearized system corresponding to (1.12) around a state y, which is a weak solution to (1.12); i.e., $y \in L^2(0,T;V)$ and $y' \in L^1(0,T;V^*)$, even though such a weak solution exists for each $u \in L^2(0,T;U)$ (cf. [7] and [18]). However, as we know (cf. [7] and [18]), for each $y_0 \in V$, $u \in L^2(0,T;U)$, there exists $T^* \equiv T^*(y_0,u)$ such that (1.12) has a unique solution in $W^{1,2}([0,T_1];H) \cap L^2(0,T_1;D(A))$ for all $T_1 < T^*$. Thus, for $T > 0$ given, (1.7) may have no solution in $W^{1,2}([0,T];H) \cap L^2(0,T;D(A))$ for each $u \in L^2(0,T;U)$ and $y_0 \in V$. So problems (P_1), (P_2), and (P_3) are non-well-posed optimal control problems.

One of the main difficulties is that system (1.2) may have no solution in Y for each $u \in L^2(0,T;U)$, and the estimates on the operator B defined in (1.4) are weaker than those for the 2-dimensional Navier–Stokes equation. This makes us unable to apply the general techniques in, for instance, [3], [8], [17], and [19] to study the variations of the state with respect to the controls and to study the adjoint system and linearized system corresponding to (1.5). We overcome such difficulty by thinking of system (1.5) as a constraint mixed by the state and the control and then by introducing kinds of penalty functionals so that we may transfer the original non-well-posed control problems into optimization problems, where the variables y and u are independent. Thus we do not need to analyze the variations of the state with respect to the control. Instead, we use some special way involving the generic solvability theorem for linear parabolic evolution systems to construct the adjoint state. We believe that there are other ways to approach such problems. For instance, we may consider the local optimal control problem governed by (1.5), which means that the control set is taken as $U_{ad} = \{u \in L^2(0,T;U) : \|u\|_{L^2(0,T;U)} \le r\}$ for some small $r > 0$ depending on T. In such a way, system (1.12) has a unique solution $y \in Y$ corresponding to each $u \in U_{ad}$. However, we believe the methods deployed in this paper are more constructive.

Another difficulty is caused by the involvement of state constraints (1.13) and (1.14) and the integrand $g(t,y)$, which is allowed to be from $R^+ \times V$ to R (instead of from $R^+ \times H$ to R). This makes the analyses of the linearized system and the adjoint system of (1.5) more complicated.

For other literature on optimal control problems governed by Navier–Stokes equations and related to this paper, we cite [3], [4], [5], [6], [8], [9], [11], [12], and [13].

The outline of this paper is as follows. In section 2, we give and prove the necessary conditions for optimalities of problems (P_1) and (P_2). In section 3, we get the first order necessary conditions of optimality for problem (P_3) in terms of the Euler–Lagrange system. In section 4, we give some examples covered by the form of cost functionals (1.9) and the form of state constraints (1.13) and (1.14).

2. Optimal control with state constraint of integral type. Let (y^*, u^*) be optimal for problem (P_1). In this section, we shall state and prove the necessary conditions for (y^*, u^*). First we recall approximations g^ε of g and h_ε of h as follows. For the details, we refer the reader to [1, Chapter 1]. Let $g^\varepsilon : [0,T] \times V \to R^+$ be defined by

$$g^\varepsilon(t,y) = \int_{R^m} g(t, P_m y - \varepsilon \Lambda_m \tau)\rho_m(\tau)d\tau \quad \forall\ y \in V, \tag{2.1}$$

where $m = [\varepsilon^{-1}]$, ρ_m is a mollifier on R^m, $P_m : V \to X_m$ is the projection operator from H to X_m, which is the finite dimensional space generated by $\{e_i\}_{i=1}^m$, where $\{e_i\}_{i=1}^\infty$ is an orthonormal basis in V, and $\Lambda_m : R^m \to X_m$ is defined by $\Lambda_m(\tau) = \sum_{i=1}^m \tau_i e_i$, $\tau = (\tau_1, \dots, \tau_m)$. Let $h_\varepsilon : U \to R$ be defined by

$$h_\varepsilon(u) = \inf\{|u - v|_U^2/(2\varepsilon) + h(v) : v \in U\}. \tag{2.2}$$

Now, for each $\varepsilon > 0$, we define a penalty functional $L_\varepsilon : Y \times L^2(0,T;U) \to R$ by

$$\begin{aligned} L_\varepsilon(y,u) = {} & \int_0^T [g^\varepsilon(t,y) + h_\varepsilon(u)]dt + \frac{1}{4}\int_0^T \|y - y^*\|^4 dt + \frac{1}{2}\int_0^T |u - u^*|^2 dt \\ & + \frac{1}{2\varepsilon}\int_0^T |y' + \gamma Ay + By - Du - f|^2 dt + \frac{1}{2\varepsilon}[\varepsilon + d_W(F(y))]^2. \end{aligned} \tag{2.3}$$

Since $y \in Y \subset C([0,T];V)$, L_ε is well defined, and we may define $Y_0 = \{y \in Y : y(0) = y_0\}$. Consider the approximation problem $(P_{1\varepsilon})$ as follows.

$(P_{1\varepsilon})$ $\inf L_\varepsilon(y,u)$ over all $(y,u) \in Y_0 \times L^2(0,T;U)$.

We have the following existence and approximation results for problem $(P_{1\varepsilon})$.

LEMMA 2.1. *For each $\varepsilon > 0$, problem $(P_{1\varepsilon})$ has at least one solution.*

Proof. It is clear that $\inf(P_{1\varepsilon}) > -\infty$. Let $(y_n, u_n) \in Y_0 \times L^2(0,T;U)$ be such that

$$\inf(P_{1\varepsilon}) \le L_\varepsilon(y_n, u_n) \le \inf(P_{1\varepsilon}) + \frac{1}{n}, \quad n = 1, 2, \dots. \tag{2.4}$$

By (2.3) and (2.4), we imply that

$$\|u_n\|_{L^2(0,T;U)} \le C \tag{2.5}$$

and

$$\|y_n\|_{L^4(0,T;V)} \le C; \tag{2.6}$$

here and throughout the proof of Lemma 2.1, we shall denote by C several positive constants independent of n. By (2.3) and (2.4) again, there exist $v_n \in L^2(0,T;H)$, $n = 1, 2, \dots$, such that

$$\|v_n\|_{L^2(0,T;H)} \le C \tag{2.7}$$

and

$$\begin{cases} y_n' + \gamma A y_n + B y_n = D u_n + v_n \quad \text{a.e. in } (0,T), \\ y_n(0) = y_0. \end{cases} \tag{2.8}$$

Multiplying (2.8) by y_n, integrating on $(0,t)$, using Gronwall's inequality, and noting that $\langle By_n, y_n\rangle = 0$, which is from (1.6), we get that

$$|y_n(t)|^2 + \int_0^T \|y_n(t)\|^2 dt \le C \quad \forall\, t \in [0,T]. \tag{2.9}$$

It follows from (1.8) that

$$\begin{aligned}\int_0^t |\langle By_n, Ay_n\rangle| ds &\le C\int_0^t \|y_n\|^{\frac{3}{2}} |Ay_n|^{\frac{3}{2}} ds \\ &\le \frac{\gamma}{4}\int_0^t |Ay_n|^2 ds + C_\gamma \int_0^t \|y_n\|^6 ds;\end{aligned} \tag{2.10}$$

here and throughout what follows, C_γ denotes several positive constants independent of n but dependent on γ.

Multiplying (2.8) by Ay_n and integrating on $(0,t)$, by (2.10), (2.5), and (2.7) we get that

$$\|y_n(t)\|^2 + \gamma\int_0^T |Ay_n(t)|^2 dt \le C_\gamma \left(1 + \int_0^t \|y_n(s)\|^4 \|y_n(s)\|^2 ds\right).$$

Then, by (2.6) and by Gronwall's inequality, we obtain that

$$\|y_n(t)\|^2 + \gamma\int_0^T |Ay_n(t)|^2 dt \le C \quad \forall\, t \in [0,T]. \tag{2.11}$$

Now it follows from (1.8) and (2.11) that, for each $w \in H$,

$$|\langle By_n, w\rangle| = |b(y_n, y_n, w)| \le C\|y_n\|^{\frac{3}{2}} |Ay_n|^{\frac{1}{2}} |w| \le C|Ay_n|^{\frac{1}{2}} |w|,$$

which implies that

$$\int_0^T |By_n|^2 ds \le C. \tag{2.12}$$

By (2.11), (2.12), and (2.8), we obtain that

$$\|y_n'\|_{L^2(0,T;H)} \le C. \tag{2.13}$$

Thus by (2.5), (2.6), (2.11), (2.13), the Arezala–Ascoli theorem, and the Aubin compactness theorem, we conclude that there exist $(\tilde{y}, \tilde{u}) \in Y \times L^2(0,T;U)$ and subsequences of $\{y_n\}$ and $\{u_n\}$, still denoted by themselves, such that, as $n \to \infty$,

$$\begin{aligned} y_n &\to \tilde{y} \quad \text{strongly in } C([0,T];H) \cap L^2(0,T;V), \\ &\quad\ \text{weakly in } L^2(0,T;D(A)) \cap L^4(0,T;V), \end{aligned} \tag{2.14}$$

$$y_n' \to \tilde{y}' \ \text{ weakly in } L^2(0,T;H), \tag{2.15}$$

$$u_n \to \tilde{u} \ \text{ weakly in } L^2(0,T;U). \tag{2.16}$$

Next we claim that

$$By_n \to B\tilde{y} \ \text{ strongly in } L^2(0,T;H) \ \text{ as } n \to \infty. \tag{2.17}$$

To this end, we observe first that

$$\begin{aligned}|\langle By_n - B\tilde{y}, w\rangle| &\le |b(y_n - \tilde{y}, y_n, w)| + |b(y_n, y_n - \tilde{y}, w)| \\ &\le C[\|y_n - \tilde{y}\|^{\frac{1}{2}}|A(y_n - \tilde{y})|^{\frac{1}{2}}\|y_n\| + \|y_n\|^{\frac{1}{2}}|Ay_n|^{\frac{1}{2}}\|y_n - \tilde{y}\|]|w| \\ &\le C[\|y_n - \tilde{y}\|^{\frac{1}{2}}|A(y_n - \tilde{y})|^{\frac{1}{2}} + |Ay_n|^{\frac{1}{2}}\|y_n - \tilde{y}\|]|w|\end{aligned}$$

for each $w \in H$. Since $\|y_n - \tilde{y}\|_{C([0,T];V)} \le C$, which is from (2.11), the above inequality implies that

$$\begin{aligned}\int_0^T |By_n - B\tilde{y}|^2 dt &\le C\int_0^T [\|y_n - \tilde{y}\||A(y_n - \tilde{y})| + |Ay_n|\|y_n - \tilde{y}\|^2]dt \\ &\le C\left\{\left[\int_0^T \|y_n - \tilde{y}\|^2 dt\right]^{\frac{1}{2}}\left[\int_0^T |A(y_n - \tilde{y})|^2 dt\right]^{\frac{1}{2}}\right. \\ &\qquad \left. + \|y_n - \tilde{y}\|_{C([0,T];V)}\left[\int_0^T |Ay_n|^2 dt\right]^{\frac{1}{2}}\left[\int_0^T \|y_n - \tilde{y}\|^2 dt\right]^{\frac{1}{2}}\right\} \\ &\to 0 \ \text{ as } n \to \infty.\end{aligned}$$

By (2.14), (2.15), (2.16), and (2.17), we infer that

$$\underline{\lim}_{n\to\infty}\int_0^T \|y_n - y^*\|^4 dt \ge \int_0^T \|\tilde{y} - y^*\|^4 dt \tag{2.18}$$

and

$$\underline{\lim}_{n\to\infty}\int_0^T |y_n' + \gamma Ay_n + By_n - Du_n - f|^2 dt \ge \int_0^T |\tilde{y}' + \gamma A\tilde{y} + B\tilde{y} - D\tilde{u} - f|^2 dt. \tag{2.19}$$

By (1.10), (2.1), (2.11), and (2.14), we imply that

$$\int_0^T |g^\varepsilon(t, y_n) - g^\varepsilon(t, \tilde{y})|dt \le L\int_0^T \|y_n - \tilde{y}\|dt \to 0 \ \text{ as } n \to \infty, \tag{2.20}$$

where $L > 0$ is independent of n. Since h_ε is convex and continuous, it follows from (2.16) that

$$\underline{\lim}_{n\to\infty}\int_0^T h_\varepsilon(u_n)dt \ge \int_0^T h_\varepsilon(\tilde{u})dt. \tag{2.21}$$

By (2.14) and (H_6), $F(y_n) \to F(\tilde{y})$ as $n \to \infty$. Thus we have that

$$\frac{1}{2\varepsilon}(\varepsilon + d_W(F(y_n)))^2 \to \frac{1}{\varepsilon}(\varepsilon + d_W(F(\tilde{y})))^2 \ \text{ as } n \to \infty. \tag{2.22}$$

On the other hand, since $y_n(0) = y_0$ and $y_n(0) \to \tilde{y}(0)$ strongly in H, we have that $\tilde{y}(0) = y_0$, which shows that $\tilde{y} \in Y_0$. Thus it follows immediately from (2.4) and (2.18)–(2.22) that $(\tilde{y}, \tilde{u})$ is optimal for problem $(P_{1\varepsilon})$. This completes the proof.

LEMMA 2.2. *Let $(y_\varepsilon, u_\varepsilon)$ be optimal for problem $(P_{1\varepsilon})$. Then there exists a generalized subsequence of $(y_\varepsilon, u_\varepsilon)$, still denoted by itself, such that*

$$u_\varepsilon \to u^* \ \textit{strongly in } L^2(0,T;U), \quad y_\varepsilon \to y^* \ \textit{strongly in } Y \ \textit{ as } \varepsilon \to 0.$$

Remark. We say $y_\varepsilon \to y^*$ strongly in Y as $\varepsilon \to 0$ if $y_\varepsilon \to y^*$ strongly in $L^2(0,T;D(A)) \cap C([0,T];V)$ and $dy_\varepsilon/dt \to dy^*/dt$ strongly in $L^2(0,T;H)$ as $\varepsilon \to 0$.

Proof. Since $(y_\varepsilon, u_\varepsilon)$ is optimal for $(P_{1\varepsilon})$, it follows from (2.3) that

$$L_\varepsilon(y_\varepsilon, u_\varepsilon) \le L_\varepsilon(y^*, u^*) = \int_0^T [g^\varepsilon(t, y^*) + h_\varepsilon(u^*)]dt + \frac{\varepsilon}{2}. \tag{2.23}$$

By (2.1) and by the same argument as in [1, Chapter 3], we get that

$$|g^\varepsilon(t, y^*) - g(t, y^*)| \le L(\|y^* - P_m y^*\| + \varepsilon), \tag{2.24}$$

where $L > 0$ is a constant independent of ε, and the projection operator P_m was given in (2.1). By (2.23), (2.24), and the same argument as in [17], we get that

$$\overline{\lim}_{\varepsilon \to 0} L(y_\varepsilon, u_\varepsilon) \le L(y^*, u^*). \tag{2.25}$$

On the other hand, it follows from (2.3) and (2.25) that

$$\|y_\varepsilon\|_{L^4(0,T,V)} + \|u_\varepsilon\|_{L^2(0,T;U)} \le C, \tag{2.26}$$

$$\int_0^T |y'_\varepsilon + \gamma A y_\varepsilon + B y_\varepsilon - D u_\varepsilon - f|^2 dt \le 2C\varepsilon, \tag{2.27}$$

and

$$[\varepsilon + d_W(F(y_\varepsilon))]^2 \le 2C\varepsilon. \tag{2.28}$$

Here and throughout the proof of Lemma 2.2, we shall denote by C several positive constants independent of ε.

By (2.27), there exists a $v_\varepsilon \in L^2(0,T;H)$ for each $\varepsilon > 0$ such that $\|v_\varepsilon\|_{L^2(0,T;H)} \to 0$ as $\varepsilon \to 0$, and

$$\begin{cases} y'_\varepsilon + \gamma A y_\varepsilon + B y_\varepsilon = D u_\varepsilon + v_\varepsilon + f \quad \text{a.e. in } (0,T), \\ y_\varepsilon(0) = y_0. \end{cases} \tag{2.29}$$

By (2.26) and (2.29), using the same argument as in the proof of Lemma 2.1, we obtain that there exist $\tilde{y} \in Y$, $\tilde{u} \in L^2(0,T;U)$, and subsequences of $\{y_\varepsilon\}$ and $\{u_\varepsilon\}$, still denoted by themselves, such that, as $\varepsilon \to 0$,

$$y_\varepsilon \to \tilde{y} \quad \begin{array}{l} \text{strongly in } C([0,T];H) \cap L^2(0,T;V), \\ \text{weakly in } L^2(0,T;D(A)), \end{array} \tag{2.30}$$

$$\|y_\varepsilon\|_{C([0,T];V)} \le C, \tag{2.31}$$

$$y'_\varepsilon \to \tilde{y}' \ \text{ weakly in } L^2(0,T;H), \tag{2.32}$$

and

$$u_\varepsilon \to \tilde{u} \ \text{ weakly in } L^2(0,T;U). \tag{2.33}$$

By (2.30) and (2.31) and by the same argument as in the proof of Lemma 2.1, we deduce that

$$B y_\varepsilon \to B y^* \ \text{ strongly in } L^2(0,T;H) \ \text{ as } \varepsilon \to 0. \tag{2.34}$$

Thus, by (2.30)–(2.34), we may pass to the limit for $\varepsilon \to 0$ in (2.29) to derive that

$$\begin{cases} \tilde{y}' + \gamma A\tilde{y} + B\tilde{y} = D\tilde{u} + f \quad \text{a.e. in } (0,T), \\ \tilde{y}(0) = y_0. \end{cases} \tag{2.35}$$

It follows from (2.30) and (H_6) that

$$F(y_\varepsilon) \to F(\tilde{y}) \ \text{ strongly in } X \ \text{ as } \varepsilon \to 0. \tag{2.36}$$

However, by (2.28), we have that $d_W(F(y_\varepsilon)) \to 0$ as $\varepsilon \to 0$, which, combined with (2.36), indicates that

$$F(\tilde{y}) \in W, \tag{2.37}$$

since W is closed. Thus, by (2.35) and (2.37), we infer that

$$L(y^*, u^*) \le L(\tilde{y}, \tilde{u}) \tag{2.38}$$

because (y^*, u^*) is optimal for (P).

Now by (2.1), (1.10), and (2.31), one can get that

$$|g^\varepsilon(t, y_\varepsilon) - g^\varepsilon(t, \tilde{y})| \le L\|y_\varepsilon - \tilde{y}\|, \tag{2.39}$$

$$\lim_{\varepsilon \to 0} g^\varepsilon(t, \tilde{y}(t)) = g(t, \tilde{y}(t)) \quad \forall\, t \in [0,T], \tag{2.40}$$

$$|g^\varepsilon(t, \tilde{y}(t)) - g(t, \tilde{y}(t))| \le L(\|\tilde{y} - P_m\tilde{y}\| + \varepsilon), \tag{2.41}$$

where $L > 0$ is independent of ε, and P_m was given in (2.1). Then, by (2.39), (2.40), (2.41), and the Lebesgue dominated convergence theorem, we get that

$$\begin{aligned} \int_0^T |g^\varepsilon(t, y_\varepsilon) - g(t, \tilde{y})|dt &\le \int_0^T [|g^\varepsilon(t, y_\varepsilon) - g^\varepsilon(t, \tilde{y})| + |g^\varepsilon(t, \tilde{y}) - g(t, \tilde{y})|]dt \\ &\le \int_0^T [L\|y_\varepsilon(t) - \tilde{y}(t)\| + |g^\varepsilon(t, \tilde{y}) - g(t, \tilde{y})|]dt \\ &\to 0 \ \text{ as } \varepsilon \to 0. \end{aligned} \tag{2.42}$$

By the same argument as in [1, Chapter 5], we deduce that

$$\underline{\lim}_{\varepsilon \to 0} \int_0^T \left[h_\varepsilon(u_\varepsilon) + \frac{1}{2}|u_\varepsilon - u^*|^2 \right] dt \ge \int_0^T \left[h(\tilde{u}) + \frac{1}{2}|\tilde{u} - u^*|^2 \right] dt. \tag{2.43}$$

Now it follows from (2.38), (2.42), and (2.43) that

$$\underline{\lim}_{\varepsilon \to 0} L_\varepsilon(y_\varepsilon, u_\varepsilon) \ge L(\tilde{y}, \tilde{u}) \ge L(y^*, u^*). \tag{2.44}$$

Thus, by (2.25) and (2.44), we infer that $\tilde{u} = u^*$, $\tilde{y} = y^*$,

$$u_\varepsilon \to u^* \ \text{ strongly in } L^2(0,T;U) \ \text{ as } \varepsilon \to 0, \tag{2.45}$$

and

$$y_\varepsilon \to y^* \ \text{ strongly in } L^4(0,T;V) \ \text{ as } \varepsilon \to 0. \tag{2.46}$$

Finally, we shall prove that $y_\varepsilon \to y^*$ strongly in Y as $\varepsilon \to 0$. To this end, we first observe that

$$\begin{cases} (y_\varepsilon - y^*)' + \gamma A(y_\varepsilon - y^*) + By_\varepsilon - By^* = D(u_\varepsilon - u^*) + v_\varepsilon \quad \text{a.e. in } (0,T), \\ (y_\varepsilon - y^*)(0) = 0. \end{cases} \tag{2.47}$$

Multiplying (2.47) by $A(y_\varepsilon - y^*)$ and integrating on $(0,t)$, we get

$$\begin{aligned} &\frac{1}{2}\|y_\varepsilon(t) - y^*(t)\|^2 + \gamma \int_0^T |A(y_\varepsilon - y^*)|^2 dt \\ &\quad \le \frac{\gamma}{2}\int_0^T |A(y_\varepsilon - y^*)|^2 dt + C_\delta \left\{ \int_0^T |By_\varepsilon - By^*|^2 dt + \int_0^T [|D(u_\varepsilon - u^*)|_U^2 + |v_\varepsilon|^2] dt \right\}. \end{aligned}$$

This together with (2.34) and (2.47) yields that

$$y'_\varepsilon \to (y^*)' \text{ strongly in } L^2(0,T;H) \text{ as } \varepsilon \to 0.$$

Hence $y_\varepsilon \to y^*$ strongly in Y as $\varepsilon \to 0$. This completes the proof.

Now we are in a position to state and prove the necessary conditions for (y^*, u^*). Let y_ε be a solution to problem $(P_{1\varepsilon})$, $\varepsilon > 0$, and define the operator $B'(y_\varepsilon(t)) : V \to V^*$ for each $t \in [0,T]$ by

$$\langle B'(y_\varepsilon(t))z, w\rangle = b(z, y_\varepsilon(t), w) + b(y_\varepsilon(t), z, w) \quad \forall\, z, w \in V. \tag{2.48}$$

It is clear that $B'(y_\varepsilon(t)) \in L(V, V^*)$ for each $t \in [0,T]$. The adjoint operator of $B'(y_\varepsilon(t))$, $[B'(y_\varepsilon(t))]^*$ is given by

$$\langle [B'(y_\varepsilon(t))]^* q, w\rangle = b(w, y_\varepsilon(t), q) + b(y_\varepsilon(t), w, q). \tag{2.49}$$

Because $\{y_\varepsilon\}$ is bounded in $C([0,T];V)$, it follows from (1.7), (2.48), and (2.49) that

$$\|B'(y_\varepsilon(t))\varphi\|_* \le C\|y_\varepsilon\|\|\varphi\| \le C\|\varphi\| \quad \forall\, \varphi \in V, \tag{2.50}$$

$$\|[B'(y_\varepsilon(t))]^*\varphi\|_* \le C\|y_\varepsilon\|\|\varphi\| \le C\|\varphi\| \quad \forall\, \varphi \in V, \tag{2.51}$$

$$|\langle B'(y_\varepsilon(t))\varphi, \varphi\rangle| \le C\|y_\varepsilon\||\varphi|^{\frac{1}{2}}\|\varphi\|^{\frac{3}{2}} \le \frac{\gamma}{2}\|\varphi\|^2 + C_\gamma|\varphi|^2 \quad \forall\, \varphi \in V, \tag{2.52}$$

and

$$|\langle [B'(y_\varepsilon)]^*\varphi, \varphi\rangle| \le \frac{\gamma}{2}\|\varphi\|^2 + C_\gamma|\varphi|^2 \quad \forall\, \varphi \in V. \tag{2.53}$$

By (2.50)–(2.53), using a standard existence result for linear evolution equations (cf. [12, Theorem 1.2, Chapter 3]), the Cauchy problems

$$\begin{cases} \varphi' + \gamma A\varphi + B'(y_\varepsilon)\varphi = g \quad \text{a.e. in } (0,T), \\ \varphi(0) = x \end{cases} \tag{2.54}$$

and

$$\begin{cases} -\psi' + \gamma A\psi + [B'(y_\varepsilon)]^*\psi = g \quad \text{a.e. in } (0,T), \\ \psi(T) = x \end{cases} \tag{2.55}$$

have unique solutions φ and ψ in $W(0,T)$ for each $g \in L^2(0,T;V^*)$ and $x \in H$, respectively. Moreover,

$$\|\varphi'\|^2_{L^2(0,T;V^*)} + \|\varphi\|^2_{L^2(0,T;V)} \le C(\|g\|^2_{L^2(0,T;V^*)} + |x|^2), \tag{2.56}$$

and

$$\|\psi'\|^2_{L^2(0,T;V^*)} + \|\psi\|^2_{L^2(0,T;V)} \le C(\|g\|^2_{L^2(0,T;V^*)} + |x|^2). \tag{2.57}$$

If $g \in L^2(0,T;H)$ and $x \in V$, then $\varphi, \psi \in Y$ (cf. [1]).

Similar to (2.48), we may define operator $B'(y^*(t)) : V \to V^*$ and obtain similar estimates to (2.56) and (2.57).

In order to get the necessary conditions for (y^*, u^*), we need one more assumption, as follows.

(H_9) The set $F'(y^*)R_r - W$ has finite codimensionality in X for some $r > 0$, where

$$M(0,r) = \{v \in L^2(0,T;U) : \|v\|_{L^2(0,T;U)} \le r\} \tag{2.58}$$

and

$$\begin{aligned} R_r = \{z \in Y : z' + \gamma Az + B'(y^*)z = Dv \quad \text{a.e. in } (0,T) \\ \text{and } z(0) = 0 \quad \text{for some } v \in M(0,r)\}. \end{aligned} \tag{2.59}$$

For the definition of a set to be finite codimensional in X and for related results, we refer the reader to [14]. Throughout what follows, we shall denote by $\partial g(t, y^*)$ the generalized derivative of g to the second variable at y^* and by $\partial h(u^*)$ the subdifferential of h at u^*. For the details, we refer the reader to [1]. We denote by $\langle \cdot, \cdot \rangle_{X^*,X}$ the pairing between X^* and X and by $[F'(y^*)]^*$ and D^* the adjoint operators of $F'(y^*)$ and D, respectively.

THEOREM 2.3. *Suppose that (H_1)–(H_6) hold. Let (y^*, u^*) be optimal for problem (P_1). Suppose further that (H_7) holds. Then there exists a triplet $(\lambda_0, \varphi, \xi_0) \in R \times L^2(0,T;V) \cap W(0,T) \times X^*$ with $(\lambda_0, \xi_0) \ne 0$ such that*

$$\begin{cases} -p' + \gamma Ap + [B'(y^*)]^*p + [F'(y^*)]^*\xi_0 \in -\lambda_0 \partial g(t, y^*) & \text{a.e. in } (0,T), \\ p(T) = 0, \end{cases} \tag{2.60}$$

$$\langle \xi_0, w - F(y^*) \rangle_{X^*,X} \le 0 \quad \forall\, w \in W, \tag{2.61}$$

and

$$D^*p(t) \in \lambda_0 \partial h(u^*(t)) \quad \text{a.e. in } (0,T). \tag{2.62}$$

Moreover, if $F'(y^*)$ is injective, then $(\lambda_0, p) \ne 0$.

Proof. Let $Z = \{y \in Y : y(0) = 0\}$. For $z \in Z$, $v \in L^2(0,T;U)$, we set $y^\rho_\varepsilon = y_\varepsilon + \rho z$, $u^\rho_\varepsilon = u_\varepsilon + \rho v$, where $(y_\varepsilon, u_\varepsilon)$ is optimal for problem $(P_{1\varepsilon})$. It is clear that $y^\rho_\varepsilon \in Y_0$, $u^\rho_\varepsilon \in L^2(0,T;U)$,

$$y^\rho_\varepsilon \to y_\varepsilon \ \text{ strongly in } Y \ \text{ as } \rho \to 0,$$

and

$$u^\rho_\varepsilon \to u_\varepsilon \ \text{ strongly in } L^2(0,T;U) \ \text{ as } \rho \to 0.$$

One can easily check that

$$(By^\rho_\varepsilon - By_\varepsilon)/\rho = B'(y_\varepsilon)z + \rho B(y_\varepsilon). \tag{2.63}$$

Hence

$$(By_\varepsilon^\rho - By_\varepsilon)/\rho \to B'(y_\varepsilon)z \text{ strongly in } L^2(0,T;H) \text{ as } \rho \to 0. \tag{2.64}$$

It follows from (2.64) that

$$\begin{aligned}\frac{1}{2\varepsilon\rho}\int_0^T [|(y_\varepsilon^\rho)' + \gamma A y_\varepsilon^\rho + By_\varepsilon^\rho - Du_\varepsilon^\rho - f|^2 - |y_\varepsilon' + \gamma A y_\varepsilon + By_\varepsilon - Du_\varepsilon - f|^2]dt \\ \to \int_0^T \langle q_\varepsilon, z' + \gamma Az + B'(y_\varepsilon)z - Dv\rangle dt \text{ as } \rho \to 0,\end{aligned} \tag{2.65}$$

where $q_\varepsilon = \frac{1}{\varepsilon}[y_\varepsilon' + \gamma A y_\varepsilon + By_\varepsilon - Du_\varepsilon - f]$. Since $\langle Ay, z\rangle = \langle y, z\rangle_V$ for all $y \in D(A), z \in V$, we infer that

$$\begin{aligned}\lim_{\rho\to 0}\frac{1}{4\rho}\int_0^T [\|y_\varepsilon^\rho - y^*\|^4 - \|y_\varepsilon - y^*\|^4]dt &= \int_0^T \|y_\varepsilon - y^*\|^2\langle y_\varepsilon - y^*, z\rangle_V dt \\ &= \int_0^T \|y_\varepsilon - y^*\|^2\langle A(y_\varepsilon - y^*), z\rangle dt.\end{aligned} \tag{2.66}$$

By the same argument as in [1, Chapter 5], we see that

$$\lim_{\rho\to 0}\frac{1}{\rho}\int_0^T [g^\varepsilon(t, y_\varepsilon^\rho) - g^\varepsilon(t, y_\varepsilon)]dt = \int_0^T \langle \nabla g^\varepsilon(t, y_\varepsilon), z\rangle dt \tag{2.67}$$

and

$$\begin{aligned}\lim_{\rho\to 0}\frac{1}{\rho}\int_0^T \left\{[h_\varepsilon(u_\varepsilon^\rho) - h_\varepsilon(u_\varepsilon)] + \frac{1}{2}[\|u_\varepsilon^\rho - u^*\|^2 - \|u_\varepsilon - u^*\|^2]\right\} dt \\ = \int_0^T \langle \nabla h_\varepsilon(u_\varepsilon) + u_\varepsilon - u^*, u\rangle_U dt.\end{aligned} \tag{2.68}$$

By the same argument as in [21], we get that

$$\begin{aligned}\lim_{\rho\to 0}[(\varepsilon + d_W(F(y_\varepsilon^\rho)))^2 - (\varepsilon + d_W(F(y_\varepsilon)))^2] \\ = \frac{\varepsilon + d_W(F(y_\varepsilon))}{\varepsilon}\langle \xi_\varepsilon, F'(y_\varepsilon)z\rangle_{X^*,X},\end{aligned} \tag{2.69}$$

where $\nabla g^\varepsilon(t, y_\varepsilon)$ denotes the gradient of g^ε to the second variable at y_ε and $\nabla h_\varepsilon(u_\varepsilon)$ denotes the gradient of h_ε at u_ε, while $\xi_\varepsilon \in \partial d_W(F(y_\varepsilon))$. Moreover,

$$\|\xi_\varepsilon\|_{X^*} = \begin{cases} 1 & \text{if } F(y_\varepsilon) \notin W, \\ 0 & \text{if } F(y_\varepsilon) \in W, \end{cases} \tag{2.70}$$

because W is convex and closed and X^* is strictly convex (cf. [15, Chapter 5]).

Since $(L_\varepsilon(y_\varepsilon^\rho, u_\varepsilon^\rho) - L_\varepsilon(y_\varepsilon, u_\varepsilon))/\rho \geq 0$ for all $\rho > 0$, it follows from (2.3) and (2.65)–(2.69) that

$$\begin{aligned}0 \leq \lambda_\varepsilon \int_0^T [&\langle \nabla g^\varepsilon(t, y_\varepsilon), z\rangle + \langle \nabla h_\varepsilon(u_\varepsilon), v\rangle_U \\ &+ \|y_\varepsilon - y^*\|^2\langle A(y_\varepsilon - y^*), z\rangle + \langle u_\varepsilon - u^*, v\rangle_U]dt \\ &+ \int_0^T \langle p_\varepsilon, z' + \gamma Az + B'(y_\varepsilon)z - Dv\rangle dt \\ &+ \int_0^T \langle [F'(y_\varepsilon)]^*\xi_\varepsilon, z\rangle dt \quad \forall\, z \in Z, v \in L^2(0,T;U),\end{aligned} \tag{2.71}$$

where

$$\lambda_\varepsilon = \frac{\varepsilon}{\varepsilon + d_W(F(y_\varepsilon))}, \quad p_\varepsilon = \lambda_\varepsilon q_\varepsilon \in L^2(0,T;H). \tag{2.72}$$

By taking $z = 0$ in (2.71), we obtain that

$$D^* p_\varepsilon = \lambda_\varepsilon \nabla h_\varepsilon(u_\varepsilon) + \lambda_\varepsilon(u_\varepsilon - u^*) \quad \text{a.e. in } (0,T), \tag{2.73}$$

while, by taking $v = 0$ in (2.71), we get that

$$\begin{aligned} 0 = &\int_0^T \langle \lambda_\varepsilon \nabla g^\varepsilon(t, y_\varepsilon) + [F'(y_\varepsilon)]^* \xi_\varepsilon + \lambda_\varepsilon \|y_\varepsilon - y^*\|^2 A(y_\varepsilon - y^*), z\rangle dt \\ &+ \int_0^T \langle p_\varepsilon, z' + \gamma A z + B'(y_\varepsilon) z\rangle dt \quad \forall\ z \in Z. \end{aligned} \tag{2.74}$$

We may regard (2.73) and (2.74) as the necessary conditions for $(y_\varepsilon, u_\varepsilon)$. Now we are in a position to pass to the limit for $\varepsilon \to 0$ in (2.73) and (2.74) to derive (2.62) and (2.60), respectively.

First we deal with (2.73). Note that $\alpha_\varepsilon \equiv \lambda_\varepsilon \nabla g^\varepsilon(t, y_\varepsilon) + [F'(y_\varepsilon)]^* \xi_\varepsilon + \lambda_\varepsilon \|y_\varepsilon - y^*\|^2 A(y_\varepsilon - y^*) \in L^2(0,T;V^*)$ and $\{\alpha_\varepsilon\}_{\varepsilon>0}$ is bounded in $L^2(0,T;V^*)$.

By (2.55), we may let $p_{\varepsilon 1} \in W(0,T)$ be the solution to

$$\begin{cases} -p'_{\varepsilon 1} + \gamma A p_{\varepsilon 1} + [B'(y_\varepsilon)]^* p_{\varepsilon 1} = -\alpha_\varepsilon \quad \text{a.e. in } (0,T), \\ p_{\varepsilon 1}(T) = 0. \end{cases} \tag{2.75}$$

Multiplying (2.75) by z and integrating on $(0,t)$, we get that

$$\int_0^T \langle p_{\varepsilon 1}, z' + \gamma A z + B'(y_\varepsilon) z\rangle dt = -\int_0^T \langle \alpha_\varepsilon, z\rangle dt.$$

This together with (2.74) implies that

$$\int_0^T \langle p_\varepsilon - p_{\varepsilon 1}, z' + \gamma A z + B'(y_\varepsilon) z\rangle dt = 0 \quad \forall\ z \in Z. \tag{2.76}$$

By (2.54), for each $g \in L^2(0,T;H)$, there exists $z \in Z$ such that $z' + \gamma A z + B'(y_\varepsilon) z = g$ in $(0,T)$. Thus it follows from (2.76) that $p_\varepsilon(t) = p_{\varepsilon 1}(t)$ a.e. in $(0,T)$. So $p_\varepsilon \in W(0,T)$ and satisfies

$$\|p'_\varepsilon\|^2_{L^2(0,T;V^*)} + \|p_\varepsilon\|^2_{L^2(0,T;V)} \le C\|\alpha_\varepsilon\|^2_{L^2(0,T;V^*)} \le C. \tag{2.77}$$

By the Aubin compactness theorem and the trace theorem (cf. [13, Theorem 3.1 of Chapter 1]), there exist $p \in W(0,T)$ and a subsequence of p_ε, still denoted by itself, such that

$$\begin{aligned} &p_\varepsilon \to p \ \text{ strongly in } L^2(0,T;H) \ \text{ weakly in } L^2(0,T;V) \ \text{ as } \varepsilon \to 0, \\ &p'_\varepsilon \to p' \ \text{ weakly in } L^2(0,T;V^*) \ \text{ as } \varepsilon \to 0, \\ &p_\varepsilon(0) \to p(0) \ \text{ weakly in } H \ \text{ as } \varepsilon \to 0. \end{aligned} \tag{2.78}$$

By (2.70) and (2.72), we have that

$$1 \le \lambda_\varepsilon + \|\xi_\varepsilon\|_{X^*} \le 2 \quad \forall\ \varepsilon > 0. \tag{2.79}$$

Thus there exist generalized subsequences of λ_ε and ξ_ε such that

$$\lambda_\varepsilon \to \lambda_0 \ \text{ as } \varepsilon \to 0, \tag{2.80}$$

and

$$\xi_\varepsilon \to \xi_0 \ \text{ weakly star in } X^* \ \text{ as } \varepsilon \to 0. \tag{2.81}$$

By Lemma 2.2 and (2.78), using the same argument as in [2, (1)], we may pass to the limit for $\varepsilon \to 0$ in (2.73) to derive (2.62).

Next we deal with (2.74), i.e., pass to the limit for $\varepsilon \to 0$ in (2.74).

By Lemma 2.2 and by the same argument as in [1, Chapter 5], we infer that

$$\nabla g^\varepsilon(t, y_\varepsilon) \to \beta \ \text{ weakly in } L^2(0,T;V^*) \ \text{ and } \ \beta(t) \in \partial g(t, y^*(t)) \ \text{ a.e. in } (0,T). \tag{2.82}$$

By (H_6), Lemma 2.2, and (2.81), we obtain that

$$[F'(y_\varepsilon)]^* \xi_\varepsilon \to [F'(y^*)]^* \xi_0 \ \text{ weakly in } L^2(0,T;V^*) \ \text{ as } \varepsilon \to 0. \tag{2.83}$$

By Lemma 2.2 again, we get that

$$\|y_\varepsilon - y^*\|^2 A(y_\varepsilon - y^*) \to 0 \ \text{ strongly in } L^2(0,T;H). \tag{2.84}$$

Now we claim that

$$[B'(y_\varepsilon)]^* p_\varepsilon \to [B'(y^*)]^* p \ \text{ weakly star in } L^2(0,T;V^*) \ \text{ as } \varepsilon \to 0. \tag{2.85}$$

Here is the argument. For any $w \in L^2(0,T;V)$, we have from (1.7) and (2.49) that

(2.86)

$$\begin{aligned}
&\int_0^T |\langle [B'(y_\varepsilon)]^* p_\varepsilon - [B'(y^*)]^* p, w\rangle| dt \\
&\quad = \int_0^T |b(w, y_\varepsilon, p_\varepsilon) + b(y_\varepsilon, w, p_\varepsilon) - b(w, y^*, p) - b(y^*, w, p)| dt \\
&\quad \le \int_0^T |b(w, y_\varepsilon - y^*, p_\varepsilon)| dt + \int_0^T |b(w, y^*, p_\varepsilon - p)| dt \\
&\qquad + \int_0^T |b(y_\varepsilon - y^*, w, p_\varepsilon)| dt + \int_0^T |b(y^*, w, p_\varepsilon - p)| dt \\
&\quad \le C \left[\int_0^T \|w\| \|p_\varepsilon\| \|y_\varepsilon - y^*\|^{\frac{1}{2}} |y_\varepsilon - y^*|^{\frac{1}{2}} dt + \int_0^T \|w\| \|y^*\| \|p_\varepsilon - p\|^{\frac{1}{2}} |p_\varepsilon - p|^{\frac{1}{2}} \right] \\
&\quad \le C \left[\|y_\varepsilon - y^*\|^{\frac{1}{2}}_{C([0,T];V)} \|y_\varepsilon - y^*\|^{\frac{1}{2}}_{C([0,T];H)} \left[\int_0^T \|w\|^2 dt \right]^{\frac{1}{2}} \left[\int_0^T \|p_\varepsilon\|^2 dt \right]^{\frac{1}{2}} \right. \\
&\qquad \left. + \|y^*\|_{C([0,T];V)} \left[\int_0^T \|w\|^2 dt \right]^{\frac{1}{2}} \left[\int_0^T \|p_\varepsilon - p\|^2 dt \right]^{\frac{1}{4}} \left[\int_0^T |p_\varepsilon - p|^2 dt \right]^{\frac{1}{4}} \right].
\end{aligned}$$

By (2.78), (2.86), and Lemma 2.2, we obtain (2.85) as claimed.

Thus, by (2.78), (2.80), and (2.82)–(2.85), we may pass to the limit for $\varepsilon \to 0$ in (2.75) to obtain that $p \in W(0,T)$ and satisfies (2.60).

On the other hand, since $\xi_\varepsilon \in \partial d_W(F(y_\varepsilon))$, we must have that

$$\langle \xi_\varepsilon, w - F(y_\varepsilon)\rangle_{X^*,X} \le 0 \quad \forall\, w \in W.$$

This implies that

$$\langle \xi_\varepsilon, w - F(y^*)\rangle_{X^*,X} \le \langle \xi_\varepsilon, F(y_\varepsilon) - F(y^*)\rangle_{X^*,X}. \tag{2.87}$$

Then, by Lemma 2.2, (H_6), and (2.81), we may pass to the limit for $\varepsilon \to 0$ in (2.87) to get (2.61).

We have proved (2.60), (2.61), and (2.62). Now we are in a position to show that $(\lambda_0, \xi_0) \neq 0$. To this end, we suppose that $\lambda_0 = 0$. Then, by (2.79) and (2.80), there exist $\varepsilon_1 > 0$ and $\delta > 0$ such that

$$2 \ge \|\xi_\varepsilon\|_{X^*} \ge \delta > 0 \quad \forall\, \varepsilon < \varepsilon_1. \tag{2.88}$$

It follows from (2.71) and (2.87) that

$$\begin{aligned} -\eta_\varepsilon(z,v) \le{}& \langle \xi_\varepsilon, F'(y^*)z - w + F(y^*)\rangle_{X^*,X} \\ &+ \int_0^T \langle p_\varepsilon, z' + \gamma Az + B'(y^*)z - Dv\rangle dt \end{aligned} \tag{2.89}$$

for all $(z,v) \in Z \times L^2(0,T;U)$ and $w \in W$, where

$$\begin{aligned} \eta_\varepsilon(z,v) = \lambda_\varepsilon &\left\{ \int_0^T [\langle \nabla g^\varepsilon(t, y_\varepsilon), z\rangle + \langle \nabla h_\varepsilon(u_\varepsilon), v\rangle_U] dt \right. \\ &+ \int_0^T \langle \|y_\varepsilon - y^*\|^2 A(y_\varepsilon - y^*), z\rangle dt \\ &\left. + \int_0^T \langle u_\varepsilon - u^*, v\rangle_U dt \right\} + \int_0^T \langle p_\varepsilon, [B'(y_\varepsilon) - B'(y^*)]z\rangle dt \\ &+ \langle \xi_\varepsilon, [F'(y_\varepsilon) - F'(y^*)]z + F(y_\varepsilon) - F(y^*)\rangle_{X^*,X}. \end{aligned} \tag{2.90}$$

For each $\varepsilon > 0$ and $v \in M(0,r)$, where $M(0,r)$ was given in (2.58) and $r > 0$ was given in (H_9), let $z_\varepsilon(v)$ be the solution to (2.54) with $g = Dv$ and $x = 0$; then $z_\varepsilon(v) \in Z$, and

$$\|z_\varepsilon(v)\|^2_{L^2(0,T;D(A))} + \|z'_\varepsilon(v)\|^2_{L^2(0,T;H)} \le C \quad \forall\, v \in M(0,r), \tag{2.91}$$

where $C > 0$ is independent of ε and v. Here we have used the estimate

$$|\langle B'(y_\varepsilon)z_\varepsilon, Az_\varepsilon\rangle| \le C\|z_\varepsilon\|^{\frac{1}{2}} \|y_\varepsilon\| |Az_\varepsilon|^{\frac{3}{2}} \le \frac{\gamma}{4}|Az_\varepsilon|^2 + C_\gamma \|z_\varepsilon\|^2,$$

which is from (1.8), (2.48), and Lemma 2.2.

It follows from (1.8) and (2.48) that

$$\begin{aligned} &\int_0^T |\langle p_\varepsilon, [B'(y_\varepsilon) - B'(y^*)]z_\varepsilon(v)\rangle| dt \\ &\quad \le C\int_0^T [|b(z_\varepsilon(v), y_\varepsilon - y^*, p_\varepsilon)| + |b(y_\varepsilon - y^*, z_\varepsilon(v), p_\varepsilon)|] dt \\ &\quad \le C\int_0^T \|z_\varepsilon(v)\| \|p_\varepsilon\| \|y_\varepsilon - y^*\|^{\frac{1}{2}} |y_\varepsilon - y^*|^{\frac{1}{2}} dt \\ &\quad \le C\|y_\varepsilon - y^*\|^{\frac{1}{2}}_{C([0,T];V)} \|y_\varepsilon - y^*\|^{\frac{1}{2}}_{C([0,T];H)} \|z_\varepsilon(v)\|_{L^2(0,T;V)} \|p_\varepsilon\|_{L^2(0,T;V)}. \end{aligned} \tag{2.92}$$

By (2.90), (2.91), (2.92), and Lemma 2.2, one can easily check that

$$\eta_\varepsilon(z_\varepsilon(v), v) \to 0 \ \text{ as } \varepsilon \to 0 \ \text{ uniformly in } v \in M(0,r). \tag{2.93}$$

Thus it follows from (2.89), (2.9), and (2.59) that

$$\langle \xi_\varepsilon, F'(y^*)z - w + F(y^*)\rangle_{X^*,X} \geq -\eta_\varepsilon \quad \forall\ z \in R_r, w \in W, \tag{2.94}$$

where $\eta_\varepsilon \to 0$ as $\varepsilon \to 0$. By (H_9), $F'(y^*)R_r - W$ has finite codimensionality in X, and so does $F'(y^*)R_r - W + F(y^*)$. Thanks to [14, Lemma 3.6], we conclude from (2.88), (2.93), and (2.94) that $(\lambda_0, \xi_0) \neq 0$.

Finally, if $F'(y^*)$ is injective and $(\lambda_0, p) \neq 0$, then it follows from (2.60) that $[F'(y^*)]^*\xi_0 = 0$, which implies that $\xi_0 = 0$. This contradiction leads $(\lambda_0, p) \neq 0$ and completes the proof.

Now we turn to present the maximum principle for problem (P_2). Let (y^*, u^*) be optimal for problem (P_2). We set

$$Q_{r_1,r_2} = \left\{ (z_0, z_T) \in V \times V : \begin{array}{l} \exists\ z \in Y \ \text{ with } \|z\|_Y \leq r_1 \ \text{ such that} \\ \mathcal{A}z \in D(M(0,r_2)) \ \text{ and } z(0) = z_0, z(T) = z_T \end{array} \right\}, \tag{2.95}$$

where $\mathcal{A}$ is the operator from Y to $L^2(0,T;H)$ defined by

$$\mathcal{A}z = z' + \gamma Az + B'(y^*)z, \tag{2.96}$$

$M(0,r_2)$ was defined in (2.58), and $D(M(0,r_2)) = \{Dv : v \in M(0,r_2)\}$. Then we assume the following.

(H_{10}) The set $Q_{r_1,r_2} - S$ has finite codimensionality in $H \times H$ for some $r_1, r_2 > 0$.

The main result for problem (P_2) will be as follows.

THEOREM 2.4. *Suppose that (H_1)–(H_5) and (H_7) hold. Let (y^*, u^*) be optimal for problem (P_2). Suppose further that (H_{10}) holds. Then there exists $(\lambda_0, p) \in R \times W(0,T)$ with $(\lambda_0, p) \neq 0$ such that*

$$\begin{aligned} -p' + \gamma Ap + [B'(y^*)]^*p &\in -\lambda_0 g(t, y^*) \quad \text{a.e. in } (0,T), \\ \langle p(0), x_0 - y^*(0)\rangle - \langle p(T), x_1 - y^*(T)\rangle &\leq 0 \quad \forall\ (x_0, x_1) \in S, \\ D^*p(t) &\in \lambda_0 \partial h(u^*(t)) \quad \text{a.e. in } (0,T). \end{aligned}$$

In this case, we introduce the penalty functional

$$\begin{aligned} L_\varepsilon(y,u) = &\int_0^T [g^\varepsilon(t,y) + h_\varepsilon(u)]dt + \frac{1}{2}\int_0^T |u - u^*|_U^2 dt + \frac{1}{2}\|y(0) - y^*(0)\|^2 \\ &+ \frac{1}{4}\int_0^T \|y - y^*\|^4 dt + \frac{1}{2\varepsilon}[\varepsilon + d_S(y(0), y(T))]^2 \\ &+ \frac{1}{2\varepsilon}\int_0^T |y' + \gamma Ay + By - Du - f|^2 dt, \end{aligned}$$

where $d_S(\cdot,\cdot)$ denotes the distance of $(\cdot,\cdot)$ to S in $H \times H$, and we consider the following approximation problem.

$(P_{2\varepsilon})$ $\inf L_\varepsilon(y,u)$ over all $(y,u) \in Y \times L^2(0,T;U)$.

Since the main ideas and steps in the proof of Theorem 2.4 are similar to those in the proof of Theorem 2.3, we omit the proof of Theorem 2.4 here.

3. Optimal control with periodic inputs. The form of state constraint (1.14) covers the periodic case; we will see this in Example 4.7 in section 4. However, in Theorem 2.4, we cannot get $\lambda_0 \neq 0$. In this section, we shall derive the qualified maximum principle for the periodic case (i.e., $\lambda_0 \neq 0$) by a different method.

Let $X = \{y \in Y : y(0) = y(T)\}$. Then X is dense in $L^2(0,T;H)$ (cf. [3]). Suppose that (H_8) holds and $g^\varepsilon : [0,T] \times H \to R^+$ is defined in the similar way as in (2.1), where we replace V by H, and h_ε is defined by (2.2). We introduce $L_\varepsilon : Y \times L^2(0,T;U) \to R$ by

$$\begin{aligned} L_\varepsilon(y,u) = &\int_0^T [g^\varepsilon(t,y) + h_\varepsilon(u)]dt + \frac{1}{4}\int_0^T \|y - y^*\|^4 dt \\ &+ \frac{1}{2}\int_0^T |u - u^*|_U^2 dt + \frac{1}{2\varepsilon}\int_0^T |y' + \gamma Ay + By - Du - f|^2 dt \end{aligned} \tag{3.1}$$

and consider the following approximation problem $(P_{3\varepsilon})$.

$(P_{3\varepsilon})$ $\inf L_\varepsilon(y,u)$ over all $(y,u) \in X \times L^2(0,T;U)$.

Similar to Lemmas 2.1 and 2.2, we may have the following existence and approximation results for problem $(P_{3\varepsilon})$. We omit the proofs here.

LEMMA 3.1. *For each $\varepsilon > 0$, problem $(P_{3\varepsilon})$ has at least one solution.*

LEMMA 3.2. *Let $(y_\varepsilon, u_\varepsilon) \in X \times L^2(0,T;U)$ be optimal for problem $(P_{3\varepsilon})$. Then*

$$\begin{aligned} &y_\varepsilon \to y^* \ \text{ strongly in } Y \ \text{ as } \varepsilon \to 0, \\ &u_\varepsilon \to u^* \ \text{ strongly in } L^2(0,T;U) \ \text{ as } \varepsilon \to 0. \end{aligned}$$

In the space $L^2(0,T;H)$, we define the operators

$$\mathcal{A}_\varepsilon \varphi = \varphi' + \gamma A\varphi + B'(y_\varepsilon)\varphi \quad \forall\ \varphi \in D(\mathcal{A}_\varepsilon) = X \tag{3.2}$$

and

$$\mathcal{A}_\varepsilon{}^*\varphi = -\varphi' + \gamma A\varphi + [B'(y_\varepsilon)]^*\varphi \quad \forall\ \varphi \in X, \tag{3.3}$$

where $B'(y_\varepsilon)$ and $[B'(y_\varepsilon)]^*$ were defined in (2.48) *and* (2.49), *respectively. It is readily seen that*

$$\int_0^T \langle \mathcal{A}_\varepsilon{}^* q, \varphi\rangle dt = \int_0^T \langle \mathcal{A}_\varepsilon \varphi, q\rangle dt \quad \forall\ \varphi, q \in D(\mathcal{A}_\varepsilon) = D(\mathcal{A}_\varepsilon{}^*) = X. \tag{3.4}$$

The operators $\mathcal{A}$ and $\mathcal{A}^*$ are defined by the same formulae (3.2) *and* (3.3), *where $y_\varepsilon = y^*$.*

LEMMA 3.3. *The operators $\mathcal{A}_\varepsilon$, $\mathcal{A}_\varepsilon{}^*$, $\mathcal{A}$, and $\mathcal{A}^*$ are closed, densely defined, and have closed ranges in $L^2(0,T;H)$. Moreover, $\dim N(\mathcal{A}_\varepsilon)$, $\dim N(\mathcal{A}_\varepsilon{}^*) \le n_0$, independent of ε, $\mathcal{A}_\varepsilon{}^*$ is the adjoint operator of $\mathcal{A}_\varepsilon$, and the following estimates hold:*

$$\|\mathcal{A}_\varepsilon{}^{-1} g\|_{L^2(0,T;D(A))\cap W^{1,2}([0,T];H)} \le C\|g\|_{L^2(0,T;H)} \quad \forall\ g \in R(\mathcal{A}_\varepsilon), \tag{3.5}$$

$$\|(\mathcal{A}_\varepsilon{}^*)^{-1} g\|_{L^2(0,T;D(A))\cap W^{1,2}([0,T];H)} \le C\|g\|_{L^2(0,T;H)} \quad \forall\ g \in R(\mathcal{A}_\varepsilon{}^*). \tag{3.6}$$

Similarly, the operator $\mathcal{A}^*$ and $\mathcal{A}$ are mutually adjoint and estimates (3.5) *and* (3.6) *remain true for $\mathcal{A}$ and $\mathcal{A}^*$. Here we have used the symbols N and R to denote the null space and the range of the corresponding operators.*

Remark. Lemma 3.3 was obtained by Barbu (cf. [3, Lemma 2]) for 2-dimensional Navier–Stokes equations. We observe that it works for 3-dimensional Navier–Stokes equations, and the proof is similar to that in [3].

Proof of Lemma 3.3. Consider the linear evolution equation

$$\begin{cases} \varphi' + \gamma A\varphi + B'(y_\varepsilon)\varphi = g \quad \text{a.e. in } (0,T), \\ \varphi(0) = x. \end{cases} \tag{3.7}$$

By (2.50) and (2.52), system (3.7) has a unique solution $\varphi \equiv \varphi_\varepsilon(t;x,g) \in W(0,T)$ for each $x \in H$, $g \in L^2(0,T;H)$, satisfying the estimate (2.56) with $\|g\|_{L^2(0,T;V^*)}$ replaced by $\|g\|_{L^2(0,T;H)}$. Moreover, if $x \in V$, then it follows that $\varphi \in Y \subset C([0,T];V)$.

In the latter case, if we multiply (3.7) by $tA\varphi(t)$, integrate on $(0,t)$, and observe that (by (1.8) and (2.48))

$$\begin{aligned} |\langle B'(y_\varepsilon)\varphi, A\varphi\rangle| &\le |b(\varphi, y_\varepsilon, A\varphi)| + |b(y_\varepsilon, \varphi, A\varphi)| \\ &\le C[\|\varphi\|^{\frac12}|A\varphi|^{\frac12}\|y_\varepsilon\| + \|y_\varepsilon\|\|\varphi\|^{\frac12}|A\varphi|^{\frac12}]|A\varphi| \\ &\le C\|\varphi\|^{\frac12}|A\varphi|^{\frac32} \\ &\le \frac{\gamma}{4}|A\varphi|^2 + C_\gamma\|\varphi\|^2, \end{aligned}$$

we get that

$$\begin{aligned} &\frac12 t\|\varphi(t)\|^2 + \gamma\int_0^t s|A\varphi(s)|^2 ds \\ &\quad\le \frac{\gamma}{4}\int_0^t s|A\varphi(s)|^2 ds + C_\gamma\int_0^T |g(t)|^2 dt + \frac{\gamma}{4}\int_0^t s|A\varphi(s)|^2 ds + C_\gamma\int_0^T \|\varphi\|^2 dt, \end{aligned}$$

which, together with (2.56), implies that

$$t\|\varphi(t)\|^2 \le C[|x|^2 + \|g\|^2_{L^2(0,T;H)}] \quad \forall\, t \in [0,T].$$

This estimate extends to all solutions to (3.7), where $x \in H$, and we have, therefore, that

$$\varphi_\varepsilon(T;x,g) \in V; \quad \|\varphi_\varepsilon(T;x,g)\|^2 \le C[|x|^2 + \|g\|^2_{L^2(0,T;H)}] \quad \forall\, \varepsilon > 0, \tag{3.8}$$

where $C > 0$ is independent of ε.

We define $G_\varepsilon : L^2(0,T;U) \to H$ by $G_\varepsilon(g) = \varphi_\varepsilon(T;0,g)$ and define $\Gamma_\varepsilon : H \to H$ by $\Gamma_\varepsilon x = \varphi_\varepsilon(T;x,0)$. It is clear that

$$\varphi_\varepsilon(T;x,g) = \Gamma_\varepsilon x + G_\varepsilon g, \tag{3.9}$$

and estimate (3.8) yields that

$$\|\Gamma_\varepsilon\|_{L(H,V)} + \|G_\varepsilon\|_{L(L^2(0,T;H),V)} \le C \quad \forall\, \varepsilon > 0. \tag{3.10}$$

Since the injection of V into H is compact, we infer that Γ_ε is completely continuous. Let $(y,g) \in \mathcal{A}_\varepsilon$; i.e., $\mathcal{A}_\varepsilon y = g$. We have, therefore, that $y(t) = \varphi_\varepsilon(t;x,g)$, where $(I - \Gamma_\varepsilon)x = G_\varepsilon g$. By the Fredholm–Riesz theory (cf. [22, Chapter X]), we know that $R(I - \Gamma_\varepsilon)$ is closed and $\dim N(I - \Gamma_\varepsilon) < \infty$. Hence $R(\mathcal{A}_\varepsilon)$ is closed in $L^2(0,T;H)$, and $N(\mathcal{A}_\varepsilon)$ is finite dimensional. Moreover, if $(\varphi_n, g_n) \in \mathcal{A}_\varepsilon$ and $\varphi_n \to \varphi$, $g_n \to g$ strongly in $L^2(0,T;H)$, then we have

$$\varphi_n' + \gamma A\varphi_n + B'(y_\varepsilon)\varphi_n = g_n \quad \text{a.e. in } (0,T). \tag{3.11}$$

Multiplying (3.11) by $t\varphi_n(t)$ and integrating on $(0,t)$, by (2.52) we get that

$$t|\varphi_n(t)|^2 + \gamma\int_0^t s\|\varphi_n(s)\|^2 ds \le C_\gamma\left[1 + \int_0^t s|\varphi_n(s)|^2 ds\right].$$

By Gronwall's inequality, we get that $t|\varphi_n(t)|^2 \le C$ for all $t \in [0,T]$. This implies that $|\varphi_n(T)| \le C$, and so $|\varphi_n(0)| \le C$. Then it follows from (3.8) that $\{\varphi_n(0)\}$ is bounded in V, and, as seen earlier, this implies that $\{\varphi_n\}$ is bounded in Y.

Hence, on a subsequence, still denoted by φ_n,

$$\varphi_n \to \varphi \ \text{ strongly in } L^2(0,T;V)\cap C([0,T];H) \ \text{ as } \varepsilon \to 0,$$

which, combined with (1.8) and Lemma 3.2, indicates that

$$\begin{aligned}\int_0^T |B'(y_\varepsilon)\varphi_n - B'(y_\varepsilon)\varphi|^2 dt &\le C\int_0^T \|\varphi_n - \varphi\|\|A\varphi_n - A\varphi\|\|y_\varepsilon\| dt\\ &\le C\left[\int_0^T \|\varphi_n - \varphi\|^2\right]^{\frac12}\left[\int_0^T \|A\varphi_n - A\varphi\|^2 dt\right]^{\frac12}\\ &\to 0 \ \text{ as } \varepsilon \to 0.\end{aligned}$$

Hence we may pass to the limit for $n \to \infty$ in (3.11) to get that $(\varphi, g) \in \mathcal{A}_\varepsilon$, i.e., $\mathcal{A}_\varepsilon$ is closed.

Now let $\Gamma \in L(H,H)$ be defined by $\Gamma x = \varphi(T;x,0)$, where φ is the solution to

$$\begin{cases}\varphi' + \gamma A\varphi + B'(y^*)\varphi = g \quad \text{a.e. in } (0,T),\\ \varphi(0) = x.\end{cases}$$

As seen earlier, $\Gamma \in L(H,V)$, and so Γ is completely continuous from H into itself. Moreover, by Lemma 3.2 and the estimate (3.10), it follows that

$$\Gamma_\varepsilon \to \Gamma \ \text{ in } L(H,H) \ \text{ as } \varepsilon \to 0.$$

Since $\dim N(I-\Gamma) < \infty$, the latter implies that there exists $n_0 > 0$ such that $\dim N(I-\Gamma_\varepsilon) \le n_0$ for all $\varepsilon > 0$. Hence $\dim N(\mathcal{A}_\varepsilon) \le n_0$ for all $\varepsilon > 0$ as claimed. Moreover, we have that

$$|(I-\Gamma_\varepsilon)^{-1}g_0| \le C|g_0| \quad \forall\ g_0 \in R(I-\Gamma_\varepsilon). \tag{3.12}$$

Indeed, otherwise there exist $x_\varepsilon \in R((I-\Gamma_\varepsilon)^*)$, $f_\varepsilon \in R(I-\Gamma_\varepsilon)$ such that $(I-\Gamma_\varepsilon)x_\varepsilon = f_\varepsilon$ and $|f_\varepsilon| = 1$, $|x_\varepsilon| \to \infty$. Let $\tilde{x}_\varepsilon = \frac{x_\varepsilon}{|x_\varepsilon|}$ and $\tilde{f}_\varepsilon = \frac{f_\varepsilon}{|x_\varepsilon|}$. Then $\tilde{f}_\varepsilon \to 0$ and $(I-\Gamma_\varepsilon)\tilde{x}_\varepsilon = \tilde{f}_\varepsilon$. We have that $\Gamma_\varepsilon\tilde{x}_\varepsilon = \varphi(T;\tilde{x}_\varepsilon,0)$. It follows from (3.8) that $\|\Gamma_\varepsilon\tilde{x}_\varepsilon\| \le C|\tilde{x}_\varepsilon| \le C$. This implies that $\{\Gamma_\varepsilon\tilde{x}_\varepsilon\}$ has a subsequence which converges in H. Since $\tilde{x}_\varepsilon = \Gamma_\varepsilon\tilde{x}_\varepsilon + \tilde{f}_\varepsilon$, we infer that there exists a subsequence of $\{\tilde{x}_\varepsilon\}$, still denoted by itself, such that $\tilde{x_\varepsilon} \to x_0$ in H and $|x_0| = 1$. We have that $x_0 \in R((I-\Gamma)^*)$ and $(I-\Gamma)x_0 = 0$, which contradicts the fact that

$$R((I-\Gamma)^*) \oplus N(I-\Gamma) = H.$$

Recall that $\varphi = \varphi_\varepsilon(t;x,g)$, where $(I-\Gamma_\varepsilon)x = G_\varepsilon g$ is a solution to $\mathcal{A}_\varepsilon\varphi_\varepsilon = g$. It follows from (3.12) that

$$|\varphi_\varepsilon(0)| \le |(I-\Gamma_\varepsilon)^{-1}(G_\varepsilon g)| \le C|G_\varepsilon g|. \tag{3.13}$$

Then, as seen above, we have that

$$\|\varphi_\varepsilon\|_{W^{1,2}([0,T];H)} + \|\varphi_\varepsilon\|_{L^2(0,T;D(A))} \le C\|g\|_{L^2(0,T;H)} \quad \forall\, g \in R(\mathcal{A}_\varepsilon),$$

which implies (3.5).

The corresponding properties of the operator ${\mathcal{A}_\varepsilon}^*$ follow from the previous arguments because, in this case, (3.7) is replaced by

$$\varphi' + \gamma A\varphi + [B'(y_\varepsilon)]^*\varphi = g, \quad \varphi(0) = x,$$

and so the previous estimates remain valid. In particular, it follows that the operator ${\mathcal{A}_\varepsilon}^*$ is closed, and so we conclude from (3.4) that its adjoint is precisely $\mathcal{A}_\varepsilon$.

Similarly, we may prove the corresponding results for the operators $\mathcal{A}$ and $\mathcal{A}^*$. This completes the proof.

THEOREM 3.4. *Let (H_1)–(H_4) and (H_8) hold. Suppose that (y^*, u^*) is optimal for problem (P_3). Then there exists $p \in X$ such that*

$$p' - \gamma Ap - [B'(y^*)]^*p \in \partial g(t, y^*) \quad \text{a.e. in } (0,T), \tag{3.14}$$

and

$$D^*p(t) \in \partial h(u^*(t)) \quad \text{a.e. in } (0,T). \tag{3.15}$$

Proof. Let $(y_\varepsilon, u_\varepsilon)$ be optimal for $(P_{3\varepsilon})$. For any $z \in X$, $v \in L^2(0,T;U)$ fixed, we set $y_\varepsilon^\rho = y_\varepsilon + \rho z$ and $u_\varepsilon^\rho = u_\varepsilon + \rho v$. Then $(y_\varepsilon^\rho, u_\varepsilon^\rho) \in X \times L^2(0,T;U)$. By the same argument as in the proof of Theorem 2.3, we have that

$$\begin{aligned} 0 \le &\int_0^T [\langle \nabla g^\varepsilon(t, y_\varepsilon), z\rangle + \langle \nabla h_\varepsilon(u_\varepsilon), v\rangle_U]dt \\ &+ \int_0^T \langle \|y_\varepsilon - y^*\|^2 A(y_\varepsilon - y^*), z\rangle dt \\ &+ \int_0^T \langle u_\varepsilon - u^*, v\rangle_U dt + \int_0^T \langle p_\varepsilon, z' + \gamma Az + B'(y_\varepsilon)z - Dv\rangle dt \end{aligned} \tag{3.16}$$

for all $z \in X$ and $v \in L^2(0,T;U)$, where $p_\varepsilon = \frac{1}{\varepsilon}[y_\varepsilon' + \gamma Ay_\varepsilon + By_\varepsilon - Du_\varepsilon - f]$.

By taking $v = 0$ in (3.17), we get that

$$\begin{aligned} &\int_0^T \langle \nabla g^\varepsilon(t, y_\varepsilon) + \|y_\varepsilon - y^*\|^2 A(y_\varepsilon - y^*), z\rangle dt \\ &\qquad + \int_0^T \langle p_\varepsilon, \mathcal{A}_\varepsilon z\rangle dt = 0 \quad \forall\, z \in Z. \end{aligned} \tag{3.17}$$

Hence $p_\varepsilon \in D({\mathcal{A}_\varepsilon}^*) = X$, and

$${\mathcal{A}_\varepsilon}^* p_\varepsilon = -\nabla g^\varepsilon(t, y_\varepsilon) - \|y_\varepsilon - y^*\|^2 A(y_\varepsilon - y^*). \tag{3.18}$$

By taking $z = 0$ in (3.16), we deduce that

$$\int_0^T [\langle \nabla h_\varepsilon(u_\varepsilon), v\rangle_U - \langle D^*p_\varepsilon, v\rangle_U + \langle u_\varepsilon - u^*, v\rangle_U]dt \ge 0 \quad \forall\, v \in L^2(0,T;U).$$

This yields that

$$D^*p_\varepsilon(t) = \nabla h_\varepsilon(u_\varepsilon(t)) + u_\varepsilon(t) - u^*(t) \quad \text{a.e. in } (0,T). \tag{3.19}$$

By (3.19) and Lemma 3.2 and using the same argument as in [2, Chapter 1], we have

$$\|D^*p_\varepsilon\|_{L^2(0,T;H)} \le C \quad \forall\, \varepsilon > 0. \tag{3.20}$$

Now, by Lemma 3.3 and the closed range theorem (cf. [20, p. 205]), we may write

$$p_\varepsilon = p_\varepsilon^1 + p_\varepsilon^2,$$

where $p_\varepsilon^1 \in R(\mathcal{A}_\varepsilon)$ and $p_\varepsilon^2 \in N(\mathcal{A}_\varepsilon{}^*)$. By Lemma 3.3 again, we get that

$$\|p_\varepsilon^1\|_{L^2(0,T;D(A))\cap W^{1,2}([0,T];H)} \le C \quad \forall\, \varepsilon > 0. \tag{3.21}$$

On the other hand, since the space $N(\mathcal{A}_\varepsilon{}^*)$ is finite dimensional, we infer that the restriction of D to $N(\mathcal{A}_\varepsilon{}^*)$, still denoted by D, has closed range. Then by the closed range theorem again, we deduce that

$$p_\varepsilon^2 = p_\varepsilon^3 + p_\varepsilon^4,$$

where $p_\varepsilon^3 \in R(D)$ and $p_\varepsilon^4 = N(D^*)$. Then, by (3.20), we see that $\{p_\varepsilon^3\}$ is bounded in $L^2(0,T;H)$. Since $\{p_\varepsilon^3\} \subset N(\mathcal{A}_\varepsilon{}^*)$ and $\dim(\mathcal{A}_\varepsilon{}^*) \le n_0$, there exist $p^3 \in L^2(0,T;H)$ and a subsequence of p_ε^3, still denoted by itself, such that

$$p_\varepsilon^3 \to p^3 \ \text{ strongly in } L^2(0,T;H) \ \text{ as } \varepsilon \to 0. \tag{3.22}$$

By (3.21), we may assume that (without loss of generality)

$$p_\varepsilon^1 \to p^1 \ \text{ strongly in } L^2(0,T;H) \ \text{ as } \varepsilon \to 0. \tag{3.23}$$

By Lemma 3.2, one can easily check that

$$\mathcal{A}_\varepsilon z \to \mathcal{A}z \ \text{ weakly in } L^2(0,T;H) \quad \forall\, z \in X. \tag{3.24}$$

Since $p_\varepsilon^2 = p_\varepsilon^3 + p_\varepsilon^4 \in N(\mathcal{A}_\varepsilon)$, we may rewrite (3.18) as

$$\mathcal{A}_\varepsilon{}^*(p_\varepsilon^1 + p_\varepsilon^3) = -\nabla g^\varepsilon(t, y_\varepsilon) - \|y_\varepsilon - y^*\|^2 A(y_\varepsilon - y^*),$$

which is equivalent to

$$\begin{aligned} &\int_0^T \langle \nabla g^\varepsilon(t, y_\varepsilon) + \|y_\varepsilon - y^*\|^2 A(y_\varepsilon - y^*), z\rangle dt \\ &\qquad + \int_0^T \langle p_\varepsilon^1 + p_\varepsilon^3, \mathcal{A}_\varepsilon z\rangle dt = 0 \end{aligned} \tag{3.25}$$

for all $z \in X$.

By Lemma 3.2 and by the same argument as in [1, Chapter 5], we have that

$$\nabla g^\varepsilon(t, y_\varepsilon) \to \beta \ \text{ weakly in } \ L^2(0,T;H) \ \text{ and } \beta \in \partial g(t, y^*) \quad \text{a.e. in } (0,T). \tag{3.26}$$

Now by (3.22), (3.23), (3.24), (3.26), and Lemma 3.2, we may pass to the limit for $\varepsilon \to 0$ in (3.25) to get

$$\int_0^T \langle \beta, z\rangle dt + \int_0^T \langle p^1 + p^3, \mathcal{A}z\rangle dt = 0 \quad \forall\ z \in Z.$$

This shows that $p^1 + p^3 \in D(\mathcal{A}^*)$ and

$$\mathcal{A}(p^1 + p^3) = \beta. \tag{3.27}$$

By (3.22), (3.23), (3.26), and Lemma 3.2, we may pass to the limit for $\varepsilon \to 0$ in (3.19) to get that

$$D^*(p^1 + p^3) \in \partial h(u^*) \quad \text{a.e. in } (0,T). \tag{3.28}$$

Let $p = p^1 + p^3 \in X$. Then $p \in X$. By (3.27) and (3.28), we derive (3.14) and (3.15). This completes the proof.

4. Applications. In this section, we shall point out some special cases of cost functionals and state constraints covered by (1.9), (1.13), and (1.14).

Example 4.1. Let $g(t,y) = \frac{1}{2}[|y - y^0(t)|^2 + |\nabla \times y|^2]$, where $y^0 \in L^\infty(0,T;H)$ stands for reference velocity and $\nabla \times y = \operatorname{curl} y$, and let $h(u) = \frac{1}{2}|u|_U^2$.

One can check that $g : [0,T] \times V \to R^+$ and $h : U \to R$ satisfy all conditions in (H_5). In this case, our objective is to determine the control u in such a way that the velocity vector is as close as possible, in the sense of (P_1) (or (P_2)), to the desired velocity y^0, and the turbulence is minimal.

Example 4.2. In (H_6), we take $X = L^2(0,T;V)$, $F \equiv I$, the identity operator on $L^2(0,T;V)$, and $W = \{y \in L^2(0,T;V) : \|y\|_{L^2(0,T;V)} \le \rho\}$.

It is clear that W is closed and convex and has finite codimensionality in $L^2(0,T;V)$. In this case, (1.13) is equivalent to $\int_0^T \|y\|^2 dt \le \rho$, which is equivalent to (cf. [5]) $\int_0^T |\nabla \times y|^2 dt \le \rho^2$. Thus the state constraint (1.13) in this case means that the average of the turbulence in $[0,T]$ is governed by ρ^2.

Example 4.3. In (H_6), we may take $X = R$, $F(y) = \int_0^T |y|^2 dt$, and $W = (-\infty, \rho^2]$.

It is clear that W is convex and closed in $X = R$ with the codimension zero (cf. [6] and [10]) and $F'(y)$ injective for all $y \in L^2(0,T;V)$. In this case, (1.13) is equivalent to $\int_0^T |y|^2 dt \le \rho^2$. Thus the state constraint (1.13) means that the average of the energy in $[0,T]$ is governed by ρ^2.

Example 4.4. In (H_6), we may take $X = R$, $F(y) = \int_0^T [|\nabla \times y|^2 - c|y|^2] dt$, where $c \ge 0$ and $W = (-\infty, \rho^2]$.

One can easily check that F is class of C^1 and $F'(y)$ is injective (for c small enough), and W is closed and convex with the codimension zero. In this case, (1.13) is equivalent to $\int_0^T [|\nabla \times y|^2 - c|y|^2] dt \le \rho^2$ or $\int_0^T |\nabla \times y|^2 dt \le c\int_0^T |y|^2 dt + \rho^2$. Thus the state constraint (1.13) in this case means that the average of the turbulence is governed by the average of the energy and ρ^2.

Recall that the enstrophy set (cf. [6] and [10]) $K = \{y_0 \in V : |\nabla \times y_0|^2 \le \varphi(|y_0|^2) + \rho^2\}$ plays an important role in fluid mechanics. We may regard $\tilde{K}$ as a special generalized enstrophy set, where $\tilde{K} = \{y \in L^2(0,T;V) : \int_0^T |\nabla \times y|^2 dt \le c\int_0^T |y|^2 dt + \rho^2\}$. Thus (1.13) in this case means that $y \in \tilde{K}$.

Example 4.5. In (H_6), we may take $X = R$, $W = (-\infty, \rho^2]$, and $F(y) = \int_0^T \langle y, \nabla \times y\rangle^2 dt + \lambda^2 \int_0^T \|y\|^2 dt$.

One can easily check that F is class of C^1 and $F'(y)$ is injective. In this case, (1.13) is equivalent to $\int_0^T \langle y, \nabla \times y\rangle^2 dt + \lambda^2 \int_0^T \|y\|^2 dt \le \rho^2$.

Recall that the helicity set $K_1 = \{y_0 \in V : \langle y_0, \nabla \times y_0\rangle^2 + \lambda^2 \|y_0\|^2 \le \rho^2\}$ plays an important role in fluid mechanics, and, in particular, it is an invariant set of Euler's equation (cf. [9]). We may regard $\tilde{K_1}$ as a generalized helicity set, where

$$\tilde{K_1} = \left\{ y \in L^2(0,T;V) : \int_0^T \langle y, \nabla \times y\rangle^2 dt + \lambda^2 \int_0^T \|y\|^2 dt \le \rho^2 \right\}.$$

The state constraint (1.13) in this case means that $y \in \tilde{K_1}$.

Example 4.6. In (H_7), let $S = \{y_0\} \times S_1$, where $y_0 \in V$ and $S_1 \subset H$ is convex and closed. Furthermore, we assume the following.

$(H_{10,1})$ S_1 has finite codimensionality in H.

We claim that $(H_{10,1})$ implies (H_{10}). To this end, let $\Gamma : H \to H$ and $G : L^2(0,T;U) \to H$ be defined by $\Gamma x = \varphi(T;x,0)$ and $Gu = \varphi(T;0,u)$, where $\varphi(t;x,u)$ is the solution to

$$\begin{cases} \varphi' + \gamma A\varphi + B'(y^*)\varphi = Du \quad \text{a.e. in } (0,T), \\ \varphi(0) = x. \end{cases} \tag{4.1}$$

Notice that (4.1) has a unique solution $\varphi(t;x,u) \in W(0,T)$ for each $x \in H$ and $u \in L^2(0,T;U)$. Moreover, $\varphi(t;x,u) \in Y$ if $x \in V$ and $u \in L^2(0,T;U)$.

Let $M(0,r)$ be defined by (2.58), let Q_{r_1,r_2} be defined in (3.2), and let $E(0,r)$ be defined by $E(0,r) = \{z_0 \in V : \|z_0\| \le r\}$. Consider the set $\{(z_0, \Gamma z_0 + Gu) : z_0 \in E(0,r), u \in M(0,r)\}$. It is clear that $E(0,r)$ has finite codimensionality in H for each $r > 0$. Thus by $(H_{10,1})$ and by [14, Proposition 4], we deduce that $\{(z_0, \Gamma z_0 + Gu) : z_0 \in E(0,r), u \in M(0,r)\} - \{y_0\} \times S_1$ has finite codimensionality in $H \times H$.

However, for all $z_0 \in E(0,r)$ and $u \in M(0,r)$, $\Gamma z_0 + Gu = \varphi(T;z_0,u) \in V$ and $|\varphi(t;z_0,u| \le r_0$ for some $r_0 > 0$ independent of z_0 and u. Hence $Q_{r_0,r} \supset \{(z_0, \Gamma z_0 + Gu) : z_0 \in E(0,r), u \in M(0,r)\}$. So $Q_{r_0,r} - S$ has finite codimensionality. This means $(H_{10,1})$ implies (H_{10}).

Example 4.7. Let Γ, G, $M(0,r)$, and $E(0,r)$ be defined in Example 4.6.

Let $Q = \{x_1 - \Gamma x_0 : (x_0, x_1) \in S\}$, $\mathcal{R}_r = GM(0,r) = \{Gu : u \in M(0,r)\}$, and $\tilde{\mathcal{R}}_r = \{(z_0, z_T) \in H \times H : z_0 \in E(0,r), z_T = \Gamma z_0 + Gu, u \in M(0,r)\}$. We assume the following.

$(H_{10,2})$ $\mathcal{R}_r - Q$ has finite codimensionality in H.

We recall the following proposition; for its proof, we refer the reader to [20, Lemma 2.5].

PROPOSITION 4.8. *$\mathcal{R}_r - Q$ is finite codimensional in H if and only if $\tilde{\mathcal{R}}_r - S$ is so in $H \times H$.*

By the same argument as in Example 4.6, we have that $\tilde{\mathcal{R}}_r \subset Q_{r_0,r}$ for some $r_0 > 0$. Thus $(H_{10,2})$ implies (H_{10}).

Next we shall show that Theorem 2.4 works for the periodic case, i.e., the case where $S = \{(x,x,) : x \in H\}$ without assumption (H_{10}). Indeed, in this case, $Q = \{(I - \Gamma)x : x \in H\}$. By the previous discussion (cf. the proof of Lemma 3.3 in section 3), $R(I-\Gamma)^*$ is closed, $\dim N(I-\Gamma) < \infty$, and $H = R((I-\Gamma)^*) \oplus N(I-\Gamma)$. Thus Q has finite codimensionality in H. So does $\mathcal{R}_r - Q$ (cf. [14, Proposition 4 of Chapter 4]). Then it follows from Proposition 4.8 that $Q_{r_0,r}$ has finite codimensionality.

Notice that, in this case, we cannot get $\lambda_0 = 0$; i.e., the maximum principle we obtain is not qualified.

REFERENCES

[1] V. Barbu, *Optimal Control of Variational Inequalities*, Pitman Res. Notes Math. Ser. 100, Pitman, Boston, 1984.

[2] V. Barbu, *Analysis and Control of Nonlinear Infinite Dimensional Systems*, Academic Press, Boston, 1993.

[3] V. Barbu, *Optimal control of Navier-Stokes equations with periodic inputs*, Nonlinear Anal., 31 (1998), pp. 15–31.

[4] V. Barbu, *The time optimal control of Navier-Stokes equations*, Systems Control Lett., 30 (1997), pp. 93–100.

[5] V. Barbu and N. Pavel, *Flow-invariance closed set with respect to nonlinear semigroup flows*, J. Math. Anal. Appl., to appear.

[6] V. Barbu and S. Sritharan, *Flow-invariance preserving feedback controllers for Navier-Stokes equations*, J. Math. Anal. Appl., 255 (2001), pp. 281–307.

[7] P. Constantin and C. Foias, *Navier-Stokes Equations*, The University of Chicago Press, Chicago, 1998.

[8] H. O. Fattorini and S. S. Sritharan, *Necessary and sufficient conditions for optimal controls in viscous flow problems*, Proc. Royal Soc. Edinburgh Sect. A, 124 (1994), pp. 211–251.

[9] H. O. Fattorini and S. S. Sritharan, *Optimal control problems with state constraints in fluid mechanics and combustion*, Appl. Math. Optim., 38 (1998), pp. 159–192.

[10] V. I. Arnold and B. A. Khesin, *Topological Methods in Hydrodynamics*, Springer-Verlag, New York, 1998.

[11] A. V. Fursikov, *Control problems and theorems concerning unique solvability of a mixed boundary value problem for three-dimensional Navier–Stokes and Euler equations*, Mat. Sb., 115 (1981), pp. 281–306, 320 (in Russian).

[12] A. V. Fursikov, *Optimal control problems for Navier–Stokes system with distributed control function*, in Optimal Control of Viscous Flow VI, SIAM, Philadelphia, 1998, pp. 109–150.

[13] K. Ito and S. Kang, *A dissipative feedback control synthesis for systems arising in fluid dynamics*, SIAM J. Control Optim., 32 (1994), pp. 831–854.

[14] J.-L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, Heidelberg, New York, 1971.

[15] J.-L. Lions and E. Magenes, *Non-Homogeneous Boundary Valued Problems and Application*, Springer-Verlag, Berlin, Heidelberg, New York, 1972.

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

[17] X. Li and J. Yong, *Necessary conditions for optimal control of distributed parameter systems*, SIAM J. Control Optim., 29 (1991), pp. 895–908.

[18] R. Teman, *Navier-Stokes Equations*, North–Holland, Amsterdam, 1979.

[19] G. S. Wang, *Optimal control of parabolic differential equations with two point boundary state constraints*, SIAM J. Control Optim., 38 (2000), pp. 1639–1654.

[20] G. S. Wang and S. R. Chen, *Maximum principle for optimal control of some parabolic systems with two point boundary conditions*, Numer. Funct. Anal. Optim., 20 (1999), pp. 163–174.

[21] G. Wang, Y. Zhao, and W. Li, *Some optimal control problems governed by elliptic variational inequalities with control and state constraint on the boundary*, J. Optim. Theory Appl., 106 (2000), pp. 627–655.

[22] K. Yosuda, *Functional Analysis*, 5th ed., Springer-Verlag, Berlin, Heidelberg, New York, 1978.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 2, pp. 607–622

ON THE BOUNDARY CONTROL OF SYSTEMS OF CONSERVATION LAWS*

ALBERTO BRESSAN† AND GIUSEPPE MARIA COCLITE†

Abstract. This paper is concerned with the boundary controllability of entropy weak solutions to hyperbolic systems of conservation laws. We prove a general result on the asymptotic stabilization of a system near a constant state. On the other hand, we give an example showing that exact controllability in finite time cannot be achieved, in general.

Key words. boundary control, hyperbolic system, conservation laws

AMS subject classifications. 35L65, 93C20

PII. S0363012901392529

1. Introduction.

Consider an $n \times n$ system of conservation laws on a bounded interval:

$$u_t + f(u)_x = 0, \qquad t \geq 0, \quad x \in]a, b[. \tag{1.1}$$

The system is assumed to be strictly hyperbolic, each characteristic field being either linearly degenerate or genuinely nonlinear in the sense of Lax [8]. We shall also assume that all characteristic speeds are bounded away from zero. More precisely, let $f : \Omega \mapsto \mathbb{R}^n$ be a smooth map defined on an open set $\Omega \subseteq \mathbb{R}^n$. For each $u \in \Omega$, call $\lambda_1(u) < \cdots < \lambda_n(u)$ the eigenvalues of the Jacobian matrix $Df(u)$. We assume that there exist a minimum speed $c_0 > 0$ and an integer $p \in \{1, \ldots, n\}$ such that

$$\begin{cases} \lambda_i(u) < 0 & \text{if} \quad i \leq p, \\ \lambda_i(u) > 0 & \text{if} \quad i > p, \end{cases} \tag{1.2}$$

$$\big|\lambda_i(u)\big| \geq c_0 > 0, \qquad u \in \Omega. \tag{1.3}$$

By (1.2), for a solution defined on the strip $t \geq 0,\ x \in]a, b[$, there will be $n - p$ characteristics entering at the boundary point $x = a$ and p characteristics entering at $x = b$. The initial-boundary value problem is thus well posed if we prescribe $n - p$ scalar conditions at $x = a$ and p scalar conditions at $x = b$ [11]. See also [1, 2] for the case of general entropy weak solutions taking values in the space BV of functions with bounded variation.

In the present paper, we study the effect of boundary conditions on the solution of (1.1) from the point of view of control theory. Namely, given an initial condition

$$u(0, x) = \phi(x), \qquad x \in]a, b[, \tag{1.4}$$

with small total variation, we regard the boundary data as *control functions*, and we study the family of configurations

$$\mathcal{R}(T) \doteq \big\{u(T, \cdot)\big\} \subset \mathbf{L}^1\big([a, b]\,;\, \mathbb{R}^n\big), \tag{1.5}$$

which can be reached by the system at a given time $T > 0$.

*Received by the editors July 13, 2001; accepted for publication (in revised form) December 19, 2001; published electronically July 1, 2002.
http://www.siam.org/journals/sicon/41-2/39252.html
†S.I.S.S.A., Via Beirut 4, Trieste 34014, Italy (bressan@sissa.it, coclite@sissa.it).

Beginning with the simplest case, consider a strictly hyperbolic system with constant coefficients:

$$u_t + Au_x = 0, \tag{1.6}$$

where A is an $n \times n$ constant matrix, with real distinct eigenvalues

$$\lambda_1 < \cdots < \lambda_p < 0 < \lambda_{p+1} < \cdots < \lambda_n.$$

Call

$$\tau \doteq \max_i \frac{b-a}{|\lambda_i|}$$

the maximum time taken by waves to cross the interval $[a, b]$. In this case, it is easy to see that the reachable set in (1.5) is the entire space: $\mathcal{R}(T) = \mathbf{L}^1$ for all $T \geq \tau$. In other words, the system is completely controllable after time τ. Indeed, for any $T \geq \tau$ and initial and terminal data $\phi, \psi \in \mathbf{L}^1\big([a, b]; \mathbb{R}^n\big)$, one can always find a solution of (1.4), defined on the rectangle $[0, T] \times [a, b]$, such that

$$u(0, x) = \phi(x), \qquad u(T, x) = \psi(x), \qquad x \in [a, b].$$

Such a solution can be constructed as follows. Let $l_1, \ldots, l_n$ and $r_1, \ldots, r_n$ be dual bases of right and left eigenvectors of A so that $l_i \cdot r_j = \delta_{ij}$. For $i = 1, \ldots, n$, let $u_i(t, x)$ be a solution to the scalar Cauchy problem

$$u_{i,t} + \lambda_i u_{i,x} = 0,$$

$$u_i(0, x) = \begin{cases} l_i \cdot \phi(x) & \text{if} \quad x \in [a, b], \\ l_i \cdot \psi(x + \lambda_i T) & \text{if} \quad x \in [a - \lambda_i T,\ b - \lambda_i T], \\ 0 & \text{otherwise.} \end{cases}$$

Then the restriction of

$$u(t, x) = \sum_i u_i(t, x) r_i$$

to the interval $[0, T] \times [a, b]$ satisfies (1.6) and takes the required initial and terminal values. Of course, this corresponds to the solution of an initial-boundary value problem, determined by the n boundary conditions

$$\begin{cases} l_i \cdot u(t, a) = u_i(t, a), & i = p + 1, \ldots, n, \\ l_i \cdot u(t, b) = u_i(t, b), & i = 1, \ldots, p. \end{cases}$$

This result on exact boundary controllability has been extended in [9, 10] to the case of general quasi-linear systems of the form

$$u_t + A(u)u_x = 0.$$

In this case, the existence of a solution taking the prescribed initial and terminal values is obtained for all sufficiently small data $\phi, \psi \in \mathcal{C}^1$.

The aim of the present paper is to study analogous controllability properties within the context of entropy weak solutions $t \mapsto u(t, \cdot) \in BV$. For the definitions

and basic properties of weak solutions, we refer to [4]. For general nonlinear systems, it is clear that a complete controllability result within the space BV cannot hold. Indeed, already for a scalar conservation law, it was proved in [3] that the profiles $\psi \in BV$ which can be attained at a fixed time $T > 0$ are only those which satisfy the Oleinik-type conditions

$$\psi'(x) \leq \frac{f'\big(\psi(x)\big)}{(x-a)f''\big(\psi(x)\big)} \qquad \text{for almost every } x \in [a,b].$$

For general $n \times n$ systems, a complete characterization of the reachable set $\mathcal{R}(T)$ does not seem possible, due to the complexity of repeated wave-front interactions.

Our first result is concerned with stabilization near a constant state. Assuming that all characteristic speeds are bounded away from zero, we show that the system can be asymptotically stabilized to any state $u^* \in \Omega$, with quadratic rate of convergence.

THEOREM 1. *Let K be a compact, connected subset of the open domain $\Omega \subset \mathbb{R}^n$. Then there exist constants $C_0, \delta, \kappa > 0$ such that the following holds. For every constant state $u^* \in K$ and every initial data $u(0) = \phi : [a,b] \mapsto K$ with* Tot.Var.$\{\phi\} < \delta$, *there exists an entropy weak solution $u = u(t,x)$ of (1.1) such that, for all $t > 0$,*

$$\text{Tot.Var.}\big\{u(t)\big\} \leq C_0\, e^{-2^{\kappa t}}, \tag{1.7}$$

$$\big\|u(t,x) - u^*\big\|_{L^\infty} \leq C_0\, e^{-2^{\kappa t}}. \tag{1.8}$$

The proof will be given in section 2. An interesting question is whether the constant state u^* can be exactly reached in a finite time T. By the results in [9], this is indeed the case if the initial data has a small $\mathcal{C}^1$ norm. On the contrary, in the final part of this paper, we show that exact controllability in finite time cannot be attained, in general, if the initial data is only assumed to have small total variation.

Our counterexample is concerned with a class of strictly hyperbolic, genuinely nonlinear 2×2 systems of the form (1.1). More precisely, we assume the following.

(H) The eigenvalues $\lambda_i(u)$ of the Jacobian matrix $A(u) = Df(u)$ satisfy

$$-\lambda^* < \lambda_1(u) < -\lambda_* \ < 0 \ < \lambda_* < \lambda_2(u) < \lambda^*. \tag{1.9}$$

Moreover, the right eigenvectors $r_1(u)$, $r_2(u)$ satisfy the inequalities

$$D\lambda_1 \cdot r_1 > 0, \qquad D\lambda_2 \cdot r_2 > 0, \tag{1.10}$$

$$r_1 \wedge r_2 < 0, \qquad r_1 \wedge (Dr_1 \cdot r_1) < 0, \qquad r_2 \wedge (Dr_2 \cdot r_2) < 0. \tag{1.11}$$

Here $D\lambda_i$, Dr_i denote the differentials of the functions $\lambda_i(u)$, $r_i(u)$, while $\wedge$ is the wedge product: if $v = (v_1, v_2)$, $w = (w_1, w_2)$, we define

$$v \wedge w \doteq v_1 w_2 - v_2 w_1.$$

A particular system which satisfies the above assumptions is the one studied by DiPerna [7]:

$$\begin{cases} \rho_t + (u\rho)_x = 0, \\ u_t + \left(\dfrac{u^2}{2} + \dfrac{K^2}{\gamma - 1}\rho^{\gamma-1}\right)_x = 0, \end{cases}$$

with $1 < \gamma < 3$. Here $\rho > 0$ and u denote the density and the velocity of a gas, respectively.

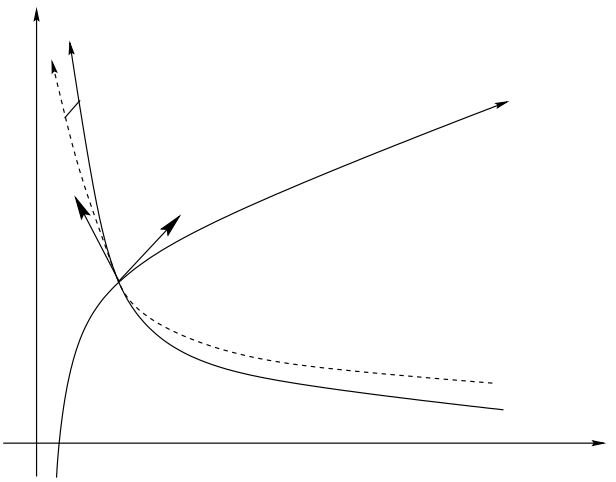


FIG. 1.

The last two inequalities in (1.11) imply that the rarefaction curves (i.e., the integral curves of the vector fields (r_1, r_2) in the (u_1, u_2) plane) turn clockwise (Figure 1). In such a case, the interaction of two shocks of the same family generates a shock in the other family.

THEOREM 2. *Consider a* 2×2 *system satisfying the assumption* (H). *Then there exist initial data* $\phi : [a, b] \mapsto \mathbb{R}^2$ *having arbitrarily small total bounded variation for which the following holds. For every entropy weak solution* u *of* (1.1), (1.4), *with* Tot.Var.$\{u(t, \cdot)\}$ *remaining small for all* t, *the set of shocks in* $u(t, \cdot)$ *is dense on* $[a, b]$ *for each* $t > 0$. *In particular,* $u(t, \cdot)$ *cannot be constant.*

As a preliminary, in section 3, we establish an Oleinik-type estimate on the decay of positive waves. This bound is of independent interest and sharpens the results in [5] for systems satisfying the additional conditions (H).

As a consequence, this implies that positive waves are "weak" and cannot completely cancel a shock within finite time. The proof of Theorem 2 is then achieved by an induction argument. We show that, if the set of 1-shocks is dense on $[0, T] \times [a, b]$, then the set of points $P_j = (t_j, x_j)$, where two 1-shocks interact and create a new 2-shock, is also dense on the same domain. Therefore, new shocks are constantly generated, and the solution can never be reduced to a constant. Details of the proof will be given in section 4.

As in [9], all of the above results refer to the case where total control on the boundary values is available. As a consequence, the problem is reduced to proving the existence (or nonexistence) of an entropy weak solution defined on the open strip $t > 0$, $x \in]a, b[$, satisfying the required conditions. This is a first step toward the analysis of more general controllability problems, where the control acts only on some of the boundary conditions. We thus leave open the case where a subset of indices $I \subset \{1, \dots, n\}$ is given, and one requires

$$l_i \cdot u(t, a) = \begin{cases} \alpha_i(t) & \text{if} \quad i \in I, \\ 0 & \text{if} \quad i \notin I, \end{cases} \qquad i = p+1, \dots, n,$$

$$l_i \cdot u(t, b) = \begin{cases} \alpha_i(t) & \text{if} \quad i \in I, \\ 0 & \text{if} \quad i \notin I, \end{cases} \qquad i = 1, \dots, p,$$

for some control functions α_i acting only on the components $i \in I$.

Throughout the following, we denote by $r_i(u), l_i(u)$ the right and left i-eigenvectors of the Jacobian matrix $A(u) \doteq Df(u)$. As in [4], we write $\sigma \mapsto R_i(\sigma)(u_0)$ for the parametrized i-rarefaction curve through the state u_0 so that

$$\frac{d}{d\sigma} R_i(\sigma) = r_i\big(R_i(\sigma)\big), \qquad R_i(0) = u_0.$$

The i-shock curve through u_0 is denoted by $\sigma \mapsto S_i(\sigma)(u_0)$. It satisfies the Rankine–Hugoniot equations

$$f\big(S_i(\sigma)\big) - f(u_0) = \lambda_i(\sigma)\,\big(S_i(\sigma) - u_0\big)$$

for some shock speed λ_i. We recall (see [4, Chap. 5]) that the general Riemann problem is solved in terms of the composite curves

$$\Psi_i(u_0)(\sigma) = \begin{cases} R_i(u_0)(\sigma) & \text{if } \ \sigma \geq 0, \\ S_i(u_0)(\sigma) & \text{if } \ \sigma < 0. \end{cases} \tag{1.12}$$

2. Proof of Theorem 1. The proof relies on the following two lemmas.

LEMMA 1. *In the setting of Theorem* 1*, there exists a time* $T > 0$ *such that the following holds. For every pair of states* $\omega, \omega' \in K$*, there exists an entropic solution* $u = u(t,x)$ *of* (1.1) *such that*

$$u(0,x) \equiv \omega, \qquad u(T,x) \equiv \omega' \qquad \textit{for all } x \in [a,b]. \tag{2.1}$$

Proof. Consider the function

$$\Phi(\sigma_1, \ldots, \sigma_n\,;\ v,\, v') \doteq \Psi_n(\sigma_n) \circ \cdots \circ \Psi_{p+1}(\sigma_{p+1})(v') - \Psi_p(\sigma_p) \circ \cdots \circ \Psi_1(\sigma_1)(v). \tag{2.2}$$

Observe that, whenever $v = v'$, the $n \times n$ Jacobian matrix $\partial\Phi/\partial\sigma_1 \cdots \sigma_n$ computed at $\sigma_1 = \sigma_2 = \cdots = \sigma_n = 0$ has full rank. Indeed, the columns of this matrix are given by the linearly independent vectors $-r_1(v), \ldots, -r_p(v)$, $r_{p+1}(v), \ldots, r_n(v)$. By the implicit function theorem and a compactness argument, we can find $\delta > 0$ such that the following holds. For every $v, v' \in K$, with $|v - v'| \leq \delta$, there exist unique values $\sigma_1, \ldots, \sigma_n$ such that

$$v'' \doteq \Psi_n(\sigma_n) \circ \cdots \circ \Psi_{p+1}(\sigma_{p+1})(v') = \Psi_p(\sigma_p) \circ \cdots \circ \Psi_1(\sigma_1)(v). \tag{2.3}$$

Defining the time as

$$\tau \doteq \max_{1 \leq i \leq n} \sup_{u \in \Omega} \frac{b-a}{\big|\lambda_i(u)\big|}, \tag{2.4}$$

we claim that there exists an entropy weak solution $u : [0, 2\tau] \times [a,b] \mapsto \Omega$ such that

$$u(0,x) \equiv v, \qquad u(2\tau, x) \equiv v'. \tag{2.5}$$

The function u is constructed as follows (see Figure 2). For $t \in [0, \tau]$, we let u be the solution of the Riemann problem

$$u(0,x) = \begin{cases} v & \text{if } \ x < b, \\ v'' & \text{if } \ x > b. \end{cases} \tag{2.6}$$

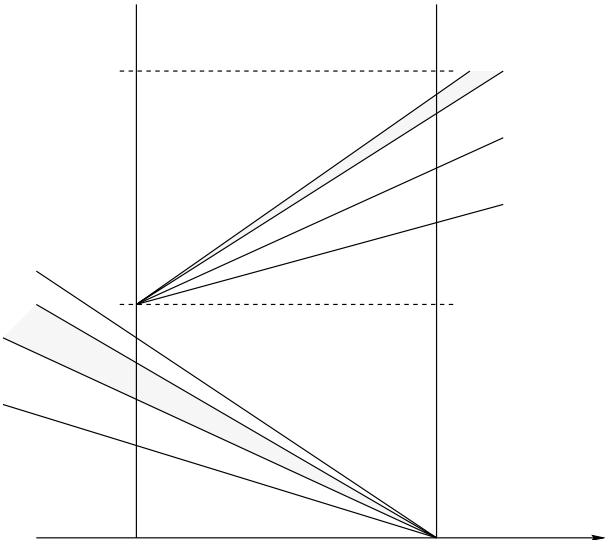


FIG. 2.

Moreover, for $t \in [\tau,\ 2\tau]$, we define u as the solution of the Riemann problem

$$u(\tau, x) = \begin{cases} v' & \text{if} \quad x < a, \\ v'' & \text{if} \quad x > a. \end{cases} \tag{2.7}$$

It is now clear that the restriction of u to the domain $[0,\ 2\tau] \times [a, b]$ satisfies the conditions in (2.5). Indeed, by (2.3), on $[0, \tau]$ the solution u contains only waves of families $\leq p$, originating at the point $(0, b)$. By (2.4), these waves cross the whole interval $[a, b]$ and exit from the boundary point a before time τ. Hence $u(\tau, x) \equiv v''$. Similarly, still by (2.3), for $t \in [\tau,\ 2\tau]$, the function u contains only waves of families $\geq p+1$, originating at the point (τ, a). By (2.4), these waves cross the whole interval $[a, b]$ and exit from the boundary point b before time 2τ. Hence $u(2\tau, x) \equiv v'$.

Next, given any two states $\omega, \omega' \in K$, by the connectedness assumption, we can find a chain of points $\omega_0 = \omega, \omega_1, \ldots, \omega_N = \omega'$ in K such that $|\omega_i - \omega_{i-1}| < \delta$ for every $i = 1, \ldots, N$. Repeating the previous construction in connection with each pair of states (ω_{i-1}, ω_i), we thus obtain an entropy weak solution $u : [0,\ 2N\tau] \times [a, b] \mapsto \Omega$ that satisfies the conclusion of the lemma, with $T = 2N\tau$. □

In the following, we shall construct the desired solution $u = u(t, x)$ as the limit of a sequence of front tracking approximations. Roughly speaking, an ε-approximate front tracking solution is a piecewise constant function u^ε, having jumps along a finite set of straight lines in the t-x plane, say, $x = x_\alpha(t)$, which approximately satisfies the Rankine–Hugoniot equations:

$$\sum_\alpha \left| f\big(u(t, x_\alpha+)\big) - f(u(t, x_\alpha-)\big) - \dot{x}_\alpha \left(u(t, x_\alpha+) - u(t, x_\alpha-)\right)\right| < \varepsilon$$

for all $t > 0$. For details, see [4, p. 125].

LEMMA 2. *In the setting of Theorem* 1*, for every state* $u^* \in \Omega$*, there exist constants* $C, \delta_0 > 0$ *for which the following holds. For any* $\varepsilon > 0$ *and every piecewise constant function* $\bar{u} : [a, b] \mapsto \Omega$ *such that*

$$\rho \doteq \sup_{x \in [a,b]} \left|\bar{u}(x) - u^*\right| \leq \delta_0, \qquad \delta \doteq \text{Tot.Var.}\{\bar{u}\} \leq \delta_0, \tag{2.8}$$

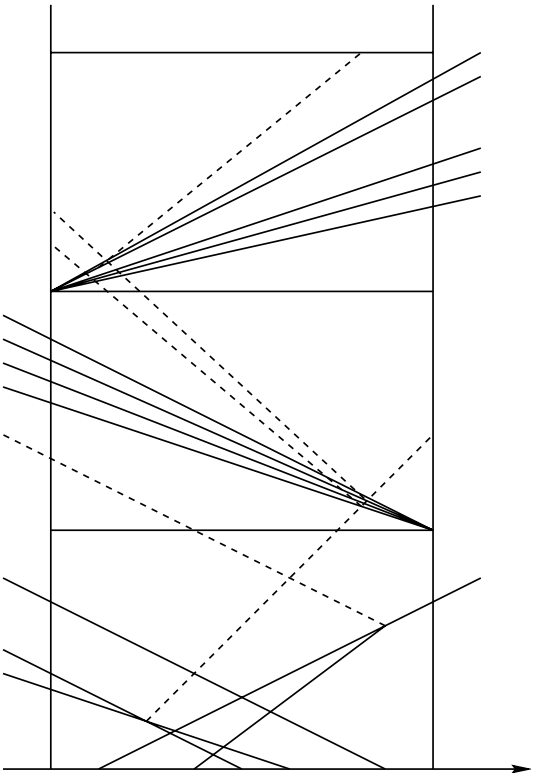


FIG. 3.

there exists an ε-approximate front tracking solution $u = u(t,x)$ of (1.1), with $u(0,x) = \bar{u}(x)$, such that

$$\sup_{x\in[a,b]} \big|u(3\tau,\, x) - u^*\big| \leq C\delta^2, \qquad \text{Tot.Var.}\big\{u(3\tau)\big\} \leq C\delta^2. \tag{2.9}$$

Proof. On the domain $(t,x) \in [0,\tau] \times [a,b]$, we construct u as an ε-approximate front tracking solution in such a way that, whenever a front hits one of the boundaries $x = a$ or $x = b$, no reflected front is ever created (see Figure 3). Since all fronts emerging from the initial data $\bar{u}$ at time $t = 0$ exit from $[a,b]$ within time τ, it is clear that $u(\tau)$ can contain only fronts of second or higher generation order. In other words, the only fronts that can be present in $u(\tau,\cdot)$ are the new ones, generated by interactions at times $t > 0$ (the dotted lines in Figure 3). Therefore, using the interaction estimate (7.69) in [4], we obtain

$$\sup_{x\in[a,b]} \big|u(\tau,\, x) - u^*\big| = \mathcal{O}(1)\cdot(\rho+\delta), \qquad \text{Tot.Var.}\big\{u(\tau)\big\} = \mathcal{O}(1)\cdot\delta^2. \tag{2.10}$$

We now apply a similar procedure as in the proof of Lemma 1 and construct a solution on the interval $[\tau,\ 3\tau]$ in such a way that $u(3\tau) \approx u^*$. More precisely, to construct u on the domain $[\tau,\ 2\tau] \times [a,b]$, consider the state v'' implicitly defined by (2.2), with $v \doteq u(\tau, b-)$, $v' \doteq u^*$. On a forward neighborhood of the point (τ,b), we let u coincide with (a front tracking approximation of) the solution to the Riemann problem

$$u(\tau,x) = \begin{cases} u(\tau,b-) & \text{if} \quad x < b, \\ v'' & \text{if} \quad x > b. \end{cases}$$

This procedure will introduce at the point (τ, b) a family of wave-fronts of families $i = 1, \dots, p$, whose total strength is $\mathcal{O}(1) \cdot (\rho + \delta)$. Because of (2.4), all of these fronts will exit from the boundary $x = a$ within time 2τ. Of course, they can interact with the other fronts present in $u(\tau, \cdot)$. In any case, the total strength of fronts in $u(2\tau, \cdot)$ is still estimated as

$$\text{Tot.Var.}\big\{u(2\tau)\big\} = \mathcal{O}(1) \cdot \delta^2. \tag{2.11}$$

Next, to define u for $t \in [2\tau,\, 3\tau]$, consider the state v''' implicitly defined by

$$\begin{cases} u(2\tau, a+) = \Psi_n(\sigma_n) \circ \cdots \circ \Psi_{p+1}(\sigma_{p+1})(v'''), \\ \qquad u^* = \Psi_p(\sigma_p) \circ \cdots \circ \Psi_1(\sigma_1)(v'''). \end{cases} \tag{2.12}$$

On a forward neighborhood of the point $(2\tau, a)$, we let u coincide with (a front tracking approximation of) the solution to the Riemann problem

$$u(2\tau, x) = \begin{cases} u(2\tau,\, a+) & \text{if} \quad x > a, \\ v''' & \text{if} \quad x < a. \end{cases}$$

This procedure introduces at the point $(2\tau,\, a)$ a family of wave-fronts of families $i = p+1, \dots, n$, whose total strength is $\mathcal{O}(1) \cdot (\rho + \delta)$. Because of (2.4), all of these fronts will exit from the boundary $x = b$ within time 3τ. Of course, they can interact with the other fronts present in $u(2\tau, \cdot)$. In any case, the total strength of fronts in $u(3\tau, \cdot)$ is still estimated as

$$\text{Tot.Var.}\big\{u(3\tau)\big\} = \mathcal{O}(1) \cdot \delta^2. \tag{2.13}$$

Moreover, the difference between the values $u(3\tau, x)$ and u^* will be of the same order of the total strength of waves in $u(\tau, \cdot)$ so that the first inequality in (2.9) will also hold. □

Proof of Theorem 1. Using the same arguments as in the proof of Lemma 1, for every $\varepsilon > 0$, we can construct an ε-approximate front tracking solution $u = u(t, x)$ on $[0,\, 2N\tau] \times [a, b]$ such that

$$\sup_{x \in [a,b]} \big|u(2N\tau, x) - u^*\big| = \mathcal{O}(1) \cdot \delta, \qquad \text{Tot.Var.}\big\{u(2N\tau)\big\} = \mathcal{O}(1) \cdot \delta. \tag{2.14}$$

Choosing $\delta > 0$ sufficiently small, we can assume that, in (2.14), $\mathcal{O}(1) \cdot \delta < \delta_0 < 1/C$, the constant in Lemma 2. Calling $T \doteq 2N\tau$, we can now repeat the construction described in Lemma 2 on each interval $\big[T + 3k\tau,\ T + 3(k+1)\tau\big]$. This yields

$$\sup_{x \in [a,b]} \big|u(T + 3k\tau,\, x) - u^*\big| \leq \delta_k, \qquad \text{Tot.Var.}\big\{u(T + 3k\tau)\big\} \leq \delta_k, \tag{2.15}$$

where the constants δ_k satisfy the inductive relations

$$\delta_{k+1} \leq C\delta_k^2. \tag{2.16}$$

Choosing a sequence of ε-approximate front tracking solutions u_ε satisfying (2.15)–(2.16) and taking the limit as $\varepsilon \to 0$, we obtain an entropy weak solution u which still satisfies the same estimates. The bounds (1.7)–(1.8) are now a consequence of (2.15)–(2.16), with a suitable choice of the constants C_0, κ. □

3. Decay of positive waves. Throughout the following, we consider a 2×2 system of conservation laws

$$u_t + f(u)_x = 0, \tag{3.1}$$

satisfying the assumptions (H). Following [6, p. 128], we construct a set of Riemann coordinates (w_1, w_2). One can then choose the right eigenvectors of $Df(u)$ so that

$$r_i(u) = \frac{\partial u}{\partial w_i}, \qquad \frac{\partial \lambda_i}{\partial w_i} = D\lambda_i \cdot r_i > 0, \qquad i = 1, 2. \tag{3.2}$$

It will be convenient to perform most of the analysis on a special class of solutions: piecewise Lipschitz functions with finitely many shocks and no compression waves. Due to the geometric structure of the system, this set of functions turns out to be positively invariant for the flow generated by the hyperbolic system. We first derive several a priori estimates concerning these solutions, in particular on the strength and location of the shocks. We then observe that any BV solution can be obtained as a limit of a sequence of piecewise Lipschitz solutions in our special class. Our estimates can thus be extended to general BV solutions.

DEFINITION 1. *We call $\mathcal{U}$ the set of all piecewise Lipschitz functions $u : \mathbb{R} \mapsto \mathbb{R}^2$ with finitely many jumps such that*

(i) *at every jump, the corresponding Riemann problem is solved only in terms of shocks (no centered rarefactions);*
(ii) *no compression waves are present; i.e., $w_{i,x}(x) \geq 0$ at almost every $x \in \mathbb{R}$, $i = 1, 2$.*

The next lemma establishes the forward invariance of the set $\mathcal{U}$.

LEMMA 3. *Consider the 2×2 system of conservation laws* (3.1) *satisfying the assumptions* (H). *Let $u = u(t, x) >$ be the solution to a Cauchy problem, with small total variation, satisfying $u(0, \cdot) \in \mathcal{U}$. Then*

$$u(t, \cdot) \in \mathcal{U} \qquad \text{for all } \ t \geq 0. \tag{3.3}$$

Proof. We have to show that, as time progresses, the total number of shocks does not increase and no compression wave is ever formed. This will be the case provided that the following hold:

(i) The interaction of two shocks of the same family produces an outgoing shock in the other family.
(ii) The interaction of a shock with an infinitesimal rarefaction wave of the same family produces a rarefaction wave in the other family.

Both of the above conditions can be easily checked by analyzing the relative positions of shocks and rarefaction curves. We will do this for the first family, leaving the verification of the other case to the reader.

Call $\sigma \mapsto R_1(\sigma)$ the rarefaction curve through a state u_0, parametrized so that

$$\lambda_1\big(R_1(\sigma)\big) = \lambda_1(u_0) + \sigma.$$

It is well known that the shock curve through u_0 has a second order tangency with this rarefaction curve. Hence there exists a smooth function $c_1(\sigma)$ such that the point

$$S_1(\sigma) \doteq R_1(\sigma) + c_1(\sigma)\frac{\sigma^3}{6} r_2(u_0)$$

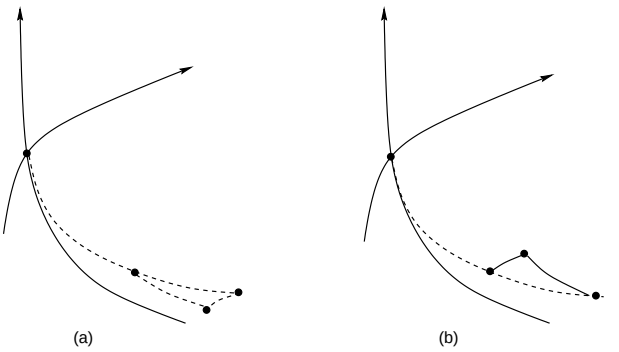


FIG. 4.

lies on this shock curve for all σ in a neighborhood of zero. From the Rankine–Hugoniot equations, it now follows that

$$\chi(\sigma) \doteq \big(f\big(R_1(\sigma)+c_1(\sigma)(\sigma^3/6)\, r_2(u_0)\big)-f(u_0)\big)\wedge\big(R_1(\sigma)+c_1(\sigma)(\sigma^3/6)\, r_2(u_0)-u_0\big) = 0. \tag{3.4}$$

Differentiating the wedge product (3.4) four times at $\sigma = 0$ and denoting derivatives with upper dots, we obtain

$$\begin{aligned}\frac{d^4\chi}{d\sigma^4}(0) &= 4\big[\lambda_1(u_0)\,\dddot{R}_1\,(0) + 2\ddot{R}_1(0) + \lambda_2(u_0)\, c_1(0) r_2(u_0)\big] \wedge \dot{R}_1(0)\\ &\quad + 6\big[\lambda_1(u_0)\ddot{R}_1(0) + \dot{R}_1(0)\big] \wedge \ddot{R}_1(0) + 4\lambda_1(u_0)\dot{R}_1(0) \wedge \big[\,\dddot{R}_1\,(0) + c(0) r_2(u_0)\big]\\ &= 4\big(\lambda_2(u_0) - \lambda_1(u_0)\big)c_1(0)\, r_2(u_0) \wedge r_1(u_0) + 2(Dr_1 \cdot r_1)(u_0) \wedge r_1(u_0)\\ &= 0.\end{aligned}$$

Hence

$$c_1(0) = \frac{(Dr_1 \cdot r_1) \wedge r_1}{2(\lambda_2 - \lambda_1)(r_1 \wedge r_2)} < 0. \tag{3.5}$$

By (3.5), the relative position of 1-shock and 1-rarefaction curves is as depicted in Figure 1. By the geometry of wave curves, the properties (i) and (ii) are now clear. Figure 4a illustrates the interaction of two 1-shocks, while Figure 4b shows the interaction between a 1-shock and a 1-rarefaction. By u_l, u_m, u_r we denote the left, middle, and right states before the interaction, while u'_m is the middle state after the interaction. In the two cases, the solution of the Riemann problem contains a 2-shock and a 2-rarefaction, respectively. □

The next lemma shows the decay of positive waves for solutions with small total variation, taking values inside $\mathcal{U}$.

LEMMA 4. *Let $u = u(t,x)$ be a solution of the Cauchy problem for the 2×2 system* (3.1) *satisfying* (H). *Assume that*

$$u(t,\cdot) \in \mathcal{U}, \qquad t \geq 0. \tag{3.6}$$

Then there exist $\kappa, \delta > 0$ such that, if Tot.Var.$(u(t,\cdot)) < \delta$ *for all t, then its Riemann coordinates (w_1, w_2) satisfy*

$$0 \leq w_{i,x}(t,x) \leq \frac{\kappa}{t}, \qquad t > 0, \quad i = 1,2. \tag{3.7}$$

Proof. We consider the case $i = 1$. Fix any point $(\bar t,\ \bar x)$. Since centered rarefaction waves are not present, there exists a unique 1-characteristic through this point, which we denote as $t \mapsto x_1(t;\ \bar t,\ \bar x)$. It is the solution of the Cauchy problem

$$\dot x(t) = \lambda_1\big(u(t, x(t))\big), \qquad x(\bar t) = \bar x. \tag{3.8}$$

The evolution of $w_{1,x}$ along this characteristic is described by

$$\frac{d}{dt} w_{1,x}\big(t, x_1(t)\big) = w_{1,xt} + \lambda_1 w_{1,xx} = -(\lambda_1 w_{1,x})_x + \lambda_1 w_{1,xx} = -\frac{\partial \lambda_1}{\partial w_1} w_{1,x}^2 - \frac{\partial \lambda_1}{\partial \omega_2} w_{1,x} w_{2,x}.$$

Since the system is genuinely nonlinear, there exists $k_1 > 0$ such that $\partial\lambda_1/\partial w_1 \geq k_1 > 0$, and hence

$$\frac{d}{dt} w_{1,x}\big(t, x_1(t)\big) \leq -k_1 w_{1,x}^2 + \mathcal{O}(1) \cdot w_{1,x} w_{2,x}. \tag{3.9}$$

Moreover, at each time t_α where the characteristic crosses a 2-shock of strength $|\sigma_\alpha|$, we have the estimate

$$w_{1,x}(t_\alpha+) \leq \big(1 + \mathcal{O}(1) \cdot |\sigma_\alpha|\big) w_{1,x}(t_\alpha-). \tag{3.10}$$

Let $Q(t)$ be the total interaction potential at time t (see, for example, [4, p. 202]), and let $V_2(t)$ be the total amount of 2-waves approaching our 1-wave located at $x_1(t)$. Repeating the arguments in [4, p. 139], we can find a constant $C_0 > 0$ such that the quantity

$$\Upsilon(t) \doteq V_1(t) + C_0\, Q(t), \qquad t > 0,$$

is nonincreasing. Moreover, for almost every t, one has

$$\dot\Upsilon(t) \leq -\big|\lambda_2 - \lambda_1\big| |w_{2,x}|\big(t, x_1(t)\big),$$

while, at times t_α, where x_1 crosses a 2-shock of strength $|\sigma_\alpha|$, there holds

$$\Upsilon(t_\alpha-) \leq \Upsilon(t_\alpha+) - |\sigma_\alpha|.$$

Call $W(t) \doteq w_{1,x}\big(t, x_1(t)\big)$. By the previous estimates, from (3.9) and (3.10), it follows that

$$\begin{aligned} \dot W(t) &\leq -k_1 W^2(t) - C\, \dot\Upsilon(t) W(t), \\ W(t_\alpha+) - W(t_\alpha-) &\leq C\big[\Upsilon(t_\alpha+) - \Upsilon(t_\alpha-)\big] W(t_\alpha-) \end{aligned} \tag{3.11}$$

for a suitable constant C. We now observe that

$$y(t) \doteq \frac{e^{-C\Upsilon(t)}}{\int_0^t k_1 e^{-C\Upsilon(s)} ds}$$

is a distributional solution of the equation

$$\dot y = -k_1 y^2 - C\, \dot\Upsilon(t) y,$$

with $y(t) \to \infty$ as $t \to 0+$. A comparison argument now yields $W(t) \leq y(t)$. Since Υ is positive and decreasing, we have

$$W(t) \leq \frac{1}{k_1} \frac{1}{\int_0^t e^{-C\Upsilon(s)} ds} \leq \frac{e^{C\Upsilon(0)}}{k_1 t}$$

for all $t > 0$. This establishes (3.7) for $i = 1$, with $\kappa \doteq e^{C\Upsilon(0)}/k_1$. The case $i = 2$ is identical. □

We conclude this section by proving a decay estimate for positive waves, valid for general BV solutions of the system (3.1). For this purpose, we need to recall some definitions introduced in [5]. See also [4, p. 201].

Let $u : \mathbb{R} \mapsto \mathbb{R}^2$ have bounded variation. By possibly changing the values of u at countably many points, we can assume that u is right continuous. The distributional derivative $\mu \doteq D_x u$ is a vector measure, which can be decomposed into a continuous and an atomic part: $\mu = \mu_c + \mu_a$. For $i = 1, 2$, the scalar measures $\mu^i = \mu^i_c + \mu^i_a$ are defined as follows. The continuous part of μ^i is the Radon measure μ^i_c such that

$$\int \phi \, d\mu^i_c = \int \phi \, l_i(u) \cdot d\mu_c \tag{3.12}$$

for every scalar continuous function ϕ with compact support. The atomic part of μ^i is the measure μ^i_a concentrated on the countable set $\{x_\alpha;\ \alpha = 1, 2, \ldots\}$, where u has a jump, such that

$$\mu^i_a\big(\{x_\alpha\}\big) = \sigma_{\alpha,i} \doteq E_i\big(u(x_\alpha-),\ u(x_\alpha+)\big) \tag{3.13}$$

is the size of the ith wave in the solution of the corresponding Riemann problem with data $u(x_\alpha\pm)$. We regard μ^i as the *measure of i-waves* in the solution u. It can be decomposed in a positive and a negative part so that

$$\mu^i = \mu^{i+} - \mu^{i-}, \qquad |\mu^i| = \mu^{i+} + \mu^{i-}. \tag{3.14}$$

The decay estimate in (3.7) can now be extended to general BV solutions. Indeed, we show that the density of positive i-waves decays as κ/t. By meas(J) we denote here the Lebesgue measure of a set J.

LEMMA 5. *Let $u = u(t, x)$ be a solution of the Cauchy problem for the 2×2 system* (3.1) *satisfying* (H). *Then there exist $\kappa, \delta > 0$ such that, if* Tot.Var.$(u(t, \cdot)) < \delta$ *for all t, then the measures μ^{1+}_t, μ^{2+}_t of positive waves in $u(t, \cdot)$ satisfy*

$$\mu^{i+}_t(J) \le \frac{\kappa}{t}\text{meas}(J) \tag{3.15}$$

for every Borel set $J \subset \mathbb{R}$ and every $t > 0$, $i = 1, 2$.

Proof. For every BV solution u of (3.1), we can construct a sequence of solutions u_ν with $u_\nu \to u$ as $\nu \to \infty$ and such that $u_\nu(t, \cdot) \in \mathcal{U}$ for all t. Calling (w^ν_1, w^ν_2) the Riemann coordinates of u_ν, by Lemma 4 we have

$$0 \le w^\nu_{i,x}(t, x) \le \frac{\kappa}{t}, \qquad t > 0, \quad i = 1, 2, \quad \nu \ge 1. \tag{3.16}$$

For a fixed $t > 0$, observe that the map $x \mapsto w^\nu_1(t, x)$ has upward jumps precisely at the points x_α where $u(t, \cdot)$ has a 2-shock. Define $\tilde\mu_\nu$ as the positive, purely atomic measure, concentrated on the finitely many points x_α where $u(t, \cdot)$ has a 2-shock, such that

$$\tilde\mu_\nu\big(\{x_\alpha\}\big) = w^\nu_1(t,\, x_\alpha+) - w^\nu_1(t,\, x_\alpha-) \le C\,|\sigma_\alpha|^3 \tag{3.17}$$

for some constant C. By possibly taking a subsequence, we can assume the existence of a weak limit $\tilde\mu_\nu \rightharpoonup \tilde\mu$. Because of the estimate in (3.17), the measure $\tilde\mu$ is purely

atomic and is concentrated on the set of points x_β, which are limits as $\nu \to \infty$ of a sequence of points x^ν_α, where $u_\nu(t,\cdot)$ has a 2-shock of uniformly positive strength $|\sigma_\nu| \geq \delta > 0$. Therefore, $\tilde\mu$ is concentrated on the set of points where the limit solution $u(t,\cdot)$ has a 2-shock and makes no contribution to the positive part of μ_t^{1+}. We thus conclude that the positive part of μ_t^{1+} is absolutely continuous with respect to Lebesgue measure, with density $\leq \kappa/t$. An analogous argument holds for μ_t^{2+}. □

COROLLARY 1. *Let $u = u(t,x)$ be a solution of the 2×2 system (1.1). Let the assumptions* (H) *hold. Fix $\varepsilon > 0$, and consider the subinterval $[a',b'] \doteq [a+\varepsilon,\ b-\varepsilon]$. Assume that, at time $t = 0$, the measures μ^{1+}, μ^{2+} of positive waves in $u(0,\cdot)$ on $[a,b]$ vanish identically. Then, for every $t > 0$, one has*

$$\mu_t^{i+}(J) \leq \frac{\kappa\lambda^*}{\varepsilon}\mathrm{meas}(J) \tag{3.18}$$

for every Borel set $J \subset [a',b']$ and every $t > 0$, $i = 1,2$.

Indeed, recalling (1.9), the values of $u(t,\cdot)$ restricted to the interval $[a',b']$ can be obtained by solving a Cauchy problem, with initial data assigned on the whole interval $[a,b]$ at time $t - \varepsilon/\lambda^*$.

4. Proof of Theorem 2.

LEMMA 6. *In the same setting as Lemma 4, assume that there exists $\kappa' > 0$ such that*

$$0 \leq w_{i,x}(t,x) \leq \kappa', \qquad t \in [0,T], \quad i = 1,2. \tag{4.1}$$

Let $t \mapsto x(t)$ be the location of a shock, with strength $\big|\sigma(t)\big|$. There exists a constant $0 < c < 1$ such that

$$\big|\sigma(t)\big| \geq c\big|\sigma(s)\big|, \qquad 0 \leq s < t \leq T. \tag{4.2}$$

Proof. To fix the ideas, let $u(t,\cdot)$ have a 1-shock located at $x(t)$, with strength $\big|\sigma(t)\big|$. Outside points of interaction with other shocks, the strength satisfies an inequality of the form

$$\frac{d}{dt}\big|\sigma(t)\big| \geq -C\cdot\big(w_{1,x}\big(t,x(t)+\big)+w_{1,x}\big(t,x(t)-\big)w_{2,x}\big(t,x(t)+\big)+w_{2,x}\big(t,x(t)-\big)\big)\big|\sigma(t)\big|. \tag{4.3}$$

At times where our 1-shock interacts with other 1-shocks, its strength increases. Moreover, at each time t_α where our 1-shock interacts with a 2-shock, say, of strength $|\sigma_\alpha|$, one has

$$\big|\sigma(t_\alpha+)\big| \geq \big|\sigma(t_\alpha-)\big|\,\big(1 - C'|\sigma_\alpha|\big) \tag{4.4}$$

for some constant C'. Assuming that the total variation remains small, the total number of 2-shocks which cross any given 1-shock is uniformly small. Hence (4.3)–(4.4) together imply (4.2). □

LEMMA 7. *Let $t \mapsto u(t,\cdot) \in \mathcal{U}$ be a solution of the Cauchy problem for a genuinely nonlinear 2×2 system satisfying (1.11). Assume that there exists $\kappa' > 0$ such that*

$$w_{i,x}(t,x) \leq \kappa', \qquad t \in [0,T], \quad i = 1,2. \tag{4.5}$$

Since no centered rarefactions are present, any two i-characteristics, say, $x(t) < y(t)$, can uniquely be traced backward up to time $t = 0$. There exists a constant $L > 0$ such that

$$y(t) - x(t) \leq L\big(y(s) - x(s)\big), \qquad 0 \leq s < t \leq T. \tag{4.6}$$

Proof. Consider the case when $i = 2$. By definition, the characteristics are solutions of

$$\dot{x}(t) = \lambda_2\big(u(t, x(t))\big), \qquad \dot{y}(t) = \lambda_2\big(u(t, y(t))\big).$$

Since the characteristic speed λ_2 decreases across 2-shocks, we can write

$$\dot{y}(t) - \dot{x}(t) \leq C \int_{x(t)}^{y(t)} \big|w_{1,x}(t,\xi)\big| + \big|w_{2,x}(t,\xi)\big|\, d\xi + C \sum_{\alpha \in \mathcal{S}_1[x,y]} \big|\sigma_\alpha(t)\big|, \tag{4.7}$$

where $\mathcal{S}_1[x, y]$ denotes the set of all 1-shocks located inside the interval $\big[x(t),\ y(t)\big]$. Introduce the function

$$\phi(t,x) \doteq \begin{cases} 0 & \text{if} \quad x \leq x(t), \\ \frac{x - x(t)}{y(t) - x(t)} & \text{if} \quad x(t) < x < y(t), \\ 1 & \text{if} \quad x \geq y(t). \end{cases}$$

Moreover, define the functional

$$\Phi(t) \doteq \sum_{\alpha \in \mathcal{S}_1} \phi\big(t, x_\alpha(t)\big)\, \big|\sigma_\alpha(t)\big| + C_0\, Q(t),$$

where the summation now refers to all 1-shocks in $u(t, \cdot)$ and Q is the usual interaction potential. Observe that the map $t \mapsto \Phi(t)$ is nonincreasing. By (4.5) and (4.7), we can now write

$$\dot{y}(t) - \dot{x}(t) \leq C'\big(1 - \dot{\Phi}(t)\big)\big(y(t) - x(t)\big)$$

for some constant C'. This implies (4.6) with $L = \exp\big\{C'T + C'\Phi(0)\big\}$. ∎

The next result is the key ingredient toward the proof of Theorem 2. It provides the density of the set of interaction points where new shocks are generated.

LEMMA 8. *Fix $\varepsilon > 0$, and define $a'' = a + 2\varepsilon$, $b'' = b - 2\varepsilon$. Consider a 2×2 system of the form* (1.1)*, satisfying* (H)*. Let u be an entropy weak solution defined on $[0, \tau] \times [a, b]$, with $\tau \doteq \varepsilon/4\lambda^*$. Let* (3.18) *hold for all $t \in [0, \tau]$, and assume that $u(0, \cdot)$ has a dense set of 1-shocks on the interval $[a'',\ b'']$. Then, for $0 \leq t \leq \tau$, the solution $u(t, \cdot)$ has a set of 1-shocks which is dense on $[a'',\ b' - \lambda^* t]$ and a set of 2-shocks which is dense on $[a'',\ b'']$.*

Proof. By the assumptions of the lemma, there exists a sequence of piecewise Lipschitz solutions $t \mapsto u_\nu(t) \in \mathcal{U}$ such that $u_\nu \to u$ in $\mathbf{L}^1$,

$$0 \leq w^\nu_{i,x}(t,x) \leq \frac{2\kappa\lambda^*}{\varepsilon}, \qquad i = 1, 2, \quad \nu \geq 1,$$

and, moreover, the following holds. For every $\rho > 0$, there exists $\delta > 0$ such that each $u_\nu(0, \cdot)$ (with ν large enough) contains at least one 1-shock of strength $\big|\sigma_\nu(0)\big| \geq \delta$ on every subinterval $J \subset [a'',\ b'']$ having length $\geq \rho$.

To prove the first statement in Lemma 8, fix $t \in [0, \tau]$, and consider any nontrivial interval $[p, q] \subseteq [a'',\ b'' - t\lambda^*]$. Call $s \mapsto p_\nu(s)$, $s \mapsto q_\nu(s)$ the backward characteristics through these points, relative to the solution u_ν. We thus have

$$\begin{cases} \dot{p}_\nu(s) = \lambda_1\big(u_\nu(s,\, p_\nu(s))\big), \\ \dot{q}_\nu(s) = \lambda_1\big(u_\nu(s,\, q_\nu(s))\big), \end{cases} \qquad \begin{cases} p_\nu(t) = p, \\ q_\nu(t) = q. \end{cases}$$

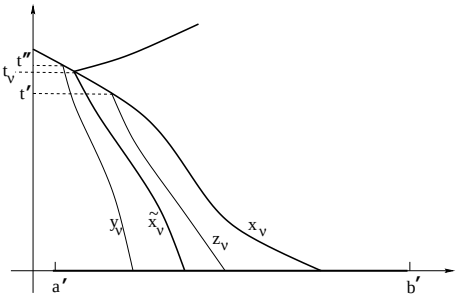


FIG. 5.

By Lemma 7, $q_\nu(0) - p_\nu(0) \geq \rho$ for some $\rho > 0$ independent of ν. Hence, each solution u_ν contains a shock of strength $\big|\sigma_\nu(s)\big| \geq \delta$ located inside the interval $\big[p_\nu(0),\ q_\nu(0)\big]$. Lemma 5 now yields $\big|\sigma_\nu(t)\big| \geq c\delta$. By possibly taking a subsequence, we conclude that the limit solution $u(t,\cdot)$ contains a 1-shock of positive strength at the point $x(t) = \lim x_\nu(t) \in [p,q]$.

To prove the second statement, we will show that the set of points where two 1-shocks in u interact and produce a new 2-shock is dense on the triangle

$$\Delta \doteq \big\{(t,x)\,;\quad t \in [0,\tau],\quad a'' < x < b'' - \lambda^* t\big\}.$$

Indeed, let $t \in [0,\tau]$ and $p < q$ be as before. For each ν sufficiently large, let $t \mapsto x_\nu(t)$ be the location of a 1-shock in u_ν, with strength $\big|\sigma_\nu(t)\big| \geq \delta > 0$. Assume $x_\nu(\cdot) \to x(\cdot)$ as $\nu \to \infty$, and $x_\nu(t) \in [p,q]$, so that $x(t)$ is the location of a 1-shock of the limit solution u, say, with strength $\big|\sigma(t)\big| > 0$.

We claim that the set of times $\hat{t}$ where some other 1-shock σ' impinges on σ and generates a new 2-shock is dense on $[0,t]$. To see this, fix $0 < t' < t'' < t$. For each ν sufficiently large, consider the backward 1-characteristics $y_\nu,\ z_\nu$ impinging from the left on the shock x_ν at times t'', t', respectively (see Figure 5). These provide solutions to the Cauchy problems

$$\begin{aligned} \dot{y}_\nu(t) &= \lambda_1\big(u_\nu\big(t, y_\nu(t)\big)\big), \qquad & y_\nu(t'') &= x_\nu(t''), \\ \dot{z}_\nu(t) &= \lambda_1\big(u_\nu(t, z_\nu(t))\big), \qquad & z_\nu(t') &= x_\nu(t'), \end{aligned}$$

respectively. Observe that

$$z_\nu(0) - y_\nu(0) \geq \rho$$

for some $\rho > 0$ independent of ν. Indeed, the genuine nonlinearity of the system implies

$$\lambda_1\big(u_\nu(t, x_\nu(t)-)\big) - \dot{x}_\nu(t) \geq \kappa\Big|u_\nu\big(t,\ x_\nu(t)+\big) - u_\nu\big(t,\ x_\nu(t)-\big)\Big| \geq \kappa\delta.$$

Therefore,

$$x_\nu(t') - y_\nu(t') \geq \rho' > 0$$

for some constant $\rho' > 0$ independent of ν. By Lemma 6, the interval $\big[y_\nu(0),\ z_\nu(0)\big]$ has uniformly positive length. Hence it contains a 1-shock of $u_\nu(0,\cdot)$ with uniformly

positive strength $\big|\sigma_\nu(0)\big| \geq \delta > 0$. By Lemma 5, every u_ν has a 1-shock with strength $\big|\sigma_\nu(t)\big| \geq c\delta$ located along some curve $t \mapsto \tilde{x}_\nu(t)$ with

$$y_\nu(t) < \tilde{x}_\nu(t) < z_\nu(t), \qquad t \in [0, t'].$$

Clearly, this second 1-shock impinges on the shock x_ν at some time $t_\nu \in [t', t'']$, creating a new 2-shock with uniformly large strength. Letting $\nu \to \infty$, we obtain the result. □

Proof of Theorem 2. Let $\delta_0 > 0$ be given. We can then construct an initial condition $u(0, \cdot) = \phi$, with Tot.Var.$\{\phi\} < \delta_0$, having a dense set of 1-shocks on the interval $[a, b]$ and no other waves. As a consequence, for any $\varepsilon > 0$, by Corollary 1, we have the estimate (3.18) on the density of positive waves away from the boundary.

Fix $\tau = \varepsilon/4\lambda^*$, and consider again the subinterval $[a'',\ b''] = [a + 2\varepsilon,\ b - 2\varepsilon]$. We can apply Lemma 8 first on the time interval $[0, \tau]$, obtaining the density of 2-shocks on the region $[0, \tau] \times [a'',\ b'']$. Then, by induction on m, the same argument is repeated on each time interval $t \in \big[m\tau,\ (m+1)\tau\big]$, proving the theorem. □

Acknowledgment. The second author warmly thanks professor Benedetto Piccoli for stimulating conversations.

REFERENCES

[1] D. Amadori, *Initial-boundary value problems for nonlinear systems of conservation laws*, NoDEA Differential Equations Appl., 4 (1997), pp. 1–42.

[2] D. Amadori and R. M. Colombo, *Continuous dependence for* 2×2 *conservation laws with boundary*, J. Differential Equations, 138 (1997), pp. 229–266.

[3] F. Ancona and A. Marson, *On the attainable set for scalar nonlinear conservation laws with boundary control*, SIAM J. Control Optim., 36 (1998), pp. 290–312.

[4] A. Bressan, *Hyperbolic Systems of Conservation Laws. The One Dimensional Cauchy Problem*, Oxford University Press, Oxford, UK, 2000.

[5] A. Bressan and R. M. Colombo, *Decay of positive waves in nonlinear systems of conservation laws*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 26 (1998), pp. 133–160.

[6] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer-Verlag, New York, 2000.

[7] R. DiPerna, *Global solutions to a class of nonlinear hyperbolic systems of equations*, Comm. Pure Appl. Math., 26 (1973), pp. 1–28.

[8] P. Lax, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[9] T. Li and B. Rao, *Exact Boundary Controllability for Quasilinear Hyperbolic Systems*, to appear.

[10] T. Li and J. Yi, *Semi-global* C^1 *solution to the mixed initial-boundary value problem for quasilinear hyperbolic systems*, Chinese Ann. Math. Ser. B, 21 (2000), pp. 165–186.

[11] T. Li and W. Yu, *Boundary Value Problems for Quasilinear Hyperbolic Systems*, Duke University Mathematics Series V, Mathematics Department, Duke University, Durham, NC, 1985.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 2, pp. 623–640

NECESSARY SUBOPTIMALITY AND OPTIMALITY CONDITIONS VIA VARIATIONAL PRINCIPLES*

BORIS S. MORDUKHOVICH† AND BINGWU WANG†

Abstract. The paper aims to develop some basic principles and tools of nonconvex variational analysis with applications to necessary suboptimality and optimality conditions for constrained optimization problems in infinite dimensions. We establish a certain subdifferential variational principle as a new characterization of Asplund spaces. This result is different from conventional support forms of variational principles and appears to be convenient for applications to nonsmooth optimization. Based on the subdifferential variational principle, we obtain new necessary conditions for suboptimal solutions in general nonsmooth optimization problems with equality, inequality, and set constraints in Asplund spaces. In this way we establish the so-called sequential normal compactness properties of constraint sets that play an essential role in infinite-dimensional variational analysis and its applications. As a by-product of our approach, we derive various forms of necessary optimality conditions for nonsmooth constrained problems in infinite dimensions, which extend known results in that direction.

Key words. variational analysis, variational principles, nonsmooth optimization, Banach and Asplund spaces, generalized differentiation, suboptimality, necessary optimality conditions

AMS subject classifications. 49J52, 49K27, 90C48

PII. S0363012900374816

1. Introduction.

This paper is devoted to variational analysis in infinite dimensions and its applications to optimization problems. The main topic is related to *variational principles* that play a crucial role in nonlinear analysis and its various applications in infinite-dimensional spaces; cf. Ekeland [7], Borwein and Preiss [2], and Deville, Godefroy, and Zizler [6].

In this paper we consider the framework of *Asplund spaces* that were originally defined as Banach spaces on which every convex continuous function is generically Fréchet differentiable. This class is sufficiently broad and convenient for the theory and applications. It includes all Banach spaces with Fréchet smooth renorms or bump functions—in particular, every reflexive space. On the other hand, there are Asplund spaces that fail to have even a Gâteaux smooth renorm. The class of Asplund spaces admits many nice geometric characterizations; see, e.g., [6] and [24].

The first result of this paper gives a new *variational characterization* of Asplund spaces that we call the *subdifferential variational principle*. The major difference between this result and variational principles in the conventional support form is that, instead of a supporting/minimization condition for the given lower semicontinuous function $f\colon X \to (-\infty, \infty]$, the subdifferential variational principle provides a dual-space condition involving the Fréchet subdifferential of f. If the space in question admits a Fréchet smooth renorm (bump function), the subdifferential variational principle implies a *smooth* variational principle in the conventional support form of Borwein–Preiss (resp., Deville–Godefroy–Zizler). Fabian and Mordukhovich [9] recently proved that the mentioned smooth renorm/bump function assumption on X

*Received by the editors July 6, 2000; accepted for publication (in revised form) November 19, 2001; published electronically July 1, 2002.
http://www.siam.org/journals/sicon/41-2/37481.html

†Department of Mathematics, Wayne State University, Detroit, MI 48202 (boris@math.wayne.edu, wangbw@math.wayne.edu). The research of the first author was partly supported by the National Science Foundation under grants DMS-9704751 and DMS-0072179 and also by the Distinguished Faculty Fellowship at Wayne State University.

is not only sufficient but also *necessary* for the validity of the corresponding smooth variational principle. In contrast, the subdifferential variational principle turns out to be a *characterization* of Asplund spaces and may be treated as an appropriate variational principle for this general class of Banach spaces. In fact, we show that the subdifferential variational principle is equivalent to the *extremal principle* established in Mordukhovich and Shao [19] as another characterization of Asplund spaces. The latter result can be viewed as a variational analogue of the convex separation principle in the case of nonconvex sets.

Next we give applications of the subdifferential variational principle to *suboptimality* conditions for problems of mathematical programming in infinite dimensions. This means that we do *not* assume the *existence* of optimal solutions and obtain conditions held for suboptimal (ε-optimal) solutions, which always exist. The latter is particularly important for infinite-dimensional problems of optimization and optimal control, where the existence of optimal solutions requires quite restrictive assumptions. As pointed out by Young [30], any theory of necessary optimality conditions is "naive" until the existence of optimal solutions is clarified. This was the primary motivation for developing theories of generalized curves/relaxed controls in problems of the calculus of variations and optimal control to automatically ensure the existence of optimal solutions. For the general optimization theory in infinite dimensions, an alternative route to avoiding trouble with the existence of optimal solutions is to find "suboptimal solutions," which are "almost" optimal and which "almost" satisfy necessary conditions for optimality.

Necessary conditions for suboptimal solutions were first obtained by Ekeland [7] for classical problems of nonlinear programming with smooth equality and inequality constraints and unconstrained optimal control. More recent developments for nondifferentiable programming are given by Loridan [16], Bustos [5], and Hamel [11] that are based mostly on Ekeland's variational principle and the usage of Clarke's generalized gradients for locally Lipschitzian functions. Suboptimality conditions for various problems of optimal control can be found in Gabasov, Kirillova, and Mordukhovich [12], Mordukhovich [17], Moussaoui and Seeger [22], Sumin [27], and their references.

In this paper we establish, based on the subdifferential variational principle, necessary conditions for suboptimal solutions in a sufficiently broad class of optimization problems, with equality and inequality constraints given by locally Lipschitzian functions, as well as geometric constraints given by closed sets. The main results are expressed in terms of our basic (limiting) normal cone and subdifferential, which may be much smaller than Clarke's counterparts even in finite dimensions and enjoy *full calculus* under general qualification conditions and the so-called *sequential normal compactness* assumptions. The latter assumptions, which are automatic in finite dimensions, are crucial to perform limiting procedures and prove the required calculus rules. As an essential part of our approach, we obtain new subdifferential conditions ensuring the sequential normal compactness of constraint sets given by Lipschitzian equalities and inequalities. We also establish a weak form of suboptimality conditions for problems with non-Lipschitzian data that do not require constraint qualifications. As a by-product of our approach, we prove necessary *optimality* conditions for the mentioned problems in strong and weak forms. In particular, the weak form of these results extends to the case of Asplund spaces the necessary optimality conditions for problems with non-Lipschitzian data recently obtained by Borwein, Treiman, and Zhu [4] in reflexive spaces.

The rest of the paper is organized as follows. Section 2 contains preliminary

results and notation used in the paper. Section 3 is devoted to the subdifferential variational principle and its relation to other basic principles of variational analysis. In section 4 we present and discuss necessary suboptimality and optimality conditions for a general constrained problem of nondifferentiable programming in Asplund spaces. Proofs of these conditions are given in section 5, where we also establish sequential normal compactness of the constraint sets, which is crucial for the proof of the main results and is certainly of independent interest.

2. Preliminaries. Our notation is basically standard and can be found in [20] with most of the definitions and results presented in this section. Finite-dimensional versions of these constructions and results are given in [17] and [26], while [29] contains recent applications to optimal control.

For any Banach space X we denote its norm by $\|\cdot\|$ and the dual space by X^* with the canonical pairing $\langle\cdot,\cdot\rangle$; $\mathbb{B}$ and $\mathbb{B}^*$ stand for the closed unit balls in the space and dual space in question. We use $\mathbb{R}^+$ to denote the interval $[0,\infty)$, $\overline{\mathbb{R}}$ the interval $(-\infty,\infty]$, and $\mathbb{R}^+\Omega := \{\alpha x |\, \alpha \ge 0,\ x \in \Omega\}$ for a given subset Ω of X.

Let $\bar{x} \in \Omega$ and $\varepsilon \ge 0$. Then

$$\widehat{N}_\varepsilon(\bar{x};\Omega) := \left\{ x^* \in X^* \,\middle|\, \limsup_{x \xrightarrow{\Omega} \bar{x}} \frac{\langle x^*, x-\bar{x}\rangle}{\|x-\bar{x}\|} \le \varepsilon \right\} \tag{2.1}$$

is the set of *ε-normals* to Ω at $\bar{x}$, where $x \xrightarrow{\Omega} \bar{x}$ means that $x \to \bar{x}$ with $x \in \Omega$. In particular, $\widehat{N}_0(\bar{x};\Omega)$ is a cone that is called the *prenormal cone* or the *Fréchet normal cone* to Ω at $\bar{x}$ and is denoted by $\widehat{N}(\bar{x};\Omega)$ for simplicity. When Ω is convex, $\widehat{N}(\bar{x};\Omega)$ reduces to the normal cone of convex analysis.

Given a multifunction $F\colon X \rightrightarrows X^*$, the expression

$$\begin{aligned}\operatorname*{Lim\,sup}_{x\to\bar{x}} F(x) := \{x^* \in X^* |\, &\exists \text{ sequences } x_k \to \bar{x} \text{ and } x_k^* \xrightarrow{w^*} x^* \\ &\text{with } x_k^* \in F(x_k) \text{ for all } k \in \mathbb{N}\}\end{aligned}$$

signifies the *sequential* Painlevé–Kuratowski upper (outer) limit with respect to the norm topology in X and the weak* topology w^* in X^*.

Now we define the (basic, limiting) *normal cone* to Ω at $\bar{x} \in \Omega$ by

$$N(\bar{x};\Omega) := \operatorname*{Lim\,sup}_{x \xrightarrow{\Omega} \bar{x}, \varepsilon \downarrow 0} \widehat{N}_\varepsilon(x;\Omega). \tag{2.2}$$

If X is Asplund, this is equivalent to

$$N(\bar{x};\Omega) = \operatorname*{Lim\,sup}_{x \xrightarrow{\Omega} \bar{x}} \widehat{N}(x;\Omega). \tag{2.3}$$

For an extended-real-valued lower semicontinuous (l.s.c.) function $f\colon X \to \overline{\mathbb{R}}$, the *Fréchet subdifferential* $\widehat{\partial} f(\bar{x})$ at $\bar{x} \in \operatorname{dom} f$ is given by

$$\widehat{\partial} f(\bar{x}) := \left\{ x^* \in X^* \,\middle|\, \liminf_{x\to\bar{x}} \frac{f(x) - f(\bar{x}) - \langle x^*, x-\bar{x}\rangle}{\|x-\bar{x}\|} \ge 0 \right\}$$

or, equivalently, by

$$\widehat{\partial} f(\bar{x}) = \left\{ x^* \in X^* \,\middle|\, (x^*,-1) \in \widehat{N}((\bar{x}, f(\bar{x})); \operatorname{epi} f) \right\}, \tag{2.4}$$

where $\operatorname{epi} f$ denotes the epigraph of f. For any $\bar{x} \in \operatorname{dom} f$, the sets

$$\partial f(\bar{x}) := \{x^* \in X^* \,|\, (x^*, -1) \in N((\bar{x}, f(\bar{x})); \operatorname{epi} f)\}, \tag{2.5}$$

$$\partial^\infty f(\bar{x}) := \{x^* \in X^* \,|\, (x^*, 0) \in N((\bar{x}, f(\bar{x})); \operatorname{epi} f)\} \tag{2.6}$$

are called, respectively, the *basic subdifferential* and the *singular subdifferential* of f at $\bar{x}$. We have $\partial^\infty f(\bar{x}) = \{0\}$ when f is Lipschitz continuous around $\bar{x}$. If X is Asplund, the following relations hold for any l.s.c. function f:

$$\partial f(\bar{x}) = \operatorname*{Lim\,sup}_{x \xrightarrow{f} \bar{x}} \widehat{\partial} f(x) \quad \text{and} \quad \partial^\infty f(\bar{x}) = \operatorname*{Lim\,sup}_{x \xrightarrow{f} \bar{x}, \lambda \downarrow 0} \lambda \widehat{\partial} f(x), \tag{2.7}$$

where $x \xrightarrow{f} \bar{x}$ means that $x \to \bar{x}$ with $f(x) \to f(\bar{x})$. Note that, due to (2.4)–(2.6), one has

$$\widehat{N}(\bar{x}; \Omega) = \widehat{\partial} \delta(\bar{x}; \Omega) \quad \text{and} \quad N(\bar{x}; \Omega) = \partial \delta(\bar{x}; \Omega) = \partial^\infty \delta(\bar{x}; \Omega)$$

for any $\Omega \subset X$ and $\bar{x} \in \Omega$, where $\delta(x; \Omega)$ is the indicator function of Ω.

Recall [21] that Ω is *sequentially normally compact* at $\bar{x} \in \Omega$ if for any sequences $x_k \xrightarrow{\Omega} \bar{x}$ and $x_k^* \in \widehat{N}(x_k; \Omega)$ with $x_k^* \xrightarrow{w^*} 0$ one has $\|x_k^*\| \to 0$. It always holds when Ω is compactly epi-Lipschitzian at $\bar{x}$ in the sense of [3] (see also [15] for a dual counterpart of the latter property), in particular, if either $X = \mathbb{R}^n$ or Ω is epi-Lipschitzian at $\bar{x}$ in the sense of [25]. A function $f \colon X \to \overline{\mathbb{R}}$ is *sequentially normally epi-compact* at $\bar{x} \in \operatorname{dom} f$ if $\operatorname{epi} f$ is sequentially normally compact at $(\bar{x}, f(\bar{x}))$. Note that f is sequentially normally epi-compact at $\bar{x}$ if it is Lipschitz continuous around this point.

The following basic calculus result for subdifferentials and its corollary for normal cones were proved in [20] by using the extremal principle; see below.

PROPOSITION 2.1. *Let X be an Asplund space and let $f_i \colon X \to \overline{\mathbb{R}}$, $i = 1, \dots, n$, be l.s.c. functions. Assume that all but one of these functions are sequentially normally epi-compact at a common point $\bar{x}$ of their domains. Suppose that the following qualification condition holds:*

$$\left[x_i^* \in \partial^\infty f_i(\bar{x}) \ (1 \le i \le n) \text{ and } \sum_{i=1}^n x_i^* = 0 \right] \Longrightarrow x_1^* = \dots = x_n^* = 0.$$

Then one has the subdifferential sum rule

$$\partial(f_1 + \dots + f_n)(\bar{x}) \subset \partial f_1(\bar{x}) + \dots + \partial f_n(\bar{x}).$$

In particular, if Ω_1 and Ω_2 are closed subsets of X such that one of them is sequentially normally compact at $\bar{x} \in \Omega_1 \cap \Omega_2$ and $N(\bar{x}; \Omega_1) \cap (-N(\bar{x}; \Omega_2)) = \{0\}$, then $N(\bar{x}; \Omega_1 \cap \Omega_2) \subset N(\bar{x}; \Omega_1) + N(\bar{x}; \Omega_2)$.

The next result follows from Proposition 2.1 and provides useful subdifferential representations of the basic normal cone to functional constraint sets given by equalities and inequalities.

PROPOSITION 2.2. *The following assertions hold in any Asplund space X.*

(a) *Let $\Omega := \{x \in X | \, f(x) \le 0\}$, where $f \colon X \to \overline{\mathbb{R}}$ is l.s.c. around $\bar{x} \in \Omega$. Suppose that there is no $\alpha \ne 0$ with $(0, \alpha) \in N((\bar{x}, 0); \operatorname{epi} f)$. Then*

$$N(\bar{x}; \Omega) \subset \{x^* \in X^* | \, (x^*, -\alpha) \in N((\bar{x}, 0); \operatorname{epi} f) \text{ for some } \alpha \ge 0\}, \tag{2.8}$$

which is equivalent to $N(\bar{x};\Omega) \subset \partial^\infty f(\bar{x}) \cup \mathbb{R}^+\partial f(\bar{x})$ *if* $f(\bar{x}) = 0$. *If, in addition,* f *is locally Lipschitzian around* $\bar{x}$ *with* $f(\bar{x}) = 0$, *then*

$$N(\bar{x};\Omega) \subset \bigcup_{\alpha\geq 0} \alpha\partial f(\bar{x}).$$

(b) *Let* $\Omega := \{x \in X |\, f(x) = 0\}$, *where* $f\colon X \to \mathbb{R}$ *is continuous around* $\bar{x} \in \Omega$. *Then*

$$N(\bar{x};\Omega) \subset \partial^\infty f(\bar{x}) \cup \partial^\infty(-f)(\bar{x}) \cup \mathbb{R}^+\partial f(\bar{x}) \cup \mathbb{R}^+\partial(-f)(\bar{x}) \tag{2.9}$$

if $0 \notin \partial f(\bar{x}) \cup \partial(-f)(\bar{x})$. *In particular, if* f *is locally Lipschitzian around* $\bar{x}$, *then*

$$N(\bar{x};\Omega) \subset \bigcup_{\alpha\geq 0} \alpha\Big(\partial f(\bar{x}) \cup \partial(-f)(\bar{x})\Big).$$

The main tool of our analysis in this paper is the *extremal principle*, which provides necessary conditions for set extremality in terms of a generalized Euler equation and can be treated as a local variational extension of the classical convex separation to systems of nonconvex sets. We refer the reader to [19] and the recent survey [18] for more information about the extremal principle and its various applications. The version of the extremal principle that we need in what follows is formulated in Theorem 3.1 of the next section, where it is used to derive the subdifferential variational principle. Now we recall that, given two subsets Ω_1 and Ω_2 of a Banach space X, a point $\bar{x} \in \Omega_1 \cap \Omega_2$ is *locally extremal* for the system $\{\Omega_1, \Omega_2\}$, named in this case an "extremal system," if there are sequences $a_{1k} \to 0$ and $a_{2k} \to 0$ in X and a neighborhood U of $\bar{x}$ such that

$$(\Omega_1 - a_{1k}) \cap (\Omega_2 - a_{2k}) \cap U = \emptyset \qquad \text{for all } k \in \mathbb{N}.$$

Various examples of extremal systems in variational analysis, optimization, and related topics can be found in [18] and in the rest of this paper.

3. Variational principles. In this section we establish the subdifferential variational principle as a characterization of Asplund spaces. In the next theorem we derive the subdifferential variational principle from a version of the extremal principle formulated below, and then we discuss its relationships with smooth variational principles in conventional support forms.

THEOREM 3.1. *Let* X *be a Banach space. Then the following assertions are equivalent:*

(a) (Extremal principle) *For every locally extremal point* $\bar{x}$ *of a closed set system* $\{\Omega_1, \Omega_2\}$ *in* X *and any* $\varepsilon > 0$ *there exist* $x_i \in \Omega_i \cap (\bar{x} + \varepsilon\mathbb{B})$ *and* $x_i^* \in \widehat{N}(x_i;\Omega_i) + \varepsilon\mathbb{B}^*$ $(i = 1, 2)$ *such that*

$$\|x_1^*\| + \|x_2^*\| = 1, \qquad x_1^* + x_2^* = 0.$$

(b) (Subdifferential variational principle) *For any l.s.c. function* $f\colon X \to \overline{\mathbb{R}}$ *bounded from below and every* $\varepsilon > 0$, $\lambda > 0$, *and* $\bar{x} \in \operatorname{dom} f$ *with* $f(\bar{x}) < \inf_X f + \varepsilon$ *there exist* $\widehat{x} \in X$ *and* $\widehat{x}^* \in \widehat{\partial} f(\widehat{x})$ *such that*

(i) $\|\widehat{x} - \bar{x}\| < \lambda$,

(ii) $f(\widehat{x}) < \inf_X f + \varepsilon$,

(iii) $\|\widehat{x}^*\| < \varepsilon/\lambda$.

(c) *X is an Asplund space.*

Proof. Implication (c)⇒(a) was first proved in [19]; another proof is given in [10]. Let us justify the remaining parts of the theorem. We begin with (b)⇒(c), and then derive (a)⇒(b), which is the main part.

(b)⇒(c). It is well known and easy to prove that a function $f\colon X \to \mathbb{R}$ is Fréchet differentiable at x if and only if the sets $\widehat{\partial} f(x)$ and $\widehat{\partial}(-f)(x)$ are nonempty simultaneously. Take an arbitrary convex continuous function $f\colon X \to \mathbb{R}$. Then $\widehat{\partial} f(x)$ agrees with the subdifferential of convex analysis and is nonempty at every $x \in X$. To establish the Asplund property of X, it is sufficient to show, due to [24, Proposition 1.25], that there is a dense subset $S \subset X$ such that $\widehat{\partial}(-f)(x) \neq \emptyset$ for every $x \in S$.

Denote $g(x) := -f(x)$ and fix $\bar{x} \in X$ and $\varepsilon > 0$. Since g is continuous, there exists a positive number $\widetilde{\varepsilon} < \varepsilon$ such that $g(x) > g(\bar{x}) - \varepsilon$ whenever $\|x - \bar{x}\| \le \widetilde{\varepsilon}$. Thus we have $h(\bar{x}) < \inf_X h(x) + 2\varepsilon$ for all $x \in X$, where $h(x) := g(x) + \delta(x; \bar{x} + \widetilde{\varepsilon}\mathbb{B})$ is obviously l.s.c. on X. Applying the subdifferential variational principle to the latter function, we find a point $\widehat{x} \in X$ with $\|\widehat{x} - \bar{x}\| < \widetilde{\varepsilon}$ such that $\widehat{\partial} h(\widehat{x}) \neq \emptyset$. This clearly implies that $\widehat{\partial} g(\widehat{x}) \neq \emptyset$, i.e., the set of points $x \in X$ with $\widehat{\partial}(-f)(x) \neq \emptyset$ is dense in X. Hence X must be Asplund.

(a)⇒(b). First we choose $\varepsilon_1 > 0$ such that $f(\bar{x}) < \inf_X f + \varepsilon - \varepsilon_1$ and let $\lambda_1 := (2\varepsilon)^{-1}(2\varepsilon - \varepsilon_1)\lambda$. Applying Ekeland's variational principle, we find $\widetilde{x} \in X$ satisfying $\|\widetilde{x} - \bar{x}\| < \lambda_1$, $f(\widetilde{x}) \le \inf_X f + \varepsilon - \varepsilon_1$, and

$$f(\widetilde{x}) < f(x) + \lambda_1^{-1}(\varepsilon - \varepsilon_1)\|x - \widetilde{x}\| \qquad \text{for all } \ x \in X \backslash \{\widetilde{x}\}. \tag{3.1}$$

Define two closed subsets of $X \times \mathbb{R}$ by

$$\Omega_1 := \operatorname{epi} f, \qquad \Omega_2 := \big\{(x, \mu) \in X \times \mathbb{R} \big| \, \mu \le f(\widetilde{x}) - \lambda_1^{-1}(\varepsilon - \varepsilon_1)\|x - \widetilde{x}\|\big\}.$$

It is easy to conclude from (3.1) that $(\widetilde{x}, f(\widetilde{x}))$ is a locally extremal point of the set system $\{\Omega_1, \Omega_2\}$; thus we can use the extremal principle.

Consider the norm $\|(x, \mu)\| := \|x\| + |\mu|$ on $X \times \mathbb{R}$ and observe that the corresponding dual norm on $X^* \times \mathbb{R}$ is given by $\|(x^*, \mu^*)\| = \max\{\|x^*\|, |\mu^*|\}$. Applying the extremal principle to the above system, for any $\varepsilon_2 > 0$ we find $(x_i, \mu_i) \in \Omega_i$ and $(x_i^*, \mu_i^*) \in \widehat{N}((x_i, \mu_i); \Omega_i)$, $i = 1, 2$, satisfying

$$\|x_i - \widetilde{x}\| + |\mu_i - f(\widetilde{x})| < \varepsilon_2, \quad i = 1, 2, \tag{3.2}$$

$$\frac{1}{2} - \varepsilon_2 < \max\{\|x_i^*\|, |\mu_i^*|\} < \frac{1}{2} + \varepsilon_2, \quad i = 1, 2, \tag{3.3}$$

$$\max\{\|x_1^* + x_2^*\|, |\mu_1^* + \mu_2^*|\} < \varepsilon_2. \tag{3.4}$$

Observe that $(x_2^*, \mu_2^*) \neq 0$ when ε_2 is sufficiently small. It follows from the structure of Ω_2 that $\mu_2 = f(\widetilde{x}) - \lambda_1^{-1}(\varepsilon - \varepsilon_1)\|x_2 - \widetilde{x}\|$, which yields $\mu_2^* > 0$ and thus implies that

$$\frac{x_2^*}{\mu_2^*} \in \widehat{\partial}\left(\lambda_1^{-1}(\varepsilon - \varepsilon_1)\|\cdot - \widetilde{x}\|\right)(x_2) \quad \text{and} \quad \frac{\|x_2^*\|}{\mu_2^*} \le \lambda_1^{-1}(\varepsilon - \varepsilon_1).$$

Taking (3.3) into account, the latter gives the estimate

$$\mu_2^* \ge \min\left\{\frac{(1 - 2\varepsilon_2)\lambda_1}{2(\varepsilon - \varepsilon_1)}, \frac{1}{2} - \varepsilon_2\right\}, \tag{3.5}$$

which ensures by (3.4) that $\mu_1^* < 0$ when ε_2 is sufficiently small. This allows us to show that $\mu_1 = f(x_1)$, since the opposite implies $\mu_1^* = 0$ due to $(x_1^*, \mu_1^*) \in \widehat{N}((x_1, \mu_1); \Omega_1)$ and the definition of Fréchet normals. Consequently, $-x_1^*/\mu_1^* \in \widehat{\partial} f(x_1)$.

It follows from (3.5) that $\varepsilon_2/\mu_2^* \to 0$ as $\varepsilon_2 \downarrow 0$. Putting all the above together, we conclude that

$$\frac{\|x_1^*\|}{|\mu_1^*|} < \frac{\|x_2^*\| + \varepsilon_2}{\mu_2^* - \varepsilon_2} = \left(\frac{\frac{\|x_2^*\|}{\mu_2^*} + \frac{\varepsilon_2}{\mu_2^*}}{1 - \frac{\varepsilon_2}{\mu_2^*}} \right) < \frac{\varepsilon}{\lambda}$$

when ε_2 is sufficiently small. On the other hand, it follows from (3.2) and the choice of λ_1 that

$$\|x_1 - \bar{x}\| < \lambda_1 + \varepsilon_2 \quad \text{and} \quad f(x_1) = \mu_1 < \inf_X f + \varepsilon - \varepsilon_1 + \varepsilon_2.$$

Finally, letting $\widehat{x} := x_1$ and $\widehat{x}^* := -x_1^*/\mu_1^*$, we get all the conclusions (i)–(iii) in (b) and finish the proof of the theorem. □

Remark 3.1. If f is smooth (Fréchet differentiable on its domain), then relation (iii) of the subdifferential variational principle reduces to $\|f'(\widehat{x})\| \le \varepsilon/\lambda$, which can be viewed as an approximate version of Fermat's stationary principle for suboptimal solutions in unconstrained optimization and was first obtained by Ekeland [7] in arbitrary Banach spaces. The next step was made by Rockafellar [25], who established, employing the sum rule for Clarke's generalized gradients, the corresponding generalized gradient version of assertion (b) in Theorem 3.1 for l.s.c. functions on Banach spaces. The same device, invoking the sum rule for two functions in Proposition 2.1, one of which is Lipschitz continuous, leads to the counterpart of Theorem 3.1(b) in terms of basic subgradients (2.5) in Asplund spaces; cf. [26, Proposition 10.44] in finite dimensions. Such a proof does not work for Fréchet subgradients in infinite dimensions; however, one can get the required conclusions of (b) in Asplund spaces by employing the "strong fuzzy sum rule" of Fabian [8, Theorem 3], which is actually equivalent to the extremal principle; see [19].

Remark 3.2. If X admits a Fréchet smooth renorm, conclusions (b) of Theorem 3.1 follow from the Borwein–Preiss smooth variational principle [2], which is equivalent, in this case, to the extremal principle by [1, Theorem 3.1]. Note that the *smooth renorm* assumption is not only sufficient but also *necessary* for the fulfillment of the smooth variational principle; see [9, Theorem 4.2]. On the other hand, the subdifferential variational principle is proved to be a characterization of Asplund spaces, and thus it can be viewed as an appropriate variational principle for this class of Banach spaces with no smoothness assumptions.

Remark 3.3. It follows from Theorem 4.6 in Fabian and Mordukhovich [9] that the subdifferential variational principle (b) in Theorem 3.1 directly implies *enhanced versions* of smooth variational principles in conventional support forms if the Asplund space X in question satisfies certain smoothness assumptions related to the existence of either smooth renorms or smooth bump functions of several types; see [9] for more details. Note that the combination of Theorem 3.1(b) and [9, Theorem 4.6] ensures the fulfillment of smooth variational principles that provide some additional information on supporting functions in comparison with the classical results of Borwein–Preiss and Deville–Godefroy–Zizler.

4. Necessary suboptimality and optimality conditions. Let us consider a general optimization problem of mathematical programming with equality, inequality,

and set constraints:

$$\text{(P)} \qquad \begin{cases} \text{minimize } f_0(x) \\ \text{subject to} \\ x \in \Delta \subset X, \\ f_i(x) \le 0, & i = 1, \dots, p, \\ f_i(x) = 0, & i = p+1, \dots, p+q, \end{cases}$$

where $f_i \colon X \to \overline{\mathbb{R}}$ are functions on a Banach space X. The main objective of this section is to present and discuss necessary conditions for *suboptimal solutions* to problem (P), proofs of which are given in the next section. In these results we do *not* assume the *existence of optimal solutions* that is an underlying assumption in the theory of necessary optimality conditions. The latter assumption is rather restrictive for problems of infinite-dimensional optimization and optimal control, while suboptimal solutions always exist. Based on the subdifferential variational principle, we find suboptimal solutions to (P) that *approximately* satisfy necessary conditions for optimality in problems with nonsmooth data. As a by-product of our approach, we derive refined necessary conditions for *optimal solutions* to (P), provided that they exist.

We obtain two types of results in this direction. Results of the first type justify "strong" necessary suboptimality and optimality conditions in both qualified and nonqualified forms for problems with Lipschitzian functions in terms of the basic normals and subgradients discussed in section 2. Proofs of these results essentially exploit the calculus rules and representations given in Propositions 2.1 and 2.2, as well as new subdifferential conditions for the sequential normal compactness of constraint sets derived in section 5. Independent results of the second type provide a "weak" form of necessary suboptimality and optimality conditions in terms of Fréchet normals and subgradients for problems with non-Lipschitzian data.

Let us consider the constraint sets

$$\Omega_i := \{x \in X | \, f_i(x) \le 0\}, \quad i = 1, \dots, p, \tag{4.1}$$

$$\Omega_i := \{x \in X | \, f_i(x) = 0\}, \quad i = p+1, \dots, p+q, \tag{4.2}$$

and denote by

$$\mathfrak{C} := \bigcap_{i=1}^{p+q} \Omega_i \cap \Delta, \qquad I(x) := \{i = 1, \dots, p+q | \, f_i(x) = 0\}$$

the sets of *feasible solutions* to (P) and *active constraint indices*, respectively. We always assume that $\mathfrak{C} \ne \emptyset$. To compress the statements of necessary suboptimality and optimality conditions formulated below, it is convenient to introduce the following sets of *generalized multipliers*.

DEFINITION 4.1. *Let $x \in \mathfrak{C}$ and $\mu \in \{0, 1\}$ be given. Then*

(a) *$M_\mu(x)$ denotes the collection of all tuples*

$$(x_0^*, \dots, x_{p+q}^*, x_\Delta^*, \alpha_1, \dots, \alpha_{p+q}) \in (X^*)^{2+p+q} \times [0, \infty)^{p+q} \tag{4.3}$$

satisfying the conditions

$$\begin{aligned} & x_0^* \in \partial f_0(x), \qquad x_\Delta^* \in N(x; \Delta), \\ & x_i^* \in \partial f_i(x) \quad \textit{for} \quad i \in \{1, \dots, p\} \cap I(x), \\ & x_i^* \in \partial f_i(x) \cup \partial(-f_i)(x) \quad \textit{for} \quad i = p+1, \dots, p+q, \\ & \alpha_i = 0 \quad \textit{for} \quad i \notin I(x), \end{aligned} \tag{4.4}$$

and

$$\mu x_0^* + \sum_{i\in I(x)} \alpha_i x_i^* + x_\Delta^* = 0. \tag{4.5}$$

(b) $M_\mu^\infty(x)$ *denotes the collection of all tuples* (4.3) *for which* (4.4) *and* (4.5) *hold, except that the inclusion required for* x_0^* *in* (4.4) *is replaced by* $x_0^* \in \partial^\infty f_0(x)$.

(c) *Given* $r > 0$, *let* $M_\mu(x; r)$ *denote the set of all tuples* (4.3) *obeying* (4.4) *and*

$$\left\| \mu x_0^* + \sum_{i\in I(x)} \alpha_i x_i^* + x_\Delta^* \right\| \le r, \qquad \mu + \sum_{i\in I(x)} \alpha_i \ge 1. \tag{4.6}$$

The next theorem gives *strong* necessary *suboptimality* conditions in both *nonqualified* (Fritz John) and *qualified* (Kuhn–Tucker) forms.

THEOREM 4.2. *Let* X *be an Asplund space. Assume that* Δ *is closed, that all of* $f_1, \dots, f_{p+q}$ *are locally Lipschitzian around each point of* $\mathfrak{C}$, *and that* $\inf_{\mathfrak{C}} f_0 > -\infty$. *Given an arbitrary* $\varepsilon > 0$, *let* $\bar{x} \in \mathfrak{C}$ *be an* ε*-optimal solution to* (P), *i.e.,* $f_0(\bar{x}) < \inf_{\mathfrak{C}} f_0 + \varepsilon$. *The following three assertions hold:*

(a) *Suppose that* f_0 *is also locally Lipschitzian around each point of* $\mathfrak{C}$. *Then for every* $\lambda > 0$ *there are* $\widehat{x} \in \mathfrak{C} \cap (\bar{x} + \lambda\mathbb{B})$ *and* $\mu \in \{0, 1\}$ *such that* $f_0(\widehat{x}) \le \inf_{\mathfrak{C}} f_0 + \varepsilon$ *and* $M_\mu(\widehat{x}; \varepsilon/\lambda) \ne \emptyset$.

(b) *Suppose that* f_0 *is l.s.c. on* $\mathfrak{C}$ *and that each* $x \in \mathfrak{C}$ *obeys both conditions* (i) *and* (ii):

  (i) *the basic qualification condition, namely, for every tuple of form* (4.3) *in* $M_1^\infty(x)$ *one has* $x_0^* = 0$, $x_\Delta^* = 0$, *and* $\alpha_i = 0$ *for all* $i \in I(x)$;

  (ii) *either* f_0 *is sequentially normally epi-compact at* x *or* Δ *is sequentially normally compact at* x *(in particular,* $\Delta = X$*).*

  Then for every $\lambda > 0$ *there is* $\widehat{x} \in \mathfrak{C} \cap (\bar{x} + \lambda\mathbb{B})$ *such that* $f_0(\widehat{x}) \le \inf_{\mathfrak{C}} f_0 + \varepsilon$ *and* $M_1(\widehat{x}; \varepsilon/\lambda) \ne \emptyset$.

(c) *Conversely, if the suboptimality conditions in* (b) *hold for any problem* (P) *with* $\mathfrak{C} = X$ *and an l.s.c. function* $f_0 \colon X \to \overline{\mathbb{R}}$ *concave on its domain, then* X *must be Asplund.*

Remark 4.1. If f_0 is locally Lipschitzian around every $x \in \mathfrak{C}$, then $\partial^\infty f_0(x) = \{0\}$ and the basic qualification condition in (i) is a *constraint qualification*. Moreover, if $\Delta = X$ and all f_i are *strictly differentiable* at x, then (i) is equivalent to the classical *Mangasarian–Fromovitz constraint qualification*:

(1) $f'_{p+1}(x), \dots, f'_{p+q}(x)$ are linearly independent.

(2) There exists $z \in X$ such that

$$\begin{aligned} &\langle f_i'(x), z\rangle < 0 \quad \text{for all} \quad i \in \{1, \dots, p\} \cap I(x), \\ &\langle f_i'(x), z\rangle = 0 \quad \text{for all} \quad i = p+1, \dots, p+q. \end{aligned}$$

In this smooth case, the suboptimality conditions of Theorem 4.2(b) reduce to those obtained by Ekeland [7] under a more restrictive constraint qualification. Namely, it was assumed in [7, Theorem 3.1] that *all* $\{f_i'(x) |\ i \in I(x)\}$ are linearly independent for each $x \in \mathfrak{C}$.

Next we present *necessary optimality* conditions in (P), which can be viewed as the limiting case of Theorem 4.2 as $\varepsilon = 0$. Note that $M_1(x; 0) = M_1(x)$.

THEOREM 4.3. *Let* X *be an Asplund space, and let* $\bar{x}$ *be an optimal solution to* (P). *Suppose that the corresponding assumptions in* (a) *and* (b) *of Theorem* 4.2 *are*

imposed only at $x = \bar{x}$. Then the conclusions of these assertions hold with $\varepsilon = 0$ for $\widehat{x} = \bar{x}$.

The last theorem of this section contains both necessary suboptimality and optimality conditions in problems (P) with *no* Lipschitzian, qualification, and/or sequential normal compactness assumptions. We obtain results in a *weak* approximate form of nonsmooth Lagrange multipliers with the replacement of relations (4.4)–(4.6) by their weaker counterparts.

To compress the statements of weak necessary suboptimality and optimality conditions in (P), it is convenient to introduce the following set. Given $\varepsilon > 0$, a weak* neighborhood V^* of the origin in X^*, and $x \in \mathfrak{C}$, we denote by $W_{\varepsilon,V^*}(x)$ the collection of all tuples

$$\begin{aligned} w &:= (x_0, \dots, x_{p+q}, x_\Delta, x_0^*, \dots, x_{p+q}^*, x_\Delta^*, \alpha_0, \dots, \alpha_{p+q}) \\ &\in X^{2+p+q} \times (X^*)^{2+p+q} \times [0,\infty)^{1+p+q} \end{aligned}$$

satisfying the conditions

$$\begin{aligned} &x_\Delta \in \Delta \cap (x + \varepsilon\mathbb{B}), \\ &f_0(x_0) - f_0(x)| \le \varepsilon, \quad \|x_i - x\| \le \varepsilon \quad \text{for } i = 0, \dots, p+q, \\ &x_\Delta^* \in \widehat{N}(x_\Delta;\Delta), \quad x_i^* \in \widehat{\partial} f_i(x_i) \quad \text{for } i = 0, \dots, p, \\ &x_i^* \in \widehat{\partial} f_i(x_i) \cup \widehat{\partial}(-f_i)(x_i) \quad \text{for } i = p+1, \dots, p+q, \end{aligned} \tag{4.7}$$

and

$$0 \in \sum_{i=0}^{p+q} \alpha_i x_i^* + x_\Delta^* + V^*, \qquad \sum_{i=0}^{p+q} \alpha_i = 1. \tag{4.8}$$

THEOREM 4.4. *Let X be an Asplund space, and let V^* be an arbitrary weak* neighborhood of the origin in X^*. The following assertions hold for problem* (P):

(a) *Assume that Δ is closed, that $f_0, \dots, f_p$ are l.s.c., and that $f_{p+1}, \dots, f_{p+q}$ are continuous around each $x \in \mathfrak{C}$. Suppose also that $\inf_{\mathfrak{C}} f_0 > -\infty$. Then there is a number $\bar{\varepsilon} > 0$ such that for every $0 < \varepsilon < \bar{\varepsilon}$ and every $\bar{x} \in \mathfrak{C}$ with $f_0(\bar{x}) < \inf_{\mathfrak{C}} f_0 + \varepsilon^2$ one has $W_{\varepsilon,V^*}(\bar{x}) \ne \emptyset$.*

(b) *Let $\bar{x}$ be an optimal solution to* (P). *Suppose that the assumptions in* (a) *are satisfied locally around $\bar{x}$. Then for every $\varepsilon > 0$ there is a $w \in W_{\varepsilon,V^*}(\bar{x})$, which obeys in addition the estimates*

$$|f_i(x_i) - f_i(\bar{x})| \le \varepsilon, \quad i = 1, \dots, p+q, \tag{4.9}$$

for the corresponding vectors x_i.

In the case of reflexive Banach spaces X, necessary optimality conditions in Theorem 4.4(b) reduce the main result of Borwein, Treiman, and Zhu [4, Theorem 2.1], obtained by a different technique based on "fuzzy" representations of the Fréchet normal cone, to the constraint sets (4.1) and (4.2). Recently these results were extended to the case of Asplund spaces by Ngai and Théra [23], who independently derived a version of Theorem 4.4(b) using Treiman's approach in [28] and a fuzzy chain rule for Fréchet subgradients. Furthermore, the result of [23] claims that all multipliers $\alpha_0, \dots, \alpha_{p+q}$ are nonzero, but it is actually equivalent to the statement of Theorem 4.4(b), since we can always cause them to be nonzero by small perturbations.

5. Proofs and auxiliary results. In this section we give proofs of the suboptimality and optimality results presented in section 4. These proofs apply the subdifferential variational principle of Theorem 3.1. Beyond that principle and the calculus rules given in section 2, our arguments strongly involve the following theorem ensuring the sequential normal compactness of the constraint sets (4.1) and (4.2). The proof of this theorem is based on the extremal principle.

THEOREM 5.1. *Let X be an Asplund space, and let $f\colon X \to \mathbb{R}$ be a function Lipschitz continuous around a given point $\bar{x} \in X$. Then the following assertions hold:*

(a) *The set $\Omega := \{x \in X | \, f(x) \le 0\}$ is sequentially normally compact at $\bar{x}$, provided that $f(\bar{x}) = 0$ and $0 \notin \partial f(\bar{x})$.*

(b) *The set $\Omega := \{x \in X | \, f(x) = 0\}$ is sequentially normally compact at $\bar{x}$, provided that $0 \notin \partial f(\bar{x}) \cup \partial(-f)(\bar{x})$.*

Proof. We prove both assertions (a) and (b) in a parallel way. In what follows, the set $\Theta \subset X \times \mathbb{R}$ stands for either $\operatorname{epi} f$ in (a) or $\operatorname{gph} f$ in (b). Choose arbitrary sequences $(x_k, x_k^*) \in X \times X^*$ such that $x_k \in \Omega$, $x_k^* \in \widehat{N}(x_k; \Omega)$ for all $k \in \mathbb{N}$, and $x_k \to \bar{x}$, $x_k^* \overset{w^*}{\to} 0$ as $k \to \infty$. It is required to prove that $\|x_k^*\| \to 0$ as $k \to \infty$. We are going to show that there exists a subsequence of $\{x_k^*\}$ with $\|x_k^*\| \to 0$. Since this result can be applied to any subsequence of $\{x_k^*\}$, it ensures the required convergence of the whole sequence.

Fix a sequence $\varepsilon_k \downarrow 0$ as $k \to \infty$. By the definition of $\widehat{N}(x_k; \Omega)$, we find a neighborhood U_k of x_k such that

$$\langle x_k^*, x - x_k \rangle - \varepsilon_k \|x - x_k\| \le 0 \qquad \text{for all } x \in U_k \cap \Omega. \tag{5.1}$$

Consider the sets

$$\begin{aligned} \Lambda_{1k} &:= \{(x, 0, \gamma) \in X \times \mathbb{R} \times \mathbb{R} | \, \gamma \ge 0\}, \\ \Lambda_{2k} &:= \{(x, \mu, \gamma) \in X \times \mathbb{R} \times \mathbb{R} | \, (x, \mu) \in \Theta, \\ &\qquad \gamma \le \langle x_k^*, x - x_k \rangle - \varepsilon_k (\|x - x_k\| + |\mu|)\}. \end{aligned}$$

Obviously these sets are closed in $X \times \mathbb{R} \times \mathbb{R}$ and $(x_k, 0, 0) \in \Lambda_{1k} \cap \Lambda_{2k}$. It follows from (5.1) that

$$\Lambda_{1k} \cap (\Lambda_{2k} - (0, 0, \nu)) \cap (U_k \times \mathbb{R} \times \mathbb{R}) = \emptyset \qquad \text{for all } \nu > 0.$$

This means that $(x_k, 0, 0)$ is a locally extremal point of the system $\{\Lambda_{1k}, \Lambda_{2k}\}$. Since X is an Asplund space, so is the product space $X \times \mathbb{R} \times \mathbb{R}$; see [24]. Thus we can apply the *extremal principle* in Theorem 3.1(a). Using this result, we find $(x_{ik}, \mu_{ik}, \gamma_{ik}) \in \Lambda_{ik}$ and

$$(x_{ik}^*, \mu_{ik}^*, \gamma_{ik}^*) \in \widehat{N}((x_{ik}, \mu_{ik}, \gamma_{ik}); \Lambda_{ik}) \quad \text{for } \ i = 1, 2, \tag{5.2}$$

satisfying the relations

$$\|x_{ik} - x_k\| + |\mu_{ik}| + |\gamma_{ik}| \le \varepsilon_k, \quad i = 1, 2, \tag{5.3}$$

$$\frac{1}{2} - \varepsilon_k \le \max\{\|x_{ik}^*\|, |\mu_{ik}^*|, |\gamma_{ik}^*|\} \le \frac{1}{2} + \varepsilon_k, \quad i = 1, 2, \tag{5.4}$$

$$\max\{\|x_{1k}^* + x_{2k}^*\|, |\mu_{1k}^* + \mu_{2k}^*|, |\gamma_{1k}^* + \gamma_{2k}^*|\} \le \varepsilon_k. \tag{5.5}$$

It easily follows from (5.2) as $i = 1$ that $x_{1k}^* = 0$ and $\gamma_{1k}^* \le 0$. Then (5.5) implies $\|x_{2k}^*\| \le \varepsilon_k$. Let us show that there exists a constant $c > 0$ such that

$$(x_{2k}^* + \gamma_{2k}^* x_k^*, \mu_{2k}^*) \in \widehat{N}_{c\varepsilon_k}((x_{2k}, \mu_{2k}); \Theta). \tag{5.6}$$

Using (5.2) with $i=2$ and the definition of $\widehat{N}$, we have

$$\limsup_{(x,\mu,\gamma)\overset{\Lambda_{2k}}{\to}(x_{2k},\mu_{2k},\gamma_{2k})} \frac{\langle x_{2k}^*, x-x_{2k}\rangle + \mu_{2k}^*(\mu-\mu_{2k}) + \gamma_{2k}^*(\gamma-\gamma_{2k})}{\|x-x_{2k}\|+|\mu-\mu_{2k}|+|\gamma-\gamma_{2k}|} \le 0, \tag{5.7}$$

where

$$\gamma_{2k} \le \langle x_k^*, x_{2k}-x_k\rangle - \varepsilon_k(\|x_{2k}-x_k\|+|\mu_{2k}|)$$

due to the construction of Λ_{2k}. If this inequality is strict, (5.7) implies that $\gamma_{2k}^*=0$ by letting $x=x_{2k}$, $\mu=\mu_{2k}$ and passing to the limit as $\gamma\to\gamma_{2k}$. Furthermore, setting $\gamma=\gamma_{2k}$ in (5.7), we get $(x_{2k}^*,\mu_{2k}^*)\in\widehat{N}((x_{2k},\mu_{2k});\Theta)$, which ensures (5.6) in this case. The other case of

$$\gamma_{2k} = \langle x_k^*, x_{2k}-x_k\rangle - \varepsilon_k(\|x_{2k}-x_k\|+|\mu_{2k}|)$$

is more difficult and can be handled as follows. We substitute

$$\gamma := \langle x_k^*, x-x_k\rangle - \varepsilon_k(\|x-x_k\|+|\mu|)$$

into (5.7) and find a neighborhood V_k of (x_{2k},μ_{2k}) such that

$$\frac{\langle x_{2k}^*, x-x_{2k}\rangle + \mu_{2k}^*(\mu-\mu_{2k}) + \gamma_{2k}^*(\gamma-\gamma_{2k})}{\|x-x_{2k}\|+|\mu-\mu_{2k}|+|\gamma-\gamma_{2k}|} \le \varepsilon_k \tag{5.8}$$

for all $(x,\mu)\in\Theta\cap V_k$. In this case

$$\begin{aligned}
&|\gamma-\gamma_{2k}-\langle x_k^*, x-x_{2k}\rangle| \\
&\quad = |-\varepsilon_k(\|x-x_k\|+|\mu|)+\varepsilon_k(\|x_{2k}-x_k\|+|\mu_{2k}|)| \\
&\quad \le \varepsilon_k(\|x-x_{2k}\|+|\mu-\mu_{2k}|), \\
&|\gamma-\gamma_{2k}| \le (1+\|x_k^*\|)(\|x-x_{2k}\|+|\mu-\mu_{2k}|).
\end{aligned}$$

Then (5.8) gives the estimate

$$\frac{\langle x_{2k}^*+\gamma_{2k}^*x_k^*, x-x_{2k}\rangle + \mu_{2k}^*(\mu-\mu_{2k})}{\|x-x_{2k}\|+|\mu-\mu_{2k}|} \le (2+|\gamma_{2k}^*|+\|x_k^*\|)\varepsilon_k \le c\varepsilon_k \tag{5.9}$$

for all $(x,\mu)\in\Theta\cap V_k$ and $c:=\sup_k\{2+|\gamma_{2k}^*|+\|x_k^*\|\}$. Note that $c<\infty$, since the sequence $\{x_k^*\}$ is w^*-convergent, and hence it is bounded due to the classical uniform boundedness theorem. Thus we get (5.6) from (5.9) and definition (2.1).

Note that the sequences of real numbers $\{\mu_{ik}^*\}$ and $\{\gamma_{ik}^*\}$, $i=1,2$, are bounded due to (5.4); hence we may assume that each of them converges. Due to (5.5) we have

$$-\lim_{k\to\infty}\mu_{1k}^* = \lim_{k\to\infty}\mu_{2k}^* := \bar{\mu}^*, \qquad -\lim_{k\to\infty}\gamma_{1k}^* = \lim_{k\to\infty}\gamma_{2k}^* := \bar{\gamma}^* \ge 0.$$

Let us show that $\bar{\mu}^*=0$ under the assumptions made. Indeed, assume the contrary and pass to the limit in (5.6) as $k\to\infty$. Taking into account definition (2.2) and also that $x_{2k}\to\bar{x}$ and $\mu_{2k}\to 0=f(\bar{x})$ by (5.3), we get $(0,\bar{\mu}^*)\in N((\bar{x},f(\bar{x}));\Theta)$ due to $x_k^*\overset{w^*}{\to}0$. This implies that $0\in\partial f(\bar{x})$ when $\Theta=\operatorname{epi} f$, and $0\in\partial f(\bar{x})\cup\partial(-f)(\bar{x})$ when $\Theta=\operatorname{gph} f$. Each of these conclusions contradicts the assumptions made in (a) and (b), respectively. Thus $\bar{\mu}^*=0$. Consequently, (5.4) and $x_{1k}^*=0$, $\|x_{2k}^*\|\le\varepsilon_k$ imply

that $\bar{\gamma}^* \neq 0$ and thus $\bar{\gamma}^* > 0$. Without loss of generality we assume that $\gamma^*_{2k} \geq d > 0$ for some constant d and all $k \in \mathbb{N}$.

To finish the proof, let us consider the following two cases for each $k \in \mathbb{N}$. (Only case (i) applies when $\Theta = \operatorname{gph} f$.)

Case (i). $\mu_{2k} = f(x_{2k})$. Substituting $\mu = f(x)$ and $\mu_{2k} = f(x_{2k})$ into (5.9), we get the estimates:

$$\begin{aligned}\frac{|\langle x^*_k, x - x_{2k}\rangle|}{\|x - x_{2k}\|} &\leq \frac{|\langle x^*_{2k}, x - x_{2k}\rangle|}{\gamma^*_{2k}\|x - x_{2k}\|} + \frac{|\mu^*_{2k}| \cdot |f(x) - f(x_{2k})|}{\gamma^*_{2k}\|x - x_{2k}\|}\\ &\quad + \frac{c\varepsilon_k}{\gamma^*_{2k}}\left(1 + \frac{|f(x) - f(x_{2k})|}{\|x - x_{2k}\|}\right)\\ &\leq \frac{\|x^*_{2k}\|}{d} + \frac{|\mu^*_{2k}|L}{d} + \frac{c\varepsilon_k}{d}(1 + L)\end{aligned}$$

for all x in a neighborhood of x_{2k}, where L is a Lipschitz modulus of f around $\bar{x}$. The latter yields

$$\|x^*_k\| \leq \frac{\|x^*_{2k}\|}{d} + \frac{|\mu^*_{2k}|L}{d} + \frac{c\varepsilon_k}{d}(1 + L). \tag{5.10}$$

Case (ii). $\mu_{2k} > f(x_{2k})$. (This only applies when $\Theta = \operatorname{epi} f$.) In this case $(x, \mu_{2k}) \in \Theta \cap V_k$ when x is near x_{2k}. Substituting $\mu = \mu_{2k}$ into (5.9), we get $\|x^*_{2k} + \gamma^*_{2k}x^*_k\| \leq c\varepsilon_k$. This implies the estimate

$$\|x^*_k\| \leq \frac{c\varepsilon_k}{d} + \frac{\|x^*_{2k}\|}{d}. \tag{5.11}$$

Summarizing both cases (i) and (ii), we see that for each $k \in \mathbb{N}$ either (5.10) or (5.11) holds. This finally implies that $\|x^*_k\| \to 0$ as $k \to \infty$, since $\|x^*_{2k}\| \leq \varepsilon_k$ for all $k \in \mathbb{N}$. The proof of the theorem is complete. □

Remark 5.1. The assumptions $0 \notin \partial f(\bar{x})$ and $0 \notin \partial f(\bar{x}) \cup \partial(-f)(\bar{x})$ are essential in Theorem 5.1. A corresponding counterexample is provided by the function $f(x) = \|x\|^2$ at $\bar{x} = 0$ for assertion (a) and by the function $f(x) = -\|x\|$ at the same point for assertion (b).

Remark 5.2. It is proved by Rockafellar [25] that if $0 \notin \overline{\partial} f(\bar{x})$, where $\overline{\partial}$ denotes the Clarke subdifferential, then the set $\Omega = \{x |\, f(x) \leq 0\}$ is epi-Lipschitzian at $\bar{x}$, provided that $f(\bar{x}) = 0$ and f is Lipschitz continuous around $\bar{x}$. It is well known that

$$\overline{\partial} f(\bar{x}) = \operatorname{cl}^* \operatorname{co} \partial f(\bar{x})$$

for locally Lipschitz functions in Asplund spaces; see [20, Theorem 8.11]. Thus, can we expect Ω to be epi-Lipschitzian at $\bar{x}$ under the weaker condition $0 \notin \partial f(\bar{x})$ ensuring the sequential normal compactness property of this set due to Theorem 5.1(a)? The answer is *negative* even in finite dimensions. A counterexample is provided by the function $f\colon \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x_1, x_2) := |x_1| - |x_2|$. One can check that

$$0 \notin \partial f(0,0) = \Big\{(x_1, x_2) \in \mathbb{R}^2 |\, |x_1| \leq 1 \text{ and either } x_2 = 1 \text{ or } x_2 = -1\Big\},$$

while the set $\{(x_1, x_2) \in \mathbb{R}^2 |\, |x_1| \leq |x_2|\}$ is obviously not epi-Lipschitzian at $(0, 0)$.

Proof of Theorem 4.2. Let us first prove assertion (b). Define an l.s.c. function f by

$$f(x) := f_0(x) + \delta(x; \mathfrak{C}) \qquad \text{for all } x \in X, \tag{5.12}$$

which is obviously bounded from below on X. Since $\inf_X f = \inf_{\mathfrak{C}} f_0$, we have $f(\bar{x}) < \inf_X f + \varepsilon$ for any $\bar{x}$ and $\varepsilon > 0$ given in the theorem. Applying the subdifferential variational principle of Theorem 3.1(b) to the function f in (5.12) with the given $\lambda > 0$, we find $\widehat{x} \in \mathfrak{C}$ and $\widehat{x}^* \in \partial f(\widehat{x})$ such that

$$\|\widehat{x} - \bar{x}\| < \lambda, \qquad f_0(\widehat{x}) = f(\widehat{x}) < \inf_X f + \varepsilon = \inf_{\mathfrak{C}} f_0 + \varepsilon, \qquad \|\widehat{x}^*\| < \varepsilon/\lambda. \tag{5.13}$$

Due to the structure of the feasible set $\mathfrak{C}$ in (P), we get from (5.12) that

$$\partial f(\widehat{x}) = \partial \left(f_0 + \sum_{i \in I(\mathbb{b})} \delta(\cdot; \Omega_i) + \delta(\cdot; \Delta) \right)(\widehat{x}), \tag{5.14}$$

with the sets Ω_i defined in (4.1) and (4.2). Now let us apply to (5.14) the basic calculus rules of Proposition 2.1 as well as the subdifferential conditions for the sequential normal compactness of the sets Ω_i obtained in Theorem 5.1. In this way, using assumptions (i) and (ii) and the normal cone representations of Proposition 2.2, we ensure that $M_1(\widehat{x}; \varepsilon/\lambda) \neq \emptyset$), which proves (b).

Next let us justify (a). By the proof of (b) we get $\widehat{x}$ and $\widehat{x}^* \in \partial f(\widehat{x})$ satisfying (5.13) and (5.14). If the basic qualification condition in (i) holds at $\widehat{x}$, then (a) follows, with $\mu = 1$, from the subsequent proof of (b). Suppose that (i) does not hold at $\bar{x}$. This means that there are numbers $\alpha_i \geq 0$ for $i \in I(\widehat{x})$, not all zero, and vectors $x^*_\Delta \in N(\widehat{x}; \Delta)$, $\widehat{x}^*_i \in \partial f_i(\widehat{x})$ for $i \in \{1, \ldots, p\} \cap I(\widehat{x})$, and $\widehat{x}^*_i \in \partial f_i(\widehat{x}) \cup \partial(-f_i)(\widehat{x})$ for $i = p+1, \ldots, p+q$ satisfying

$$\sum_{i \in I(\mathbb{b})} \alpha_i \widehat{x}^*_i + x^*_\Delta = 0.$$

Dividing the latter equality by $\alpha := \sum_{i \in I(\mathbb{b})} \alpha_i > 0$, we arrive at the conclusions of (a) with $\mu = 0$.

It remains to prove (c). Let X be a Banach space, and let $f \colon X \to \mathbb{R}$ be an arbitrary concave continuous function. Due to the continuity of f, for any $\bar{x} \in X$ and $\varepsilon > 0$ there is $0 < \varepsilon_1 < \varepsilon$ such that $f(\bar{x}) < f(x) + 2\varepsilon$ for all $x \in \bar{x} + \varepsilon_1 \mathbb{B}$. Consider the following unconstrained optimization problem of type (P):

$$\text{minimize} \quad f_0(x) \quad \text{on} \quad X, \quad \text{with} \quad f_0(x) := f(x) + \delta(x; \bar{x} + \varepsilon_1 \mathbb{B}), \tag{5.15}$$

where $f_0 \colon X \to \overline{\mathbb{R}}$ obviously satisfies the assumptions in (c). Applying to (5.15) the suboptimality conditions in (b), we find $\widehat{x} \in \bar{x} + \frac{\varepsilon_1}{2}\mathbb{B}$ such that $\partial f_0(\widehat{x}) = \partial f(\widehat{x}) \neq \emptyset$. It is well known (see [20]) that the basic subdifferential (2.5) admits the representation

$$\partial f(\widehat{x}) = \operatorname*{Lim\,sup}_{x \to \mathbb{b}, \gamma \downarrow 0} \widehat{\partial}_\gamma f(x)$$

for any continuous function f on a Banach space X, where

$$\widehat{\partial}_\gamma f(\bar{x}) := \left\{ x^* \in X^* \,\middle|\, \liminf_{x \to \bar{x}} \frac{f(x) - f(\bar{x}) - \langle x^*, x - \bar{x} \rangle}{\|x - \bar{x}\|} \geq -\gamma \right\}.$$

So for every $\gamma > 0$ there is a $x_\gamma \in \bar{x} + \varepsilon \mathbb{B}$ with $\widehat{\partial}_\gamma f(x_\gamma) \neq \emptyset$. This implies that, for any concave continuous function $f \colon X \to \mathbb{R}$ and any $\gamma > 0$, the set $\{x \in X |\ \widehat{\partial}_\gamma f(x) \neq \emptyset\}$

is dense in X. Now the Asplund property of X follows from [13, Proposition 1]. This ends the proof of the theorem. □

Proof of Theorem 4.3. Obviously $\bar{x}$ is an optimal solution to the problem of unconstrained minimization of the function $f\colon X \to \overline{\mathbb{R}}$ defined in (5.12). Due to Fermat's stationary principle in terms of the basic subdifferential (2.5), we have $0 \in \partial f(\bar{x})$. Let us first consider the case in which the assumptions of Theorem 4.2(b) hold around $\bar{x}$. Then, by Theorem 5.1, each of the sets $\Omega_1, \ldots, \Omega_{p+q}$ defined in (4.1) and (4.2) is sequentially normally compact at $\bar{x}$. Now employing Propositions 2.1 and 2.2, we get necessary optimality conditions of the form $M_1(\bar{x}; 0) \neq \emptyset$. Similarly to the proof of Theorem 4.2(a), these conditions imply the nonqualified form of $M_\mu(\bar{x}; 0) \neq \emptyset$ with $\mu \in \{0, 1\}$ when the basic qualification condition is not imposed. □

Proof of Theorem 4.4. First let us justify the weak necessary suboptimality conditions in (a). For any $z \in X$ and $\gamma > 0$ we consider a family of the w^*-neighborhoods

$$V^*(z; \gamma) := \big\{x^* \in X^* |\ |\langle x^*, z\rangle| < \gamma\big\}$$

of the origin in X^* that form a basis of the weak* topology. Taking an arbitrary w^*-neighborhood V^* in the theorem, we find constants $\bar{\gamma} > 0$, $l \in \mathbb{N}$ and vectors $z_j \in X$ with $\|z_j\| = 1,\ 1 \le j \le l$, such that

$$\bigcap_{j=1}^{l} V^*(z_j; 2\bar{\gamma}) \subset V^*.$$

Let us show that the conclusions of the theorem hold for every ε satisfying

$$0 < \varepsilon < \bar{\varepsilon} := \min\{\bar{\gamma}, 1\}.$$

Indeed, take any $\bar{x} \in \mathfrak{C}$ with $f_0(\bar{x}) < \inf_{\mathfrak{C}} f_0 + \varepsilon^2$, and find $\nu \in (0, \varepsilon)$ such that

$$f_0(\bar{x}) < \inf_{\mathfrak{C}} f_0 + (\varepsilon - \nu)^2.$$

Then for the function $f : X \to \overline{\mathbb{R}}$ defined in (5.12) one has

$$f(\bar{x}) < \inf_{X} f + (\varepsilon - \nu)^2.$$

Applying the subdifferential variational principle of Theorem 3.1(b), we find $\widehat{x} \in \mathfrak{C}$ and $\widehat{x}^* \in \widehat{\partial} f(\widehat{x})$ such that

$$\|\widehat{x} - \bar{x}\| < \varepsilon - \nu, \quad \|\widehat{x}^*\| < \varepsilon - \nu < \bar{\gamma}, \quad \text{and}$$

$$f_0(\widehat{x}) < \inf_{\mathfrak{C}} f_0 + (\varepsilon - \nu)^2 < \inf_{\mathfrak{C}} f_0 + \varepsilon - \nu.$$

The latter implies that

$$|f_0(\widehat{x}) - f_0(\bar{x})| < \varepsilon - \nu.$$

Now let us take $\gamma := \bar{\gamma}/(p + q + 1)$ and consider the w^*-neighborhood

$$\widehat{V}^* := \bigcap_{j=1}^{l} V^*(z_j; \gamma).$$

Given $\widehat{V}^*$ and ν, we employ the "weak fuzzy sum rule" of [8, Theorem 2] to $\widehat{x}^* \in \widehat{\partial} f(\widehat{x})$ and find

$$\begin{gathered} x_\Delta \in \Delta, \quad x_0 \in X, \quad y_i \in X \ \text{ for } \ i = 1, \dots, p+q, \\ \widehat{x}^*_\Delta \in \widehat{\partial}\delta(x_\Delta; \Delta) = \widehat{N}(x_\Delta; \Delta), \quad x^*_0 \in \widehat{\partial} f_0(x_0), \\ y^*_i \in \widehat{\partial}\delta(y_i; \Omega_i) = \widehat{N}(y_i; \Omega_i) \ \text{ for } \ i = 1, \dots, p+q \end{gathered}$$

satisfying the relations

$$\begin{gathered} \|x_0 - \widehat{x}\| < \nu, \quad |f_0(x_0) - f_0(\widehat{x})| < \nu, \quad \|x_\Delta - \widehat{x}\| < \nu, \\ \|y_i - \widehat{x}\| < \nu/2 \ \text{ for } \ i = 1, \dots, p+q, \quad \text{and} \\ \widehat{x}^* \in x^*_0 + \sum_{i=1}^{p+q} y^*_i + \widehat{x}^*_\Delta + \widehat{V}^*. \end{gathered}$$

To finish the proof, we need to consider the following two cases.

Case (1). There exist either $i \in \{1, \dots, p\}$ and $\alpha \neq 0$ satisfying $(0, \alpha) \in N((y_i, 0); \operatorname{epi} f_i)$ or $i \in \{p+1, \dots, p+q\}$ satisfying $0 \in \partial f_i(y_i) \cup \partial(-f_i)(y_i)$. If this happens for some $i \in \{1, \dots, p\}$, we use the basic normal cone representation (2.3) and find elements $(x_i, \mu_i) \in \operatorname{epi} f_i$ and $(x^*_i, -\mu^*_i) \in \widehat{N}((x_i, \mu_i); \operatorname{epi} f_i)$ with

$$\|x_i - y_i\| < \nu/2, \qquad \mu^*_i > 0, \qquad \text{and } x^*_i \in \mu^*_i V^*.$$

It is easy to check that $\widehat{N}((x_i, \mu_i); \operatorname{epi} f_i) \subset \widehat{N}((x_i, f_i(x_i)); \operatorname{epi} f_i)$, since $\mu_i \geq f_i(x_i)$. So we have $(x^*_i, -\mu^*_i) \in \widehat{N}((x_i, f_i(x_i)); \operatorname{epi} f_i)$, and hence

$$x^*_i/\mu^*_i \in \widehat{\partial} f_i(x_i) \quad \text{with} \quad x^*_i/\mu^*_i \in V^*.$$

If $i \in \{p+1, \dots, p+q\}$, we use the basic subdifferential representation (2.7) for continuous functions and find $x_i \in X$ and $x^*_i \in \widehat{\partial} f_i(x_i) \cup \widehat{\partial}(-f_i)(x_i)$ such that

$$\|x_i - y_i\| < \nu/2 \quad \text{and} \quad x^*_i \in V^*.$$

So in both situations of case (i) we get all the required relations of the theorem with $\alpha_i = 1$ (the other α_i are 0) and $x^*_\Delta = 0$.

Case (2). Otherwise. In this case we can use the normal cone representations (2.8) and (2.9), since the qualification assumptions of Proposition 2.2 hold. Employing (2.9) and then (2.7) for the equality constraints at y_i, we find

$$x_i \in X, \quad x^*_i \in \widehat{\partial} f_i(x_i) \cup \widehat{\partial}(-f_i)(x_i), \quad \text{and } \ \beta_i \geq 0$$

for $i = p+1, \dots, p+q$ satisfying

$$\|x_i - y_i\| < \nu/2 \quad \text{and} \quad \beta_i x^*_i \in y^*_i + \widehat{V}^*, \qquad 1 \leq i \leq p+q.$$

Now let us consider the inequality constraints and use representation (2.8) of the basic normal cone for each $i = 1, \dots, p$. Taking (y_i, y^*_i) above, we find $\mu^*_i \geq 0$ such that $(y^*_i, -\mu^*_i) \in N((y_i, 0); \operatorname{epi} f_i)$. Then using the limiting representation (2.3) of the basic normal cone in Asplund spaces, we approximate $(y^*_i, -\mu^*_i)$ in the weak* topology of $X^* \times \mathbb{R}$ by elements $(z^*_i, -r^*_i) \in \widehat{N}((z_i, r_i); \operatorname{epi} f_i)$, with (z_i, r_i) sufficiently close to $(y_i, 0)$. Without loss of generality we may assume that $r_i = f_i(z_i)$; cf. case (1). If

$r_i^* \neq 0$, we set $\beta_i := r_i^*$, $x_i := z_i$, and $x_i^* := z_i^*/\beta_i \in \widehat{\partial} f_i(x_i)$ to get the required relations. If $r_i^* = 0$, then, using the techniques developed in [14, proof of Theorem 4], we find an X^*-norm approximation $\beta_i x_i^*$ of z_i^* with some $\beta_i \geq 0$, and an X-norm approximation x_i of z_i such that $(x_i^*, -1) \in \widehat{N}((x_i, f_i(x_i)); \operatorname{epi} f_i)$, which means $x_i^* \in \widehat{\partial} f_i(x_i)$.

Combining all of the above relationships and taking into account that $\nu < \varepsilon < \bar{\gamma}$, we get

$$
\begin{gathered}
|f_0(x_0) - f_0(\bar{x})| < \varepsilon, \\
\|x_\Delta - \bar{x}\| < \varepsilon, \quad \|x_i - \bar{x}\| < \varepsilon \ \text{ for } \ i = 0, \dots, p+q, \\
0 \in x_0^* + \sum_{i=1}^{p+q} \beta_i x_i^* + \widehat{x}_\Delta^* + V^*.
\end{gathered}
$$

Finally, setting $\beta := 1/(1 + \sum_{i=1}^{p+q} \beta_i)$ and noting that $\beta V^* \subset V^*$, we arrive at (4.7) and (4.8) with

$$
x_\Delta^* := \beta \widehat{x}_\Delta^*, \quad \alpha_0 := \beta, \quad \text{and } \alpha_i := \beta\beta_i \qquad \text{for } \ i = 1, \dots, p+q.
$$

This justifies assertion (a) of the theorem.

It remains to prove the weak necessary optimality conditions in (b). Since $\bar{x}$ provides a local minimum to the function f in (5.12), we have $0 \in \widehat{\partial} f(\bar{x})$. Now following the (simplified) scheme in the proof of assertion (a), we arrive at all the conclusions of (b) except estimates (4.9) for $i = 1, \dots, p$. If $f_i(\bar{x}) = 0$ for some $i \in \{1, \dots, p\}$, then (4.9) follows directly from the lower semicontinuity of f_i at $\bar{x}$. Otherwise, if $f_i(\bar{x}) < 0$ for some $i \in \{1, \dots, p\}$, we substitute this constraint by $g_i(x) := f_i(x) - f_i(\bar{x}) \leq 0$ and observe that $\bar{x}$ is an optimal solution to the new problem with $g_i(\bar{x}) = 0$ and $\widehat{\partial} g_i(\bar{x}) = \widehat{\partial} f_i(\bar{x})$. Thus we get the desired necessary optimality conditions under the general assumptions of the theorem. □

Acknowledgments. The authors are grateful to M. Fabian, A. Kruger, and two anonymous referees for valuable suggestions and remarks that helped us to improve the original presentation.

REFERENCES

[1] J. M. Borwein, B. S. Mordukhovich, and Y. Shao, *On the equivalence of some basic principles in variational analysis*, J. Math. Anal. Appl., 229 (1999), pp. 228–257.

[2] J. M. Borwein and D. Preiss, *A smooth variational principle with applications to subdifferentiability and to differentiability of convex functions*, Trans. Amer. Math. Soc., 303 (1987), pp. 517–527.

[3] J. M. Borwein and H. M. Strojwas, *Tangential approximations*, Nonlinear Anal., 9 (1985), pp. 1347–1366.

[4] J. M. Borwein, J. S. Treiman, and Q. J. Zhu, *Necessary conditions for constrained optimization problems with semicontinuous and continuous data*, Trans. Amer. Math. Soc., 350 (1998), pp. 2409–2429.

[5] M. Bustos, *ε-Gradients pour functions localements Lipschitziennes et applications*, Numer. Funct. Anal. Optim., 15 (1994), pp. 435–453.

[6] R. Deville, G. Godefroy, and V. Zizler, *Smoothness and Renorming in Banach Spaces*, Wiley, New York, 1993.

[7] I. Ekeland, *On the variational principle*, J. Math. Anal. Appl., 47 (1974), pp. 324–353.

[8] M. Fabian, *Subdifferentiability and trustworthiness in the light of a new variational principle of Borwein and Preiss*, Acta Univ. Carolinae, 30 (1989), pp. 51–56.

[9] M. Fabian and B. S. Mordukhovich, *Nonsmooth characterizations of Asplund spaces and smooth variational principles*, Set-Valued Anal., 6 (1998), pp. 381–406.

[10] M. Fabian and B. S. Mordukhovich, *Separable reduction and extremal principles in variational analysis*, Nonlinear Anal., 49 (2002), pp. 265–292.
[11] A. Hamel, *An ε-Lagrange multiplier rule for a mathematical programming problem on Banach spaces*, Optimization, 49 (2001), pp. 137–150.
[12] R. Gabasov, F. M. Kirillova, and B. S. Mordukhovich, *The ε-maximum principle for suboptimal controls*, Soviet Math. Dokl., 27 (1983), pp. 95–99.
[13] A. D. Ioffe, *On subdifferentiability spaces*, N. Y. Acad. Sci., 410 (1983), pp. 107–119.
[14] A. D. Ioffe, *Proximal analysis and approximate subdifferentials*, J. London Math. Soc., 41 (1990), pp. 175–192.
[15] P. D. Loewen, *Limits of Fréchet normals in nonsmooth analysis*, in Optimization and Nonlinear Analysis, A. D. Ioffe et al., eds., Pitman Research Notes Math. Ser. 244, Longman, Harlow, UK, 1992, pp. 178–188.
[16] P. Loridan, *Necessary conditions for ε-optimality*, Math. Prog. Study, 19 (1982), pp. 140–152.
[17] B. S. Mordukhovich, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988.
[18] B. S. Mordukhovich, *The extremal principle and its applications to optimization and economics*, in Optimization and Related Topics, A. Rubinov and B. Glover, eds., Appl. Optim. 47, Kluwer, Dordrecht, The Netherlands, 2001, pp. 343–370.
[19] B. S. Mordukhovich and Y. Shao, *Extremal characterizations of Asplund spaces*, Proc. Amer. Math. Soc., 124 (1996), pp. 197–205.
[20] B. S. Mordukhovich and Y. Shao, *Nonsmooth sequential analysis in Asplund spaces*, Trans. Amer. Math. Soc., 348 (1996), pp. 1235–1280.
[21] B. S. Mordukhovich and Y. Shao, *Nonconvex differential calculus for infinite-dimensional multifunctions*, Set-Valued Anal., 4 (1996), pp. 205–236.
[22] M. Moussaoui and A. Seeger, *Epsilon-maximum principle of Pontryagin type and perturbation analysis of convex optimal control problems*, SIAM J. Control Optim., 34 (1996), pp. 407–427.
[23] H. V. Ngai and M. Théra, *A fuzzy optimality condition for non-Lipschitz optimization in Asplund spaces*, SIAM J. Optim., 12 (2002), pp. 656–668.
[24] R. R. Phelps, *Convex Functions, Monotone Operators and Differentiability*, 2nd ed., Lecture Notes in Math. 1364, Springer-Verlag, Berlin, 1993.
[25] R. T. Rockafellar, *Directionally Lipschitzian functions and subdifferential calculus*, Proc. London Math. Soc., 39 (1979), pp. 331–335.
[26] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[27] M. S. Sumin, *Optimal control of semilinear elliptic equations with state constraints: Maximum principle for minimizing sequences, regularity, normality, sensitivity*, Control and Cybernetics, 29 (2000), pp. 449–472.
[28] J. S. Treiman, *The linear nonconvex generalized gradient and Lagrange multipliers*, SIAM J. Optim., 5 (1995), pp. 670–680.
[29] R. B. Vinter, *Optimal Control*, Birkhäuser Boston, Cambridge, MA, 2000.
[30] L. C. Young, *Lectures on the Calculus of Variations and Optimal Control Theory*, Saunders, Philadelphia, 1969.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 2, pp. 641–658

SINGULARLY PERTURBED CONTROL SYSTEMS WITH ONE-DIMENSIONAL FAST DYNAMICS*

ZVI ARTSTEIN† AND ARIE LEIZAROWITZ‡

Abstract. The order reduction approach to singularly perturbed control systems suggests employing as a variational limit the differential algebraic system obtained when the small parameter is set to be zero. It is known that the method is valid only under restrictive convergence conditions on the fast dynamics. We verify in this paper that, when the fast state variable is one-dimensional, the order reduction method is valid in general. This is true, however, when appropriate relaxation is allowed in the reduced-order system. We also indicate how to extract near optimal solutions to the original system from optimal solutions of the order reduction one along the traditional reasoning of separating time scales. Examples are displayed, showing that, without allowing the relaxation, the order reduction may not provide the correct limit.

Key words. singular perturbations, order reduction, slow and fast motions, relaxed controls

AMS subject classifications. 49J15, 93C15, 34E15

PII. S0363012901390889

1. Introduction.

Consider the optimal control problem with singularly perturbed fast dynamics as follows:

$$\text{minimize} \quad \int_0^1 c(x(t), y(t), u(t))dt \tag{1.1}$$

subject to

$$\begin{aligned} \frac{dx}{dt} &= f(x, y, u), \\ \varepsilon \frac{dy}{dt} &= g(x, y, u), \end{aligned} \tag{1.2}$$

with initial conditions

$$x(0) = x_0, \quad y(0) = y_0, \tag{1.3}$$

where $x \in R^n$, $y \in R^m$, and $u \in R^k$ (see Remark 8.1 for extensions). The understanding is that the parameter ε multiplying the derivative of the y variable is small. Hence y is referred to as the fast variable. One is interested, in fact, in the limit behavior of the value and of the solutions to the problem as $\varepsilon \to 0$. This limit structure may be drawn from a limit problem. Hence we may try to identify an optimal control problem whose optimal value and solutions are limits, as $\varepsilon \to 0$, of the value and optimal solutions to the original system. Such a limit system is referred to as a variational limit.

*Received by the editors June 14, 2001; accepted for publication (in revised form) January 7, 2002; published electronically July 1, 2002.
http://www.siam.org/journals/sicon/41-2/39088.html

†Department of Mathematics, The Weizmann Institute of Science, Rehovot 76100, Israel (zvi.artstein@weizmann.ac.il). Incumbent of the Hettie H. Heineman Professorial Chair. The research of this author was supported by grants from the Israel Science Foundation, by the MINERVA Foundation, Germany, and by INTAS, Belgium.

‡Department of Mathematics, Technion, Haifa 32000, Israel (la@techunix.technion.ac.il). The research of this author was supported by the Fund for the Promotion of Research at the Technion.

The order reduction approach offers a candidate for a variational limit for the singularly perturbed system. It asserts that the behavior, when the small parameter tends to zero, is captured by the system arrived at when the parameter is set to be zero, namely, when (1.2) is replaced by

$$\begin{aligned} \frac{dx}{dt} &= f(x, y, u), \\ 0 &= g(x, y, u). \end{aligned} \tag{1.4}$$

The initial condition of the fast variable may not be compatible with the algebraic equation in (1.4); then a boundary layer is permitted.

The order reduction method yields suitable variational limits for a broad variety of situations, extending beyond optimal control problems. For the theory and many important applications, consult Kokotovic, Khalil, and O'Reilly [14]. See also Kokotovic and Khalil [13] and O'Malley [15]. However, for the order reduction method to apply, the optimal solutions have to satisfy quite restrictive conditions. In particular, on the fast scale, the dynamics has to converge to a stationary point. Many systems fail to have this property, and there are examples demonstrating that (1.4) may not be an appropriate variational limit for (1.2). Examples for this phenomenon, and alternative variational limits, were offered in Artstein [2], Artstein and Gaitsgory [3], [4], Gaitsgory [10], [11], and Vigodner [16].

In this paper, we establish, under mild conditions, that, for a one-dimensional fast variable y, namely a scalar, the system (1.4) is an appropriate variational limit for (1.2), provided that an appropriate relaxation of the order reduction system is allowed. No restrictions are put on the dimensions of the slow variable and the control variable. The relaxation is needed, however, for both the control variable and the stationary limit of the fast variable and the control. Indeed, the stationary solution of the reduced-order system serves as a control on the slow time scale; hence it may be subject to relaxation. In the particular case where the slow dynamics is not present, it is enough to employ relaxed controls, and, moreover, a stationary solution exists. A similar observation concerning a one-dimensional fast variable holds and is easy to verify for the uncontrolled version (see Remark 8.2). The result for the optimal control problem, even without slow dynamics, is not that apparent. Indeed, we provide examples (see Examples 8.3 and 8.4) demonstrating that, without allowing relaxed controls, the order reduction may not be a suitable variational limit.

Once an optimal solution to the relaxed version of the order reduction is detected, one can construct a near optimal solution to the original problem (1.1)–(1.3). This can be carried out along the traditional reasoning of separating the slow and fast components in the order reduction method. Consult Kokotovic, Khalil, and O'Reilly [14].

The paper is organized as follows. In the next section, we display some terminology and the technical assumptions. The interpretation of relaxed solutions for the order reduction problem is displayed in section 3, where the main general result is given. In section 4, we state the main result concerning the special case where there is no slow dynamics in the optimization process. The information we get in this case is sharper than in the general case. Two auxiliary lemmas are stated and verified in section 5. They are independent of optimality considerations. They capture and reveal, however, the role of the condition that y is a scalar variable. The proofs of the two main results, the special case with no slow dynamics and the general coupled

dynamics, are given in sections 6 and 7, respectively. In the closing section, we display some comments, extensions, and examples. In particular, counterexamples are given (Examples 8.5 and 8.7), showing that the relaxation of the fast stationary limit must be allowed, and results are exhibited (Proposition 8.9 and Example 8.10), addressing the situation where there is no need to use such a relaxation.

2. The setting. In this section, we set the terminology and display the conditions under which the main result is verified.

The control functions which are admissible for the optimal control system (1.1)–(1.3) are functions $u(\cdot) : [0,1] \to R^k$ which are Lebesgue measurable. We assume that, if there exists a solution to (1.2)–(1.3) when an admissible control $u(\cdot)$ is applied, this solution is unique. We denote by $\text{cost}_\varepsilon(u(\cdot))$ the cost of applying the admissible control function, namely, the outcome of the integral in (1.1) evaluated with the control function and with the resulting trajectories satisfying (1.2) and (1.3) for a given value of ε. We denote by $\text{val}(\varepsilon)$ the infimum of $\text{cost}_\varepsilon(u(\cdot))$ over all admissible controls.

Assumption 2.1.

(i) Continuity: The functions $f(x,y,u)$, $g(x,y,u)$, and $c(x,y,u)$ are continuous in their respective domains.

(ii) Uniqueness: Plugging a bounded and measurable control function $u(\cdot)$ in (1.2) with any initial conditions yields a unique solution $(x_\varepsilon(\cdot), y_\varepsilon(\cdot))$ on $[0,1]$.

(iii) Controllability of the fast flow: Consider the controlled equation $\frac{dy}{ds} = g(x,y,u)$ for x fixed and for $s \in [0,\infty)$. Any initial fast state y_1 can be steered to any other state y_2 by an admissible bounded control on some interval $[0,S]$.

(iv) Boundedness: There exists a uniformly bounded family of admissible controls $u_\varepsilon(\cdot)$, parameterized by ε, such that the resulting trajectories $(x_\varepsilon(\cdot), y_\varepsilon(\cdot))$ are also uniformly bounded, and such that $\text{val}(\varepsilon) - \text{cost}_\varepsilon(u_\varepsilon(\cdot))$ tends to zero as $\varepsilon \to 0$.

Conditions (i) and (ii) of Assumption 2.1 are standard. Condition (iii) is a simple one when y is scalar, as it assumes, roughly, that, for x and y fixed, the function $g(x,y,u)$ takes both positive and negative values. Condition (iv) can be derived from growth conditions on the cost function. It holds in broad classes of optimal control problems, let alone in practically all of the applications.

The term "near optimal control" is used throughout the paper, meaning that the near optimal control in question, which depends typically on a parameter, yields an arbitrarily good approximation to the optimal cost as the parameter tends to its limit.

3. The general result. In this section, we clarify how the order reduction system is solved and state the main result.

As mentioned already, the main result assumes that relaxed controls may be used. Relaxed controls were originated by Warga and are heavily employed in optimal control theory. For the theory, background, and applications of relaxed controls, consult Warga [17], Young [18], or Berkovitz [7]. Here we recall some essential facts. A relaxed control is a function which at each point t takes as a value a probability measure, say, μ, on the control space. We shall need two types of relaxation—one on the original control space and the second on the limit stationary fast dynamics, when considered as a control for the slow dynamics.

Consider first relaxed controls with values being probability measures on R^k, namely, the relaxation in the control space. The effect of the probability measure on

the right-hand side of (1.2) and on the cost function in (1.1) is via averaging. Namely, when the relaxed control μ is applied, these functions take the values

$$\int_{R^k} f(x,y,u)\mu(du), \quad \int_{R^k} g(x,y,u)\mu(du), \quad \text{and} \quad \int_{R^k} c(x,y,u)\mu(du),$$

which are denoted, respectively, by $f(x,y,\mu)$, $g(x,y,\mu)$, and $c(x,y,\mu)$. It is clear that a control value u can be regarded as the relaxed control measure supported on the singleton $\{u\}$. (As commented on in Remarks 6.3 and 7.3, for a one-dimensional y, it will be enough to consider relaxed controls whose values are supported at any given time on either 3 points or on $(n+3)^2$ points, depending on the problem.)

An admissible relaxed control is a measurable mapping $\mu(\cdot)$ defined on $[0,1]$ and with values in the space of probability measures on R^k. We denote this space by $\mathcal{P}(R^k)$. The space $\mathcal{P}(R^k)$ is endowed with the metric of weak convergence of measures (see, e.g., Billingsley [8]), namely, μ_i converge as $i \to \infty$ to μ_0 if

$$\int_{R^k} h(u)\mu_i(du) \to \int_{R^k} h(u)\mu_0(du),$$

as $i \to \infty$, holds for every bounded and continuous real valued function $h(\cdot)$. Convergence of relaxed controls is taken in the sense of weak convergence of measures over $[0,1] \times R^k$ (see Warga [17]). In particular, the relaxed controls $\mu_i(\cdot)$ converge as $i \to \infty$ to the relaxed control $\mu_0(\cdot)$ if the convergence

$$\int_0^1 \int_{R^k} h(t,u)\mu_i(t)(du)dt \to \int_0^1 \int_{R^k} h(t,u)\mu_0(t)(du)dt,$$

as $i \to \infty$, holds for every bounded and continuous real valued function $h(\cdot,\cdot)$.

The order reduction variational limit of (1.1)–(1.3), with relaxed controls, is as follows:

$$\text{minimize} \quad \int_0^1 c(x(t),y(t),\mu(t))dt \tag{3.1}$$

subject to

$$\begin{aligned} \frac{dx}{dt} &= f(x,y,\mu), \\ 0 &= g(x,y,\mu), \end{aligned} \tag{3.2}$$

with initial condition

$$x(0) = x_0. \tag{3.3}$$

We wish to emphasize the difference between (1.4) and (3.2); namely, in (3.2), relaxed controls are employed. Notice that the initial condition for the fast variable does not appear in the limit problem. Indeed, as in the standard order reduction case, the near optimal trajectories of the perturbed system may exhibit a boundary layer near $t = 0$.

In order to explain our interpretation of the order reduction system, we now display an equivalent formulation as follows.

We view the pairs (y, μ), which solve the algebraic equation in (3.2), as admissible controls for the differential equation in (3.2). To this end, we define

$$V(x) = \{(y, \mu) : 0 = g(x, y, \mu)\}. \tag{3.4}$$

It is clear from the continuity assumption that the sets $V(x)$ are closed and the set valued map $x \to V(x)$ has a closed graph. Using the notation v for the pair (y, μ), the order reduction problem can now be rephrased as follows:

$$\text{minimize} \quad \int_0^1 c(x(t), v(t))dt \tag{3.5}$$

subject to

$$\begin{aligned} \frac{dx}{dt} &= f(x, v), \\ v &\in V(x), \end{aligned} \tag{3.6}$$

with the initial condition (3.3).

An admissible trajectory of (3.2)–(3.3) (equivalently, of (3.6), (3.3)), is a triplet of functions, $(x(\cdot), y(\cdot), \mu(\cdot))$, from $[0, 1]$ to $(R^n \times R^m \times \mathcal{P}(R^k))$ (the space $\mathcal{P}(R^k)$ was described earlier) such that $x(0) = x_0$ and such that both the differential and the algebraic equations in (3.2) are satisfied (equivalently, (3.6) is satisfied). (See Remark 8.6.)

As in the standard optimal control theory, relaxed controls may be needed for solving (3.5)–(3.6). Recall that here the controls are pairs $v = (y, \mu)$ of solutions of the algebraic equation, and include, in particular, stationary points of the fast variable. Hence the relaxation amounts to employing probability measures, say, ν, on the set $V(x)$ (the effect being the convexification of this set). The meaning of $c(x, \nu)$ and $f(x, \nu)$ for the relaxation of the order reduction is drawn in the usual manner, namely,

$$c(x, \nu) = \int_{V(x)} c(x, y, \mu)\nu(dy \times d\mu), \quad f(x, \nu) = \int_{V(x)} f(x, y, \mu)\nu(dy \times d\mu). \tag{3.7}$$

With a relaxed control ν over $V(x)$, namely, over pairs (y, μ), we associate its effective distribution, say, $N(\nu)$ over $R^m \times R^k$. Namely, the probability measure $N(\nu)$ is given by

$$N(\nu)(C \times D) = \int_{C \times M} \mu(D)\nu(dy \times d\mu) \tag{3.8}$$

with M being the family of probability measures on R^k. It is clear that $c(x, \nu) = c(x, N(\nu))$ and $f(x, \nu) = f(x, N(\nu))$, which justifies the term "effective distribution."

An admissible relaxed trajectory of (3.5)–(3.6) is now a pair $(x(\cdot), \nu(\cdot))$ of measurable functions from $[0, 1]$ into the product of R^n and the collection of probability measures over the pairs (y, μ) which belong to $V(x(t))$; this is true for almost every t and such that (3.2) is satisfied with ν replacing (y, μ). (Equivalently, when in (3.6), $V(x)$ is replaced by its convex hull.)

The cost of an admissible trajectory $(x(\cdot), \nu(\cdot))$ is $\int_0^1 c(x(t), \nu(t))dt$, employing the term $c(x, \nu)$ given in (3.7) (compare with (3.1)). In view of the boundedness

assumption, Assumption 2.1(iv), we consider only relaxed trajectories with uniformly bounded supports. (In particular, $x(\cdot)$ is bounded and absolutely continuous.) The cost of the trajectory is denoted by $\text{cost}_R(x(\cdot), \nu(\cdot))$ (or by $\text{cost}_R(x(\cdot), y(\cdot), \mu(\cdot))$ if the trajectory is not a relaxed one). The subscript R stands for reduction.

As is customary, the infimum of all admissible costs is the value of the optimization problem. We need two notions. We denote by val(OR) the infimum over all admissible trajectories which use relaxation only in the control, and we denote by val(ROR) the infimum of the costs over all admissible relaxed trajectories. (Here OR and ROR stand for order reduction and relaxed-order reduction, respectively.)

In general, val(ROR) may be strictly less than val(OR). This is demonstrated in Example 8.7. In our application, it is val(ROR) which plays the key role (see also Remark 8.8). We display in Proposition 8.9 conditions under which the two values coincide.

We now state our main general result.

THEOREM 3.1. *Let Assumption* 2.1 *hold, and suppose that the fast variable y is a scalar. Then the following hold:*

(i) *The values val(ε) converge to val(ROR) as $\varepsilon \to 0$.*

(ii) *An optimal admissible relaxed trajectory of the order reduction problem exists.*

(iii) *For any admissible relaxed trajectory $(\bar{x}(\cdot), \bar{\nu}(\cdot))$, there exists a sequence $u_\varepsilon(\cdot)$ of control functions such that $\text{cost}_\varepsilon(u_\varepsilon(\cdot))$ converges to $\text{cost}_R(\bar{x}(\cdot), \bar{\nu}(\cdot))$ as $\varepsilon \to 0$, and the triplets $(x_\varepsilon(\cdot), y_\varepsilon(\cdot), u_\varepsilon(\cdot))$, resulting from the solution of* (1.2)–(1.3), *converge to $(\bar{x}(\cdot), \bar{\nu}(\cdot))$ as follows. The convergence of the slow dynamics component is uniform on $[0, 1]$. The pairs $(y_\varepsilon(\cdot), u_\varepsilon(\cdot))$ converge, in the sense of relaxed control, to the probability measure valued function $N(\bar{\nu}(\cdot))$. In particular, if the admissible relaxed solution $(x(\cdot), \nu(\cdot))$ is optimal for the order reduction system, then the controls $u_\varepsilon(\cdot)$ are near optimal for the original system when ε is small.*

4. The case of fast dynamics. As explained in the introduction, the results are sharper in the case where the slow variable x is absent from the control problem. This case is examined in the present section.

For completeness, we restate the optimal control problem without the slow variable. The problem then is as follows:

$$\text{minimize} \quad \int_0^1 c(y(t), u(t))dt \tag{4.1}$$

subject to

$$\varepsilon \frac{dy}{dt} = g(y, u), \tag{4.2}$$

with initial condition

$$y(0) = y_0. \tag{4.3}$$

The main result in this case takes a particular form as follows. Since the control problem is time-independent and the x variable, which acts as a parameter for the fast dynamics in the full problem, is absent, it is clear that the order reduction problem (4.1)–(4.3) reduces (see, however, Remark 6.2 and Examples 8.3 and 8.4) to the following optimization problem:

$$\text{minimize} \quad c(y, \mu) \tag{4.4}$$

subject to

$$0 = g(y, \mu), \tag{4.5}$$

where μ is a relaxed control. Notice that, again, the initial condition for the fast variable does not appear in the limit problem.

As in the full problem, we denote by val(ε) the optimal value of the problem (4.1)–(4.3), and we denote by val(OR) the optimal value of the problem (4.4)–(4.5). (In the present case, there is no need to use the relaxation of the fast variable; hence we do not refer to val(ROR).) The functions cost_ε and cost_R have the same meaning as in the full problem. When we refer to Assumption 2.1 in the present case, we interpret it with the absence of the x variable.

THEOREM 4.1. *Consider the optimal control problem* (4.1)–(4.3) *and the order reduction limit* (4.4)–(4.5), *both under Assumption* 2.1, *and when the fast variable y is one-dimensional. Then the following hold:*

(i) *val(ε) converges to val(OR) as $\varepsilon \to 0$.*

(ii) *An optimal solution $(\bar{y}, \bar{\mu})$ of the order reduction system exists.*

(iii) *For any solution $(\bar{y}, \bar{\mu})$ of the order reduction equation* (4.5), *there exists a bounded sequence $u_\varepsilon(\cdot)$ of control functions such that $\text{cost}_\varepsilon(u_\varepsilon(\cdot))$ converges to $\text{cost}_R(\bar{y}, \bar{\mu})$ as $\varepsilon \to 0$, and the pairs $(y_\varepsilon(\cdot), u_\varepsilon(\cdot))$, resulting from the solution of* (4.2)–(4.3), *converge to the constant function $(\bar{y}, \bar{\mu})$ as follows. The convergence of the y-component is uniform on any subinterval $[\delta, 1]$ with $\delta > 0$. The control functions $u_\varepsilon(\cdot)$ converge to $\bar{\mu}$ in the sense of relaxed controls. In particular, if $(\bar{y}, \bar{\mu})$ is an optimal solution of* (4.4)–(4.5), *then the control $u_\varepsilon(\cdot)$ forms a near optimal control for the original problem* (4.1)–(4.3) *when ε is small.*

5. Two key lemmas. In this section, we display two auxiliary results which hold the key to the main results of the paper, as they capture the role played by y being one-dimensional. The results are similar and have similar proofs, yet it is easier to provide independent proofs than to reduce one case to the other. The two results are independent of the optimality considerations of the singularly perturbed problem.

We start with a construction as follows. Let $(y_i(\cdot), u_i(\cdot))$ be a uniformly bounded family of functions from $[0, 1]$ into $R^m \times R^k$. (Indeed, we think of them as the fast trajectories and controls of a singularly perturbed system.) For each i, let P_i be the distribution of the mapping $(y_i(\cdot), u_i(\cdot))$ in $R^m \times R^k$. Namely, P_i is the probability measure given by

$$P_i(B) = \lambda(\{t : (y_i(t), u_i(t)) \in B\}) \tag{5.1}$$

for every Borel set B and where λ is the Lebesgue measure on the unit interval. Assume that the measures P_i converge in the sense of weak convergence of probability measures. Let the limit probability measure of P_i on $R^m \times R^k$ be denoted by P.

Let P^1 be the marginal measure of P on the fast coordinate space R^m; namely,

$$P^1(C) = P(C \times R^k) \tag{5.2}$$

for all Borel sets $C \subseteq R^m$. The superscript 1 indicates that the marginal is taken on the first coordinate, but also recall that we assume that, and later use, $m = 1$.

It is clear that P^1 is a probability measure on R^m. Let $\mu(\cdot)$ be the disintegration of P with respect to P^1. Namely, for P^1-almost every y, the measure $\mu(y)$ is a

probability measure on R^k, which depends measurably on y, and such that, for Borel subsets C of R^m and D of R^k,

$$P(C \times D) = \int_C \mu(y)(D) P^1(dy). \tag{5.3}$$

The construction is a standard one; see, e.g., Ash [6]. It is valid for any finite-dimensional fast dynamics. The following result depends on y being one-dimensional.

LEMMA 5.1. *Suppose that the variable y is scalar, and suppose that, for every i, the pair $(y_i(\cdot), u_i(\cdot))$ satisfies the differential equation*

$$\varepsilon_i \frac{dy}{dt} = g(y, u) \tag{5.4}$$

with $\varepsilon_i \to 0$ and with $g(\cdot, \cdot)$ continuous. Then $g(y, \mu(y)) = 0$ for P^1-almost every y.

Proof. The condition $g(y, \mu(y)) = 0$ means that

$$\int_{R^k} g(y, u) \mu(y)(du) = 0. \tag{5.5}$$

Clearly, it is enough to prove that

$$\int_{y_1}^{y_2} \int_{R^k} g(y, u) \mu(y)(du) P^1(dy) = 0 \tag{5.6}$$

whenever $[y_1, y_2]$ is an interval in the one-dimensional space, and such that y_1 and y_2 are not atoms of the marginal measure P^1. Equivalently, it is enough to verify that

$$\int_{[y_1, y_2] \times R^k} g(y, u) P(dy \times du) = 0 \tag{5.7}$$

for such intervals. Denote

$$J_i = \{t : (y_{\varepsilon_i}(t), u_{\varepsilon_i}(t)) \in [y_1, y_2] \times R^k\}. \tag{5.8}$$

The definition of P as the limit of the distributions P_i and the continuity of the function $g(y, u)$ together with the boundedness assumption on the controls and trajectories imply that the left-hand side of (5.7) is the limit, as $i \to \infty$, of

$$\int_{J_i} g(y_i(t), u_i(t)) dt. \tag{5.9}$$

For a fixed i, we divide the set J_i into three parts and perform the integration in (5.9) on each part separately.

Let $J_{i,1}$ be the subset of points t in J_i which do not belong to an interval, say, $[t_1, t_2]$, included in J_i. Then, for $t \in J_{i,1}$ (except possibly for $t = 0$ and $t = 1$), the derivative of $y_i(\cdot)$ at t, if it exists, must be equal to 0. Otherwise, on one side at least, the trajectory would enter the interval $[y_1, y_2]$. Since $(y_i(\cdot), u_i(\cdot))$ is a solution of the ordinary differential equation, this derivative exists λ-almost everywhere, and, since the differential equation is (5.4), it follows that $g(y_i(t), u_i(t)) = 0$ for λ-almost every t in $J_{i,1}$. Hence, when the integration (5.9) is performed on $J_{i,1}$, the value is 0.

Consider now a maximal interval $[t_1, t_2]$ included in J_i. Then $y_i(t_1)$ and $y_i(t_2)$ take as values either y_1 or y_2. Consider the case where $y_i(t_1) = y_i(t_2)$, say, where they both are equal to y_1. From (5.4) we deduce that

$$\int_{t_1}^{t_2} g(y_i(t), u_i(t))dt = \varepsilon_i y_1 - \varepsilon_i y_1 = 0 \tag{5.10}$$

and likewise when the common value is y_2. Let $J_{i,2}$ denote the union of such intervals in J_i. Then, when the integration (5.9) is performed on $J_{i,2}$, the value is again 0.

The rest of the set J_i (call it $J_{i,3}$) consists of intervals $[t_{1,j}, t_{2,j}]$ such that $y_i(t_{1,j}) \neq y_i(t_{2,j})$. In particular, there are only a finite number, say, r, of such intervals, and, moreover, if the index j signifies the order of these intervals, it follows that $y_i(t_{2,j}) = y_i(t_{1,j+1})$. In particular, when the integration (5.9) is performed on $J_{i,3}$, employing the differential equation (5.4), we get

$$\int_{J_{i,3}} g(y_i(t), u_i(t))dt = \sum_j \varepsilon_i(y_i(t_{2,j}) - y_i(t_{1,j})) = \varepsilon_i y_i(t_{2,r}) - \varepsilon_i y_i(t_{1,1}). \tag{5.11}$$

All in all, the integral (5.9) is the sum of the integrals on $J_{i,1}$, $J_{i,2}$, and $J_{i,3}$; namely, it is equal to the value exhibited in (5.11). Since the trajectories are uniformly bounded, it follows that, as $\varepsilon_i \to 0$, the integral (5.9) tends to zero. By the construction of J_i, the desired equality (5.7) is verified, and the proof is complete. ∎

Remark 5.2. The proof, specifically the arguments leading to (5.10) and (5.11), relies heavily on y being one-dimensional. The result does not hold in higher dimensions, as, e.g., Example 7.4 in [5], Example 10.2 in [3], or Example 10.1 in [4] show. A weaker result, however, does hold in higher dimensions, namely, that integrating $g(y, u)$ against P is equal to zero. This was established (in a slightly different context) in Artstein [1, (4.2)].

For the second auxiliary result, we need the following construction. We start again with a uniformly bounded family of functions $(y_i(\cdot), u_i(\cdot))$ from $[0, 1]$ into $R^m \times R^k$. We assume now that the functions converge in the sense of relaxed controls; namely, there exists a measure valued function $\eta(\cdot)$ on $[0, 1]$, with each value $\eta(t)$ being a probability measure on $R^m \times R^k$ such that

$$\int_0^1 h(t, y_i(t), u_i(t))dt \to \int_0^1 \int_{R^m \times R^k} h(t, y, u)\eta(t)(dy \times du)dt, \tag{5.12}$$

as $i \to \infty$, holds for every bounded and continuous real valued function $h(\cdot, \cdot, \cdot)$.

For a fixed t, we construct the marginals as in the preceding construction; namely, let $\eta^1(t)$ be the marginal of $\eta(t)$ on R^m, and let $\mu(t, \cdot)$ be the disintegration of $\eta(t)$ with respect to $\eta^1(t)$. (At this point, it is not clear how to get the measurable dependence of $\mu(\cdot, y)$, but, as we shall see, it will follow from the derivations.)

LEMMA 5.3. *Suppose that the variable y is scalar, and suppose that, for every i, the pair $(y_i(\cdot), u_i(\cdot))$ satisfies a differential equation*

$$\varepsilon_i \frac{dy}{dt} = g(x_i(t), y, u) \tag{5.13}$$

with $\varepsilon_i \to 0$, with $x_i(\cdot) : [0, 1] \to R^n$ being a uniformly converging prescribed sequence (say, the limit is $x_0(\cdot)$), and with $g(\cdot, \cdot, \cdot)$ continuous. Then, for λ-almost every t, the equality $g(x_0(t), y, \mu(t, y)) = 0$ holds for $\eta^1(t)$-almost every y.

Proof. We start as in the previous proof. Verifying $g(x_0(t), y, \mu(t, y)) = 0$ for t fixed amounts to

$$\int_{R^k} g(x_0(t), y, u)\mu(t, y)(du) = 0. \tag{5.14}$$

Clearly, it is enough to prove that

$$\int_{y_1}^{y_2} \int_{R^k} g(x_0(t), y, u)\mu(t, y)(du)\eta^1(t)(dy) = 0 \tag{5.15}$$

for intervals $[y_1, y_2]$ with a dense family of end points in the one-dimensional space. Equivalently, it is enough to verify that

$$\int_{[y_1,y_2]\times R^k} g(x_0(t), y, u)\eta(t)(dy \times du) = 0 \tag{5.16}$$

for such intervals. Since $\eta(\cdot)$ is measurable (and by this we overcome the difficulty of establishing that $\mu(\cdot, y)$ is measurable), it follows that, in order to verify the equality (5.16) for λ-almost every t, it is enough to verify that

$$\frac{1}{2\delta} \int_{t-\delta}^{t+\delta} \int_{[y_1,y_2]\times R^k} g(x_0(t), y, u)\eta(s)(dy \times du)ds \tag{5.17}$$

converges to 0 as $\delta \to 0$. The continuity of $g(\cdot, \cdot, \cdot)$ and the continuity of $x_0(\cdot)$ imply that it is enough to establish the convergence to 0 of the integration in (5.17) when the constant $x_0(t)$ is replaced by the function $x_0(\cdot)$. For convenience, we display the formula with this change:

$$\frac{1}{2\delta} \int_{t-\delta}^{t+\delta} \int_{[y_1,y_2]\times R^k} g(x_0(s), y, u)\eta(s)(dy \times du)ds. \tag{5.18}$$

Since the function $\eta(\cdot)$ is the relaxed control limit of the distributions of $(y_i(\cdot), u_i(\cdot))$, it follows from (5.12) and from Theorem 2.1(v) in Billingsley [8] that, if we assume that the points y_1, y_2 are such that the integral over $[t - \delta, t + \delta]$ of the η-measure of $\{y_1\} \times R^k$ is zero and likewise with y_2, we get that (5.18) is the limit as $i \to \infty$ of

$$\frac{1}{2\delta} \int_{J_i} g(x_0(s), y_i(s), u_i(s))ds \tag{5.19}$$

with

$$J_i = \{s \in [t - \delta, t + \delta] : (y_i(s), u_i(s)) \in [y_1, y_2] \times R^k\}. \tag{5.20}$$

It is clear that there exists a dense sequence of points in R with the properties demanded of y_1 and y_2. Since the function g is continuous, the convergence of $x_i(\cdot)$ to $x_0(\cdot)$ implies that the integral in (5.19) shares the same limit, as $i \to \infty$, with

$$\frac{1}{2\delta} \int_{J_i} g(x_i(s), y_i(s), u_i(s))ds. \tag{5.21}$$

We now show that the quantity (5.21) converges to 0 as $i \to \infty$. This, in view of the chain of arguments, would complete the proof.

To this end, we proceed in a way similar to the proof of the preceding lemma. Since the arguments are delicate, we provide the details as follows.

For a fixed i, we divide the set J_i into three parts and perform the integration in (5.21) on each part separately.

Let $J_{i,1}$ be the subset of points s in J_i which do not belong to an interval, say, $[s_1, s_2]$, included in J_i. Then, if the derivative $\frac{dy_i}{ds}(\cdot)$ exists at $s \in J_{i,1}$, it must be equal to 0 (except possibly for $s = t - \delta$ and $s = t + \delta$). Otherwise, on one side at least, the trajectory would enter the interval $[y_1, y_2]$. Since $(y_i(\cdot), u_i(\cdot))$ solves the differential equation (5.13), it follows that the derivative exists λ-almost everywhere, and then $g(x_i(s), y_i(s), u_i(s)) = 0$. Hence, when the integration (5.21) is performed on $J_{i,1}$, the value is 0.

Consider now a maximal interval $[s_1, s_2]$ included in J_i. Then $y_i(s_1)$ and $y_i(s_2)$ take as values either y_1 or y_2. Consider the case where $y_i(s_1) = y_i(s_2)$, say, where both are equal to y_1. From (5.13), we deduce that

$$\frac{1}{2\delta}\int_{s_1}^{s_2} g(x_i(s), y_i(s), u_i(s))ds = \varepsilon_i y_1 - \varepsilon_i y_1 = 0 \tag{5.22}$$

and likewise when the common value is y_2. Let $J_{i,2}$ denote the union of such intervals in J_i. Then, when the integration (5.21) is performed on $J_{i,2}$, the value is again 0.

The rest of the set J_i (call it $J_{i,3}$) consists of intervals $[s_{1,j}, s_{2,j}]$ such that $y_i(s_{1,j}) \neq y_i(s_{2,j})$. In particular, there are only a finite number, say, r, of such intervals in $[t - \delta, t + \delta]$, and, moreover, if the index j signifies the order of these intervals, it follows that $y_i(s_{2,j}) = y_i(s_{1,j+1})$. In particular, when the integration (5.21) is performed on $J_{i,3}$, employing the differential equation (5.13), we get

$$\begin{aligned}\frac{1}{2\delta}\int_{J_{i,3}} g(x_i(s), y_i(s), u_i(s))ds &= \sum_j \varepsilon_i(y_i(s_{2,j}) - y_i(s_{1,j})) \\ &= \varepsilon_i y_i(s_{2,r}) - \varepsilon_i y_i(s_{1,1}).\end{aligned} \tag{5.23}$$

All in all, the integral (5.21) is the sum of the integrals on $J_{i,1}$, $J_{i,2}$, and $J_{i,3}$; namely, it is equal to the value exhibited in (5.23). Since the trajectories are uniformly bounded, it follows that, as $i \to 0$, the integral (5.21) tends to zero. This verifies the desired convergence, and the proof is complete. □

6. Proof of Theorem 4.1. Let $u_\varepsilon(\cdot)$ be the uniformly bounded family of near optimal controls for ε small, guaranteed by Assumption 2.1(iv). Let $y_\varepsilon(\cdot)$ be the resulting trajectories of the fast variable. The assumption implies, in particular, that the pairs $(y_\varepsilon(\cdot), u_\varepsilon(\cdot))$ are uniformly bounded. We fix a subsequence ε_i such that

$$\liminf_{\varepsilon \to 0} \quad \text{val}(\varepsilon) = \lim_{\varepsilon_i \to 0} \quad \text{val}(\varepsilon_i). \tag{6.1}$$

We refer now to the construction in section 5 with $(y_i(\cdot), u_i(\cdot))$ being $(y_{\varepsilon_i}(\cdot), u_{\varepsilon_i}(\cdot))$. We may also choose the sequence such that the distributions P_i, as defined in (5.1), converge, say, to P. The continuity of the cost functional together with Assumption 2.1(iv) and (6.1) imply that

$$\liminf_{\varepsilon \to 0} \quad \text{val}(\varepsilon) = \int_{R \times R^k} c(y, u)P(dy \times du). \tag{6.2}$$

Let P^1 be the marginal of P on R, and let $\mu(\cdot)$ be the disintegration of P with respect to P^1, as was described at the beginning of section 5. Then

$$\int_{R\times R^k} c(y,u)P(dy\times du) = \int_R \int_{R^k} c(y,u)\mu(y)(du)P^1(dy). \tag{6.3}$$

The right-hand side of (6.3) can be written as $\int_R c(y,\mu(y))P^1(dy)$. Since, by Lemma 5.1, each pair $(y,\mu(y))$ in the latter integral satisfies $g(y,\mu(y))=0$, namely, it solves (4.5), it follows that

$$\liminf_{\varepsilon\to 0} \quad \mathrm{val}(\varepsilon) \geq \mathrm{val}(OR). \tag{6.4}$$

This verifies one direction of claim (i) in Theorem 4.1.

All of the values of $(y_\varepsilon(\cdot), u_\varepsilon(\cdot))$ are in one compact set, say, $\bar{B}$. Since the function $g(y,u)$ is continuous, it follows that the family of admissible pairs $(y,\mu(y))$ supported in $\bar{B}$ is compact. The continuity of $c(y,u)$ implies, therefore, that, among these pairs, a minimizer of $c(y,\mu)$ exists. Let $(\bar{y},\bar{\mu}(\bar{y}))$ be this minimizer. We show later that this minimizer is an optimal solution of the reduced-order problem (4.4)–(4.5).

At this point, we ignore the optimality property of $(\bar{y},\bar{\mu}(\bar{y}))$ and the fact that it is supported on $\bar{B}$ and regard it as a general admissible solution of (4.5). We now construct the family $u_\varepsilon(\cdot)$ promised in claim (iii) of the theorem. The controllability assumption, Assumption 2.1(iii), implies the existence of a bounded control function, say, $w(\cdot)$, defined on, say, the interval $[0,S]$, such that the solution of $\frac{dy}{ds} = g(y,w(s))$ with $y(0)=y_0$ satisfies $y(S)=\bar{y}$. We define $u_\varepsilon(t) = w(\varepsilon^{-1}t)$. It is clear then that the solution $y_\varepsilon(\cdot)$ of (4.2)–(4.3), resulting by applying this control function, satisfies $y_\varepsilon(\varepsilon S)=\bar{y}$. If relaxed controls were permissible on $[\varepsilon S,1]$, we could use $\bar{\mu}(\bar{y})$ on this interval, and claim (iii) of the theorem would be satisfied. However, since such relaxed controls are not allowed, we have to resort to the classical approximation of a relaxed control by ordinary controls (see, e.g., Warga [17]). In fact, given a fixed ε, for an arbitrary desired approximation, there exists an ordinary control function $u_\varepsilon(\cdot)$ on $[\varepsilon S,1]$ such that the resulting trajectory $y_\varepsilon(\cdot)$ stays close to $\bar{y}$ during $[\varepsilon S,1]$ and such that $u_\varepsilon(\cdot)$ approximates, as relaxed controls, the constant relaxed control $\bar{\mu}(\bar{y})$. The continuity of $c(y,u)$ implies then that $\mathrm{cost}_\varepsilon(u_\varepsilon(\cdot))$ approximates $\mathrm{cost}_R(\bar{y},\bar{\mu})$. This verifies the construction required by claim (iii) of the theorem.

The preceding construction implies, in particular, that

$$\limsup_{\varepsilon\to 0} \quad \mathrm{val}(\varepsilon) \leq \mathrm{val}(OR), \tag{6.5}$$

and, together with (6.4), claim (i) of the theorem is complete.

Recall that the specific pair $(\bar{y},\bar{\mu}(\bar{y}))$ was optimal only among those admissible relaxed controls supported on $\bar{B}$. However, the construction verifying claim (iii) of the theorem together with (6.2) imply that $\mathrm{cost}_R(\bar{y},\bar{\mu}(\bar{y}))$ is less than or equal to $\mathrm{val}(OR)$. In particular, the pair $(\bar{y},\bar{\mu}(\bar{y}))$ is actually optimal for the order reduction problem (4.4)–(4.5). Hence claim (ii) of the theorem is verified as well. This completes the proof of Theorem 4.1. □

Remark 6.1. In retrospect, (6.2) and (6.3) imply that, when the disintegration P_1 is carried out, P_1-almost every pair $(y,\mu(y))$ is an optimal solution of the order reduction problem (4.4)–(4.5). Indeed, the integral of their costs is equal to the infimal value of the cost.

Remark 6.2. The result of this section establishes the existence of an optimal solution to the order reduction problem (4.4)–(4.5) which is constant over $[0,1]$. The

case may be, however, that the minimization problem has more than one solution. In such a case, any measurable function $(y(t), \mu(t))$ which solves (4.4)–(4.5) pointwise can be regarded as a solution to the order reduction problem. In turn, such a solution would, in a manner similar to the above construction, yield near optimal solutions, which are not nearly constant, to the original singular perturbations problem.

Remark 6.3. The probability measure in the optimal pair $(\bar{y}, \bar{\mu}(\bar{y}))$ as established in the proof may be a general one. However, since it needs only to satisfy the two scalar equalities $g(\bar{y}, \bar{\mu}(\bar{y})) = 0$ and $c(\bar{y}, \bar{\mu}(\bar{y})) = \mathrm{val}(OR)$, it follows from standard arguments (see, e.g., Berkovitz [7]) that, for the same fast state $\bar{y}$, there exists an optimal probability measure which is supported on three points in R^k.

7. Proof of Theorem 3.1. We proceed along the reasoning of the proof of Theorem 4.1 in the preceding section but with the appropriate modifications.

Let $u_\varepsilon(\cdot)$ be the uniformly bounded family of near optimal controls for ε small guaranteed by Assumption 2.1(iv). Let $(x_\varepsilon(\cdot), y_\varepsilon(\cdot))$ be the resulting trajectories of the slow and fast variables. The assumption implies, in particular, that the pairs $(y_\varepsilon(\cdot), u_\varepsilon(\cdot))$ are uniformly bounded. We fix a subsequence ε_i such that

$$\liminf_{\varepsilon \to 0} \quad \mathrm{val}(\varepsilon) = \lim_{\varepsilon_i \to 0} \quad \mathrm{val}(\varepsilon_i). \tag{7.1}$$

We refer now to the construction preceding Lemma 5.3 in section 5, with $(y_i(\cdot), u_i(\cdot))$ being $(y_{\varepsilon_i}(\cdot), u_{\varepsilon_i}(\cdot))$. We may also choose the sequence such that the sequence $x_{\varepsilon_i}(\cdot)$ converges uniformly on $[0, 1]$, say, to $x_0(\cdot)$, and that $(y_i(\cdot), u_i(\cdot))$ converges in the sense of relaxed controls, say, to $\eta(\cdot)$, as defined in (5.12). The limit $(x_0(\cdot), \eta(\cdot))$ constitutes an admissible relaxed trajectory as described in section 3. The continuity of the cost functional together with Assumption 2.1(iv) and (7.1) imply that

$$\liminf_{\varepsilon \to 0} \quad \mathrm{val}(\varepsilon) = \int_0^1 \int_{R \times R^k} c(x_0(t), y, u) \eta(t)(dy \times du) dt. \tag{7.2}$$

Let $\eta^1(t)$ be the marginal of $\eta(t)$ on R, and let $\mu(t, \cdot)$ be the disintegration of $\eta(t)$ with respect to $\eta^1(t)$, as was described in section 5. Then, for every fixed t,

$$\int_{R \times R^k} c(x_0(t), y, u) \eta(t)(dy \times du) = \int_R \int_{R^k} c(x_0(t), y, u) \mu(t, y)(du) \eta^1(t)(dy). \tag{7.3}$$

The right-hand side of (7.3) can be written as $\int_R c(x_0(t), y, \mu(t, y)) \eta^1(t)(dy)$. In view of Lemma 5.3, each triplet $(x_0(t), y, \mu(t, y))$ in the latter integral satisfies $g(x_0(t), y, \mu(t, y)) = 0$; namely, it solves the algebraic equation in (3.2) for $\eta^1(t)$-almost every y. In particular, the measure $\eta^1(t)$ can be interpreted as a measure on the graph of the pairs $(y, \mu(y))$, and, as such, it constitutes a relaxation of admissible controls in $V(x_0(t))$; see (3.4). The equalities in (7.1), (7.2), and (7.3) imply now that

$$\liminf_{\varepsilon \to 0} \quad \mathrm{val}(\varepsilon) \geq \mathrm{val}(ROR). \tag{7.4}$$

This verifies one direction of claim (i) in Theorem 3.1. (Notice that we verified (7.4) with the value ROR only; this is unlike the case in Theorem 4.1, where equality with the value OR was established. Example 8.7 shows that (7.4) may not hold with $\mathrm{val}(OR)$.)

All of the values of $(x_\varepsilon(\cdot), y_\varepsilon(\cdot), u_\varepsilon(\cdot))$ are in one compact set, say, $\bar{D}$, and the trajectories $x_\varepsilon(\cdot)$ form a compact family. Since the function $g(x, y, u)$ is continuous, it

follows that the family of admissible pairs $\{(x,v) : v \in V(x)\}$, which are supported in the natural sense in $\bar{D}$, is compact. The continuity of $c(x,y,u)$ implies, therefore, that, among the admissible relaxed solutions $(x(\cdot),\nu(\cdot))$, there exists a minimizer of the cost functional. Let $(\bar{x}(\cdot),\bar{\nu}(\cdot))$ be this minimizer. We show later that this minimizer is an optimal solution of the reduced-order problem (3.1)–(3.3), with the relaxation as described in section 3.

At this point, we ignore the optimality property of $(\bar{x}(\cdot),\bar{\nu}(\cdot))$ and the fact that it is supported on $\bar{D}$, and we regard it as a general admissible relaxed solution. We now construct the family $u_\varepsilon(\cdot)$ promised in claim (iii) of the theorem.

Each $\bar{\nu}(t)$ is a probability measure on pairs $(y,\mu(y))$, which satisfy $g((\bar{x}(t),y,\mu(y))) = 0$. Recall the notion of the effective distribution of such probability measures, namely, $N(\nu)$ as defined in (3.8).

Consider for t fixed the differential equation

$$\frac{dy}{ds} = g(\bar{x}(t),y,u). \tag{7.5}$$

Claim 7.1. *The distribution $N(\bar{\nu}(t))$ can be approximated up to an arbitrarily desired approximation by the distribution of solutions $(y(\cdot),u(\cdot))$ to the differential equation* (7.5) *on some interval* $[0,S]$.

The proof provided here follows standard considerations (see, e.g., Warga [17]), employing the controllability in Assumption 2.1(iii) and the fact that $\bar{\nu}(t)$ is a probability measure on pairs which make the right-hand side of (7.5) equal to 0. (Without this property, the construction may not be possible; see Remark 7.2.) We can first approximate $\bar{\nu}(t)$ by a probability measure supported on a finite number of points, say, $(y_1,\mu_1),\dots,(y_r,\mu_r)$, with weights $p_1,\dots,p_r$ (see also Remark 7.3). Then on subsequent intervals we use the following strategy. At the beginning of the jth interval, we use the controllability to steer the trajectory to y_j. Then, for most of the interval, we use an arbitrarily close control approximation to the relaxed control $\mu_j(y_j)$. The fact that $g(\bar{x}(t),y_j,\mu_j) = 0$ implies that the solution $y(\cdot)$ will stay close to y_j on the rest of the interval. Now we take the lengths of these intervals in proportion to the weights p_j, and the claim is proved □

Notice that if, instead of (7.5), we use the original singularly perturbed equation in (3.2) and set the initial time to be t, the preceding construction will take place on the interval $[t,t+\varepsilon S]$. If t is chosen as a Lebesgue point of $N(\nu(\cdot))$, we get an approximation of the latter on a small interval. Given the relaxed trajectory $(\bar{x}(\cdot),\bar{\nu}(\cdot))$ and using the fact that the measure valued component is measurable (hence almost all its points are Lebesgue points), we can construct a piecewise approximation to the effective distribution $N(\bar{\nu}(t))$. Since the functions f and c are continuous and the measures have uniformly bounded support, the continuous dependence with respect to relaxed controls implies that the resulting slow trajectories will be uniformly close to $\bar{x}(\cdot)$. This verifies claim (iii) in Theorem 3.1.

The preceding construction implies, in particular, that

$$\limsup \quad \mathrm{val}(\varepsilon) \le \mathrm{val}(ROR), \tag{7.6}$$

and, together with (7.4), claim (i) of the theorem is complete.

Recall that the specific trajectory $(\bar{x}(\cdot),\bar{\nu}(\cdot))$ was optimal only among those admissible relaxed controls supported on $\bar{D}$. However, the construction verifying claim (iii) of the theorem together with (7.2) imply that $\mathrm{cost}_R(\bar{x}(\cdot),\bar{\nu}(\cdot))$ is less than or equal to $\mathrm{val}(ROR)$. In particular, the pair $(\bar{x}(\cdot),\bar{\nu}(\cdot))$ is actually optimal for the order

reduction problem (3.1)–(3.3). Hence claim (ii) of the theorem is verified as well. This completes the proof of Theorem 3.1 □

Remark 7.2. The approximation claimed in Claim 7.1 depends on the fact that the pairs $(y, \mu(y))$ satisfy the algebraic equation in the order reduction system. Controllability itself does not imply existence of an approximation to a given relaxed control, as the fast variable will not stay close to a constant on the long interval.

Remark 7.3. Along the reasoning of the argument in Remark 6.3, we note that the effective probability measure in the optimal admissible relaxed solution $(\bar{x}(\cdot), \bar{\nu}(\cdot))$, as established in the proof, may be a general one. However, since at each t only the values $g(\bar{x}(t), \bar{\nu}(t))$, $c(\bar{x}(t), \bar{\nu}(t))$, and $f(\bar{x}(t), \bar{\nu}(t))$ count, it follows from standard arguments (see, e.g., Berkovitz [7]) that an optimal relaxed control $\bar{\nu}(t)$ exists which, for each t, is supported on $n + 3$ points. These points are of the form (y, μ), where μ is a probability measure on R^k. For each y in the $n + 3$ points in the support, the values of $g(\bar{x}(t), y, \mu)$, $c(\bar{x}(t), y, \mu)$, and $f(\bar{x}(t), y, \mu)$ can be generated by $n + 3$ points in R^k. All in all, an optimal relaxed control can be found which, for each t, is supported on $(n + 3)^2$ points in R^k.

8. Extensions, examples, and remarks. In this section, we collect some remarks and examples concerning possible and impossible extensions of the main results.

Remark 8.1. It is clear that the choice of $[0, 1]$ as the integration domain in the problem is for convenience only, as any finite time interval will do. The restriction to a time-invariant equation is also done for convenience. If the cost and the right-hand side of (1.2) were also functions of t, we could add the time variable to the slow dynamics via the standard equation $\frac{dt}{dt} = 1$ and reduce the problem to a time-invariant one. Since no restrictions were put on the dimensionality of the slow variable, the results follow. The linear structure of the space R^k of controls is not used. It is enough to assume that $u \in U$, where U is a locally compact metric space (e.g., a closed, not necessarily convex, subset of R^k). A bounded set in U is then a set with a compact closure.

Remark 8.2. A particular case is where the dynamics does not depend on the u variable. Our result implies then that, when the variable y is one-dimensional, replacing the differential equation for the fast variable by the algebraic equation depicts the limit behavior of the solutions as $\varepsilon \to 0$. This particular observation follows easily from the approach displayed in Artstein and Vigodner [5]. In fact, what is proved there is that bounded fast solutions converge, as $\varepsilon \to 0$, to invariant measures of the fast equation (for a fixed slow state). It is easy to see that invariant measures of a one-dimensional equation are supported on equilibria.

Example 8.3. Arguments for using relaxed controls in the order reduction problem can be displayed along the reasoning of the traditional justification of relaxation. For instance, suppose that, in (4.1)–(4.3), $c(y, u) = y^2 + (1 - |u|)^2$, $g(y, u) = u$, and $y(0) = 0$. Then, already for the original singularly perturbed equation, the only optimal solution is a relaxed one, namely, the constant measure equally distributed over $\{-1, 1\}$. The next example is more telling.

Example 8.4. We display here an example where ordinary solutions for the perturbed equations exist, but their limit is not easily captured by the order reduction method. (Yet it can be interpreted along other methods suggested in the literature.) Furthermore, without relaxing the controls, the order reduction does not provide a solution at all. We also point out the solution yielded by the order reduction equation

with relaxation.

Consider the problem (4.1)–(4.3) with $c(y,u) = 0$ when $y \in [-2,2]$ and $|u| = 1$, and $c(y,u) > 0$ otherwise. Also, let $g(y,u) = u$ and $y(0) = 0$. It is easy to figure out an optimal control for each fixed ε. Indeed, the control function $u_\varepsilon(\cdot)$ should alternate between the values $+1$ and -1, with the switching occurring when the function $y_\varepsilon(\cdot)$ (which is the integral of $\varepsilon^{-1}u_\varepsilon(\cdot)$) reaches the values $+2$ and -2, respectively. The limit, as $\varepsilon \to 0$, of these solutions can be described in the language of invariant measures and limit occupational measures, as developed in Artstein [2], Artstein and Gaitsgory [3], [4], Gaitsgory and Leizarowitz [12], and Vigodner [16]. Indeed, the limit of the trajectories $(y_\varepsilon(\cdot), u_\varepsilon(\cdot))$ is an appropriate measure over $[-2,2] \times \{-1,1\}$. No single element in the support of this measure satisfies the algebraic equation in the order reduction system. In particular, this solution cannot be captured by the order reduction method. The present paper guarantees, however, that there exists a relaxed optimal solution of the order reduction system, for instance, the probability measure distributed equally on $\{-1,1\}$ applied to any stationary fast state in $[-2,2]$.

Example 8.5. The need for relaxation of the stationary fast dynamics in (1.1)–(1.3) in order to obtain an optimal solution can be derived from standard arguments. For instance, consider the problem with scalar variables, where the cost is $c(x,y,u) = x^2 + (|y|-1)^2$, and the equations are determined by $f(x,y,u) = y$ and $g(x,y,u) = u$. It is clear that an optimal solution to the reduced-order problem employs the measure on the y variable which is distributed equally on $y = 1$ and $y = -1$. This example does not demonstrate the need of relaxation in depicting the limit of the values, as here $\mathrm{val}(OR) = \mathrm{val}(ROR)$. This need is demonstrated later.

Remark 8.6. Notice that, unlike in the differential equations (1.2), the control (be it ordinary or relaxed) in the order reduction may not determine the dynamics; namely, given μ, the algebraic equation in (3.2) may have more than one solution. In particular, in the approximation, one has to specify how to drive the solution to the desired equilibrium or how to generate the chattering needed (e.g., in the preceding example) for the relaxed control approximation. In many applications, however, fixing the control determines the limit stationary state. See Kokotovic, Khalil, and O'Reilly [14]. One example for such a phenomenon is when $g(x,y,u) = -\alpha(x)y + h(x,u)$ with $\alpha(x)$ positive.

Example 8.7. We provide an example for $\mathrm{val}(ROR)$ strictly less than $\mathrm{val}(OR)$. It is a variant of known examples where using relaxed controls may strictly lower the cost. Consider the system with scalar variables where the cost is given by $c(x,y,u) = x^2 + (|y|-1)^2 + u^2$, the equations are determined by $\frac{dx}{dt} = y$ and $\varepsilon\frac{dy}{dt} = x + h(u)$ with initial conditions $x(0) = 0$, $y(0) = 0$, and where $h(u)$ is defined by $h(u) = 0$ for $u \in [-10,10]$, $h(u) = u - 10$ for $u > 10$, and $h(u) = u + 10$ for $u < 10$. The reduced equation (which in this example is independent of y) is $0 = x + h(u)$. A direct inspection reveals the optimal relaxed solution of the order reduction form. Indeed, the pairs $(y,u) = (1,0)$ and $(y,u) = (-1,0)$ are both admissible solutions of the algebraic equation, and relaxing these states with equal probabilities yields a trajectory with $x(t) = 0$ for every t and zero cost. In particular, $\mathrm{val}(ROR) = 0$. The conditions of the main result hold, and hence this relaxed optimal trajectory can be approximated by near optimal solutions of the original system.

If, however, we consider first trajectories generated by the solutions (y,μ) of the algebraic equation $0 = x + h(u)$, we see that, for $x \neq 0$, the only way to get a solution is to apply controls with absolute value greater than or equal to 10. Since, without

the relaxation of the stationary limits of the fast dynamics, the state $x = 0$ is not a solution, we deduce that val$(OR) \geq 100$.

Remark 8.8. We wish to point out that the relaxation of the order reduction system plays a role beyond guaranteeing existence of an optimal solution. Indeed, when val(OR) is strictly greater than val(ROR), there is no way to generate near optimal solutions to the perturbed system with the aid of nonrelaxed solutions of the order reduction system. As in the previous example, ordinary trajectories of the order reduction system cannot approximate optimal relaxed solutions. It is interesting to note that ordinary solutions of the perturbed system are able to approximate optimal relaxed solutions of the order reduction system.

The discrepancy between val(OR) and val(ROR) stems from the fact that the relaxation of the differential inclusion (3.6), which is determined by the set valued map with values $V(x)$, yields strictly lower values. Standard conditions are available in the literature, ensuring that this discrepancy does not occur. Since the case where one may operate with the order reduction problem (3.1)–(3.3) without alluding to relaxation may be of interest, we display conditions which guarantee that. To this end, consider the set valued map in the extended space as follows:

$$\hat{V}(x) = \{(c(x,v), f(x,v)) : v \in V(x)\}. \tag{8.1}$$

For a positive number b, let $\hat{V}_b(x)$ be the set of pairs $(c(x,v), f(x,v))$ in $\hat{V}(x)$ such that v is supported on the b-ball in $R^m \times R^k$. Distance between sets is taken as the Hausdorff distance.

PROPOSITION 8.9. *Let b be a bound on the sequence of near optimal trajectories guaranteed in Assumption* 2.1(iv). *Suppose that the mapping $x \to \hat{V}_b(x)$ is a Lipschitz function of x. Then val(OR)=val(ROR). Furthermore, if the sets $\hat{V}_b(x)$ are all convex sets, then an optimal solution to the order reduction problem can be found which does not use relaxation of the fast state variable.*

Proof. As is evident from the structure, the result alludes only to the structure of the differential inclusion. Hence the arguments go back to the fundamental observations of Filippov [9]. See Berkovitz [7, III.4 and IV.4]. □

Example 8.10. Conditions for the fulfillment of the properties listed in the previous result, thus guaranteeing that the order reduction system can be analyzed without the relaxation of the fast variable, can be demonstrated for large classes of equations. For instance, it is easy to verify that the Lipschitz property is satisfied when all of the functions involved are Lipschitz, and the equation for the fast variable is of the form $g(x,y,u) = h(x,y) + \beta(x)u$, with $\beta(x)$ not equal to 0. Indeed, the pairs (y,u) participating in the generation of $\hat{V}(x)$ are related by a Lipschitz factor. In the particular case where $g(x,y,u) = a(x)y + \beta(x)u$, with $a(x) \neq 0$ and $\beta(x) \neq 0$ (here u and $\beta(x)$ may be k-dimensional), the set valued function $\hat{V}(x)$ can be computed directly. Indeed, $\hat{V}(x) = \{(c(x, \frac{-1}{a(x)}\beta(x)\mu, \mu), f(x, \frac{-1}{a(x)}\beta(x)\mu, \mu)) : \mu \text{ is a relaxed control}\}$. Since $\hat{V}_b(x)$ is generated by controls in a bounded set, it is easy to see that, in this particular case, if the functions f, c, β, and a (note that a is continuous and $a(x) \neq 0$) are Lipschitz, then $\hat{V}_b(x)$ is Lipschitz continuous.

REFERENCES

[1] Z. ARTSTEIN, *Invariant measures of differential inclusions applied to singular perturbations*, J. Differential Equations, 152 (1999), pp. 289–307.

[2] Z. Artstein, *The chattering limit of singularly perturbed optimal control problems*, in Proceedings of the IEEE Conference on Decision and Control, Sydney, Australia, 2000, pp. 564–569.
[3] Z. Artstein and V. Gaitsgory, *Tracking fast trajectories along a slow dynamics: A singular perturbations approach*, SIAM J. Control Optim., 35 (1997), pp. 1487–1507.
[4] Z. Artstein and V. Gaitsgory, *Linear-quadratic tracking of coupled slow and fast targets*, Math. Control Signals Systems, 10 (1997), pp. 1–30.
[5] Z. Artstein and A. Vigodner, *Singularly perturbed ordinary differential equations with dynamic limits*, Proc. Roy. Soc. Edinburgh Sect. A, 126 (1996), pp. 541–569.
[6] R. B. Ash, *Real Analysis and Probability*, Academic Press, New York, 1972.
[7] L. D. Berkovitz, *Optimal Control Theory*, Springer-Verlag, New York, 1974.
[8] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1968.
[9] A. F. Filippov, *On certain questions in the theory of optimal control*, J. SIAM Control Ser. A, 1 (1962), pp. 76–84.
[10] V. Gaitsgory, *Suboptimization of singularly perturbed control systems*, SIAM J. Control Optim., 30 (1992), pp. 1228–1249.
[11] V. Gaitsgory, *Suboptimal control of singularly perturbed systems and periodic optimization*, IEEE Trans. Automat. Control, 38 (1993), pp. 888–903.
[12] V. Gaitsgory and A. Leizarowitz, *Limit occupational measures set for a control system and averaging of singularly perturbed control systems*, J. Math. Anal. Appl., 233 (1999), pp. 461–475.
[13] P. V. Kokotovic and H. K. Khalil, *Singular Perturbations in Systems and Control*, IEEE Press Selected Reprint Series, IEEE Press, New York, 1986.
[14] P. V. Kokotovic, H. K. Khalil, and J. O'Reilly, *Singular Perturbation Methods in Control: Analysis and Design*, Academic Press, London, 1986.
[15] R. E. O'Malley, Jr., *Singular perturbations and optimal control*, in Mathematical Control Theory, Lecture Notes in Math. 680, W. A. Coppel, ed., Springer-Verlag, Berlin, 1978, pp. 170–218.
[16] A. Vigodner, *Limits of singularly perturbed control problems with statistical dynamics of fast motions*, SIAM J. Control Optim., 35 (1997), pp. 1–28.
[17] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1976.
[18] L. C. Young, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, London, Toronto, 1969.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 3, pp. 659–681

SEMICONCAVE CONTROL-LYAPUNOV FUNCTIONS AND STABILIZING FEEDBACKS*

LUDOVIC RIFFORD†

Abstract. We study the general problem of stabilization of globally asymptotically controllable systems. We construct discontinuous feedback laws, and particularly we make it possible to choose these continuous outside a small set (closed with measure zero) of discontinuity in the case of control systems which are affine in the control; moreover this set of singularities is shown to be repulsive for the Carathéodory solutions of the closed-loop system under an additional assumption.

Key words. asymptotic controllability, control-Lyapunov function, feedback stabilization, nonsmooth analysis

AMS subject classifications. 93D05, 93D20, 93B05, 34D20, 49J52, 49L25, 70K15

PII. S0363012900375342

1. Introduction.

In a previous paper [23] we considered the stabilization problem for standard control systems. In particular, we proved that if a control system is globally asymptotically controllable, then one can associate to it a control-Lyapunov function which is semiconcave outside the origin. The goal of this article is to show the utility of the semiconcavity of such functions in the construction of stabilizing feedbacks.

We consider a standard control system of the general form $\dot{x} = f(x,u)$ which is globally asymptotically controllable, our objective being to design a feedback law $u : \mathbb{R}^n \to U$ such that the origin of the closed-loop system $\dot{x} = f(x,u(x))$ is globally asymptotically stable. Unfortunately, as pointed out by Sontag and Sussmann [28] and by Brockett [7], a continuous stabilizing feedback fails to exist in general. In addition to that, a smooth Lyapunov function may not exist either. As a matter of fact, although smooth Lyapunov-like techniques have been successfully used in many problems in control theory, it was shown by many authors (see Artstein [5] for the affine case, and more recently Clarke, Ledyaev, and Stern [11] for the general case) that there is no hope of obtaining a smooth Lyapunov function in the general case of globally asymptotically controllable systems. (The existence of such a function is indeed equivalent to that of a robust stabilizing feedback; see [17, 21].) These facts lead us to consider nonsmooth control-Lyapunov functions and particularly semiconcave control-Lyapunov functions; we proved the existence of such a function in our previous article [23]. This article builds on this result to derive a useful and direct construction of stabilizing feedbacks. In fact, the semiconcavity of the control-Lyapunov function allows us to give an explicit formula for the design of the stabilizing feedbacks. More particularly, this formula can be used in the context of control systems which are affine in the control to extend Sontag's formula [26] to the case of discontinuous feedback laws. Furthermore, the main result of this paper asserts that when the control system is affine in the control, we can design a feedback which is continuous on an open dense set and which stabilizes the closed-loop system in the sense of Carathéodory solutions. Surprisingly, we show that in this case, under an additional assumption on

*Received by the editors July 14, 2000; accepted for publication (in revised form) February 1, 2002; published electronically July 24, 2002.
http://www.siam.org/journals/sicon/41-3/37534.html
†Institut Girard Desargues, Université Lyon I, 43 Bd du 11 Novembre 1918, 69622 Villeurbanne Cedex, France (rifford@igd.univ-lyon1.fr).

the control-Lyapunov function, all the trajectories of the closed-loop system remain in the set of continuity for positive times; in other words, the set of singularities of the stabilized system is repulsive, and hence the feedback law is continuous (even locally Lipschitz) along the trajectories for $t > 0$.

Our paper is organized as follows: In section 2 we describe our main results. In section 3 we present some basic facts about nonsmooth analysis, semiconcavity, and discontinuous stabilizing feedbacks. In sections 4 and 5, we give the proofs of different results. Finally, the main theorems are proved in sections 6 and 7.

Throughout this paper, $\mathbb{R}_{\geq 0}$ denotes the nonnegative reals, $\|\cdot\|$ a norm on $\mathbb{R}^n$, B the open ball $B(0,1) := \{x : \|x\| < 1\}$ in $\mathbb{R}^n$, and $\overline{B}$ the closure of B.

2. Definitions and statements of the results.

2.1. General control systems. We study systems of the general form

$$\dot{x}(t) = f(x(t), u(t)), \tag{2.1}$$

where the state $x(t)$ takes values in a Euclidean space $\mathbb{R}^n$, the control $u(t)$ takes values in a given compact metric space U, and f is locally Lipschitz in x uniformly in u. We distinguish a special element "0" in U and assume that the state $x = 0$ is an equilibrium point (i.e., $f(0,0) = 0$). We are interested in globally asymptotically controllable systems, which we proceed to define.

DEFINITION 2.1. *The system* (2.1) *is globally asymptotically controllable (GAC) if there exists a nondecreasing function*

$$\tilde{\theta} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$$

such that $\lim_{r \to 0^+} \tilde{\theta}(r) = 0$, *with the property that, for each* $\xi \in \mathbb{R}^n$, *there exist a control* $u : R_{\geq 0} \to U$ *and a corresponding trajectory* $x(\cdot) : \mathbb{R}_{\geq 0} \to \mathbb{R}^n$ *such that* $x(0) = \xi$,

$$x(t) \to 0 \text{ as } t \to \infty,$$

and

$$\sup\{\|x(t)\| : 0 \leq t < \infty\} \leq \tilde{\theta}(|\xi|).$$

This definition of global asymptotic controllability is appropriate under the assumption of compactness of the control set U. When this set is not compact, we must add some conditions on the open-loop controls which stabilize the initial states; we refer to the papers of Sontag and Sussmann [29, 30] for a generalization of this definition to the general case on a noncompact control set.

REMARK 2.2. *This definition seems weaker than the one given initially in* [23]. *However, as explained by Sontag and Sussmann in* [29, 30], *a routine argument involving continuity of trajectories with respect to initial states shows that our different definitions are indeed equivalent.*

Our objective is to design a feedback law $u : \mathbb{R}^n \to U$ such that the origin of the closed-loop system (2.1) is globally asymptotically stable; that is, such that the new system

$$\dot{x}(t) = f(x(t), u(x(t))) \tag{2.2}$$

is globally asymptotically stable. Our method relies on nonsmooth Lyapunov functions, which we proceed to define; we refer to the next section for the definition of the proximal subdifferential $\partial_P V(\cdot)$.

DEFINITION 2.3. *A control-Lyapunov function for the system* (2.1) *is a continuous function* $V : \mathbb{R}^n \to \mathbb{R}$ *which is positive definite (i.e.,* $V(0) = 0$ *and* $V(x) > 0$ *for* $x \neq 0$*), proper (i.e.,* $V(x) \to \infty$ *when* $\|x\| \to \infty$*), and such that there exists a positive definite continuous function* $W : \mathbb{R}^n \longrightarrow \mathbb{R}$ *with the property that, for each* $x \in \mathbb{R}^n \setminus \{0\}$*, we have*

$$\forall \zeta \in \partial_P V(x), \quad \min_{u \in U} \langle \zeta, f(x,u) \rangle \leq -W(x). \tag{2.3}$$

The present article is based on the following theorem, which is a refinement of a result proved in [23]. The regularity of our control-Lyapunov function will be crucial for the construction of discontinuous stabilizing feedbacks.

THEOREM 2.4. *If the system* (2.1) *is GAC, then there exists a control-Lyapunov function* V *which is semiconcave on* $\mathbb{R}^n \setminus \{0\}$ *and such that*

$$\forall x \in \mathbb{R}^n \setminus \{0\}, \ \forall \zeta \in \partial_P V(x), \quad \min_{u \in U} \langle \zeta, f(x,u) \rangle \leq -V(x). \tag{2.4}$$

This theorem differs from the one given in our previous article [23] in the decrease condition (2.4). Here, we assert that we can take V as the function W of Definition 2.3. This special form of the infinitesimal decrease condition (2.4) will allow us to obtain exponential decrease for $V(x(t))$ and will make it possible to give closed-form estimates (in terms of V) on the rate of stabilization.

Now, using the concept of π-trajectories and of Euler trajectories which will be defined in the next section, we give a general result on the existence of stabilizing feedbacks; this result was announced in [25].

THEOREM 2.5. *Assume that the system* (2.1) *is GAC. Then there exists a feedback* $u : \mathbb{R}^n \to U$ *for which the system* $\dot{x} = f(x, u(x))$ *is globally asymptotically stabilizable in the sense of* π*-trajectories and in the Euler sense.*

Moreover, if we consider a control-Lyapunov function V *for the given system, then the stabilizing feedback can be designed as follows:*

- *We set* $u(0) = 0$.
- *For each* $x \in \mathbb{R}^n \setminus \{0\}$*, we choose arbitrarily* $\zeta \in \partial_L V(x)$ *and we set*

$$u := u(x) \in U, \ \textit{where } u(x) \textit{ is any point in } U \textit{ such that } \langle \zeta, f(x,u) \rangle \leq -W(x).$$

Furthermore, if the control-Lyapunov function V *is the one given by Theorem* 2.4 *(i.e., if* $W = V$*), then we have*

$$V(x(t)) \leq e^{-t} V(x_0) \tag{2.5}$$

for any Euler trajectory starting at $x_0 \in \mathbb{R}^n$.

The existence of a discontinuous feedback which is stabilizing in the sense of the π-trajectories is not new; it appeared initially in the article of Clarke et al. [10]. We refer also to Ancona and Bressan [4], who proved a slightly stronger result in the sense that their feedback stabilizes the closed-loop system in the sense of Carathéodory; their proof does not use nonsmooth control-Lyapunov functions. However, here the consideration of a semiconcave control-Lyapunov function leads to a simple proof and allows us to give an explicit formula for the design of the feedback. Moreover, we are able to design some stabilizing feedbacks which are rather regular in the case of affine systems.

2.2. Affine control systems. Let us assume now that the control system is affine in the control, that is,

$$f(x,u) = f_0(x) + \sum_{i=1}^{m} u_i f_i(x) \quad \forall (x,u) \in \mathbb{R}^n \times U, \tag{2.6}$$

where the $f_0, \dots, f_m$ are locally Lipschitz functions from $\mathbb{R}^n$ into $\mathbb{R}^n$ and where U is a strictly convex and compact set of $\mathbb{R}^m$.

REMARK 2.6. *Instead of assuming the control set U to be strictly convex, we could make a weaker assumption of convexity. As a matter of fact, if the control set is supposed to be convex, we can define a subset of it which is strictly convex and for which the control system* (2.6) *keeps the same properties of controllability. Consequently all our results hold in the case of a convex compact control set.*

First of all, assuming the knowledge of a control-Lyapunov function, we are able to give an explicit feedback law; it reduces to Sontag's formula [26] in the smooth case.

THEOREM 2.7. *Assume that V is a control-Lyapunov function for* (2.6) *and consider any selection $\zeta_V(\cdot)$ of $\partial_L V(\cdot)$. Then the feedback control defined by*

$$u_i(x) := -\phi\left(\langle f_0(x), \zeta_V(x)\rangle, \sum_{i=1}^{m} \langle f_i(x), \zeta_V(x)\rangle^2\right) \langle f_i(x), \zeta_V(x)\rangle,$$

where

$$\phi(a,b) = \begin{cases} \frac{a+\sqrt{a^2+b^2}}{b} & \text{if } b \neq 0, \\ 0 & \text{if } b = 0, \end{cases}$$

(globally asymptotically) stabilizes the control system (2.6) *in the sense of π-trajectories and in the Euler sense.*

REMARK 2.8. *The feedback given in Theorem* 2.7 *may not be with values in the control set U. However, we can project the values $u(x)$ on the unit ball $\bar{B}$ to get a stabilizing feedback which is locally bounded.*

Furthermore, we can exploit the semiconcavity property more strongly to derive some regularity properties on the feedback in the case of affine control systems. We are going to obtain continuity of our discontinuous feedback outside a set of singularity which will be proved small on account of semiconcavity.

THEOREM 2.9. *If the control affine system* (2.6) *is GAC, then there exists a subset $\mathcal{D} \subset \mathbb{R}^n$ which verifies the following properties:*

(i) *The set $\mathcal{D}$ is an open dense set.*
(ii) *The complement S of $\mathcal{D}$ has Hausdorff dimension no greater than $n-1$.*
(iii) *There exists a feedback $u : \mathbb{R}^n \to U$ which is continuous on $\mathcal{D}$ for which the closed-loop system* (2.6) *is globally asymptotically stable in the sense of Carathéodory; in particular, the Euler trajectories are solutions in the sense of Carathéodory.*

REMARK 2.10. *As in the paper of Artstein* [5, Theorem 5.2]*, if the system verifies the small-control property, then the feedback can be chosen to be continuous at the origin. More precisely, if we assume that there exists a semiconcave control-Lyapunov function such that for all $\epsilon > 0$, there exists $\delta > 0$ such that $\|x\| \leq \delta$ implies the existence of $u \in U$ with $\|u\| < \epsilon$ and satisfying* (2.4)*, then the feedback given by the previous theorem can be taken to be continuous at the origin.*

We recognize this time in Theorem 2.9 the result of Ancona and Bressan [4] in the case of control affine systems. As in their case, it turns out from the proof that the function $t \mapsto f(x(t), u(x(t))$ is left-continuous. Let us also remark that we get from the upper bound on the Hausdorff dimension of S that this set has Lebesgue measure zero. We refer to Morgan [19] for a survey of the notions of Hausdorff measure and Hausdorff dimension.

Actually, if we consider a control-Lyapunov function given by Theorem 2.4 we will see in the proof of Theorem 2.9 that the construction of the set $\mathcal{D}$ is based on the function

$$\Psi_V(x) := \min_{u \in U} \max_{\zeta \in \partial V(x)} \langle \zeta, f(x,u) \rangle, \tag{2.7}$$

where ∂V denotes the Clarke's generalized gradient of V; see section 3.1 for the definition. This function is upper semicontinuous on $\mathbb{R}^n \setminus \{0\}$; hence if we consider a continuous function $\delta : \mathbb{R}^n \to \mathbb{R}$, the set

$$\mathcal{D}_V^\delta := \{x \in \mathbb{R}^n \setminus \{0\} : \Psi_V(x) < -\delta(x)\}$$

is open. In particular, if the control-Lyapunov function V satisfies an additional assumption concerning the set $\mathcal{D}_V^\delta$ and the function Ψ_V, then we can provide a stabilizing feedback which is invariant with respect to $\mathcal{D}_V^\delta$. Let us state the result.

THEOREM 2.11. *Let there be given a GAC control system and a control-Lyapunov function V as in Theorem* 2.4. *If there exists a continuous and positive definite function*

$$\delta : \mathbb{R}^n \to \mathbb{R}$$

such that

$$\forall x \in \mathbb{R}^n \setminus \{0\}, \quad \delta(x) < V(x), \tag{2.8}$$

and

$$\forall x \in \mathbb{R}^n \setminus \{0\}, \quad \Psi_V(x) \leq -\delta(x) \Longrightarrow x \in \mathcal{D}_V^\delta, \tag{2.9}$$

then we have the following:

(i) *The set $\mathcal{D}_V^\delta$ is an open dense set and $\mathcal{D}_V^\delta \cup \{0\}$ is path-connected.*

(ii) *The complement S_V of $\mathcal{D}_V^\delta$ has Hausdorff dimension no greater than $n-1$.*

(iii) *There exists a feedback $u : \mathbb{R}^n \to U$ which is smooth on $\mathcal{D}_V^\delta$ for which the closed-loop system* (2.6) *is globally asymptotically stable in the sense of Carathéodory. Moreover, for any Carathéodory solution $x(\cdot)$ of this system, we have*

$$x(t) \in \mathcal{D}_V^\delta \quad \forall t > 0. \tag{2.10}$$

In particular, the Euler trajectories are solutions in the sense of Carathéodory.

REMARK 2.12. *Let us note that if there exists a positive definite and continuous function δ such that*

$$\forall x \in \mathbb{R}^n, \quad \Psi_V(x) < 0 \Longrightarrow \Psi_V(x) \leq -\delta(x),$$

then the function $\frac{\delta}{2}$ satisfies (2.8) *and* (2.9).

We stress that the property (2.10) implies the following facts: For each state $x_0 \in \mathbb{R}^n$, for any Carathéodory solution of the closed-loop system starting at x_0, the following hold:

- If $x_0 = 0$, then $x(t) = 0$ for all $t \geq 0$.
- If $x_0 \in \mathcal{D}_V^\delta$, then $x(t) \in \mathcal{D}_V^\delta \cup \{0\}$ for all $t \geq 0$; and consequently
$$\dot{x}(t) = f(x(t), u(x(t))) \quad \forall t \geq 0 \quad \text{such that (s.t.) } x(t) \neq 0.$$
- If $x_0 \notin \mathcal{D}_V^\delta \cup \{0\}$, then $x(t) \in \mathcal{D}_V^\delta \cup \{0\}$ for all $t > 0$; and consequently
$$\dot{x}(t) = f(x(t), u(x(t))) \quad \forall t > 0 \quad \text{s.t. } x(t) \neq 0.$$

To summarize our results, we have shown that under the additional assumptions (2.8) and (2.9), there exists a feedback which stabilizes our closed-loop system in the sense of Carathéodory, and moreover its Carathéodory trajectories are solutions in the classical sense for positive times whenever $x(t) \neq 0$. Let's also emphasize that the conclusions of Theorem 2.11 imply some topological properties for the set $\mathcal{D}_V^\delta$. As a matter of fact, the set $\mathcal{D}_V^\delta \cup \{0\}$ is invariant with respect to a locally Lipschitz vector field (since the functions $f_0, f_1, \ldots, f_m$ are locally Lipschitz), which is asymptotically stabilizing to the origin; therefore it is contractible.

We present in the two following sections two simple examples where the hypotheses (2.8) and (2.9) of Theorem 2.11 are fulfilled.

2.3. One-dimensional systems. Let us assume that the control system is of the form

$$\dot{x} = ug(x), \tag{2.11}$$

where the control u belongs to the interval $[a, b]$ and g is a locally Lipschitz vector field on $\mathbb{R}^n$. In this case, the condition (2.9) is always ensured. Let us consider a semiconcave control-Lyapunov V for the system (2.11) and set

$$\mathcal{D}_V := \{x \in \mathbb{R}^n \setminus \{0\} : \Psi_V(x) < 0\},$$

where

$$\forall x \in R^n, \quad \Psi_V(x) := \min_{u \in U} \max_{\zeta \in \partial V(x)} \langle \zeta, f(x, u) \rangle.$$

We have the following.

LEMMA 2.13. *For any $x \in \mathcal{D}_V$,*

$$\Psi_V(x) \leq -V(x).$$

Proof. Let $x \in \mathcal{D}_V$. Thus $\Psi_V(x) < 0$, and there exists $u \in [a, b]$ such that

$$\forall \zeta \in \partial V(x), \quad u\langle \zeta, g(x) \rangle < 0. \tag{2.12}$$

Without loss of generality we treat the case where $u > 0$.

We know by assumption on V that for each $\zeta \in \partial_L V(x)$, there exists $u(\zeta) \in [a, b]$ such that

$$u(\zeta)\langle \zeta, g(x) \rangle \leq -V(x).$$

Since $\partial_L V(x) \subset \partial V(x)$, we deduce immediately from (2.12) that $u(\zeta) > 0$ and that $\Psi(x) \leq -V(x)$. ∎

From this lemma and Remark 2.12 we deduce that the conclusions of Theorem 2.11 apply in the case of one-dimensional systems. We add that we will clearly define the shape of the set of singularities (i.e., S_V the complement of $\mathcal{D}_V$) in the forthcoming paper [22].

REMARK 2.14. *Actually, it is not difficult to see that the system* (2.11) *is GAC and locally stabilizable by a continuous feedback if and only if it is globally stabilizable by a smooth feedback.*

2.4. The nonholonomic integrator. The nonholonomic integrator control system

$$\begin{aligned} \dot{x}_1 &= u_1, \\ \dot{x}_2 &= u_2, \\ \dot{x}_3 &= x_1 u_2 - x_2 u_1 \end{aligned} \tag{2.13}$$

appeared in [7] as an example of a nonlinear control system which does not satisfy Brockett's condition and cannot be stabilized with continuous feedback. It was shown in [16] that the nonsmooth function

$$V = \max\left\{\sqrt{x_1^2 + x_2^2}, |x_3| - \sqrt{x_1^2 + x_2^2}\right\} \tag{2.14}$$

is a control-Lyapunov function for the nonholonomic integrator system (2.13). As before,

$$\Psi_V(x) := \min_{u \in U} \max_{\zeta \in \partial V(x)} \langle \zeta, (u_1, u_2, x_1 u_2 - x_2 u_1)^t \rangle,$$

and we remark that $\Psi_V(x) = 0$ on the set

$$S := \left\{ x \in \mathbb{R}^n : x_1^2 + x_2^2 = 0 \right\}.$$

In addition, the function V is differentiable outside S, and thus by the results given in [16] we get that for any $x \in \mathcal{D}_V$,

$$\Psi_V(x) \leq -\frac{V(x)}{\sqrt{4 + V(x)^2}} =: \delta(x).$$

Theorem 2.11 and Remark 2.12 now imply the existence of a stabilizing feedback satisfying the properties given in its statement.

2.5. A counterexample. We give in this section for every Euclidean space $\mathbb{R}^n$ with $n \geq 2$ an example of a control affine system which is GAC and which does not verify the conditions (2.8) and (2.9) for any control-Lyapunov function V and any continuous positive definite function δ, and for which the conclusions of Theorem 2.11 do not hold.

Let $n \geq 2$ and $x_0 \in R^n \setminus 3\overline{B}$ be fixed. There exists a locally Lipschitz vector field on $\mathbb{R}^n$ such that

$$f_0(x) = \begin{cases} -x & \text{if } x \in \overline{B}, \\ x - x_0 & \text{if } \frac{1}{4} \leq \|x - x_0\| \leq \frac{1}{2}. \end{cases}$$

Let us also define two auxiliary functions g_0 and g_1. We set for any $x \in \mathbb{R}^n$,

$$g_0(x) := \max\{0, 1 - d_{K_1}(x)\},$$

where $d_{K_1}(\cdot)$ denotes the distance function corresponding to the set

$$K_1 := \overline{B} \bigcup \left(x_0 + \frac{1}{2}\overline{B} \right) \setminus \left(x_0 + \frac{1}{4}\overline{B} \right).$$

Then we set for any $x \in \mathbb{R}^n, g_1(x) := d_{K_2}(x)$ with

$$K_2 := \frac{1}{2}\overline{B} \bigcup \mathcal{A} \subset K_1,$$

where $\mathcal{A}$ denotes the annulus $\mathcal{A} = \left(x_0 + \frac{7}{16}\overline{B}\right) \setminus \left(x_0 + \frac{5}{16}\overline{B}\right)$.

We now present the dynamics which will form our counterexample; we consider the following control system:

$$\dot{x} = f(x,u) := g_0(x)f_0(x) + g_1(x)u, \quad (x,u) \in \mathbb{R}^n \times \overline{B}. \tag{2.15}$$

Let us notice the following facts:

$$f(x,u) = f_0(x) \quad \text{if } (x,u) \in K_2 \times \overline{B}$$

and

$$f(x,u) = g_1(x)u \quad \text{if } (x,u) \in \left(\mathbb{R}^n \setminus (K_1 + \overline{B})\right) \times \overline{B}.$$

It is straightforward to show that this affine control system is GAC; let us notice that this is true since we are in a dimension greater than 2. Now, let us assume that there exists an open dense set $\mathcal{D}$ of $\mathbb{R}^n$ which is invariant with respect to some smooth (on $\mathcal{D}$) stabilizing feedback $k(\cdot)$ and such that its complement $S := (\mathbb{R}^n \setminus \{0\}) \setminus \mathcal{D}$ is repulsive (see (2.10)).

First, since our dynamics reduce to f_0 around the origin, we can assume that $0 \in \mathcal{D}$ and hence that $\mathcal{D}$ is contractible. On the other hand, the control system (2.15) coincides with the dynamical system $\dot{x} = f_0(x)$ on the interior of $\mathcal{A}$. Consequently, by repulsivity the set S cannot intersect $\mathrm{int}(\mathcal{A})$; hence we deduce that S meets the ball $x_0 + \frac{5}{16}\overline{B}$. (If the vector field $f(\cdot, k(\cdot))$ were continuous on this ball, it would have an equilibrium on it by Brouwer's theorem; as a matter of fact the ball would be invariant under the dynamic $\dot{x} = -f(x,k(x))$.) In other words, there exists a nonempty compact set K such that

$$K \subset x_0 + \frac{3}{8}\overline{B} \subset x_0 + \frac{7}{16}B$$

and

$$K \cap \mathcal{D} \neq \emptyset.$$

This means that we can write our set $\mathcal{D}$ as follows:

$$\mathcal{D} = \left(\mathcal{D} \bigcup x_0 + \frac{7}{16}B\right) \setminus K.$$

In fact, we can see $\mathcal{D}$ as an open set minus a compact subset of itself. Such a set can't be contractible (we refer to algebraic topology for the proof of this result).

In particular, this shows by Theorem 2.11 that the control system defined above does not possess a semiconcave (outside the origin) control-Lyapunov function V (with, for instance, $W = V$) and a continuous positive definite function verifying (2.8) and (2.9).

3. Complementary definitions.

3.1. Some facts in nonsmooth analysis. We recall briefly some notions of nonsmooth analysis which are essential for this article. We first define $\partial_P V(x)$ as the *proximal subdifferential* of V at x where the function $V : \mathbb{R}^n \to \mathbb{R}$ is assumed to be locally Lipschitz: ζ belongs to $\partial_P V(x)$ if and only if there exists σ and $\eta > 0$ such that

$$V(y) - V(x) + \sigma\|y - x\|^2 \geq \langle \zeta, y - x \rangle \quad \forall y \in x + \eta B. \tag{3.1}$$

We further state that this object can be empty at some points. Nevertheless it can be proved that the proximal subdifferential is nonempty on a dense set of $\mathbb{R}^n$. Such a property leads us to define the *limiting subdifferential* and the *generalized gradient* which will be nonempty at every point. For all x in $\mathbb{R}^n$, the limiting subdifferential of V at x is defined as follows:

$$\partial_L V(x) := \{\lim \zeta_k : x_k \to x, \zeta_k \in \partial_P V(x_k)\}. \tag{3.2}$$

REMARK 3.1. *Of course, by the construction of the limiting subdifferential and by continuity of $f(\cdot,\cdot)$, the property* (2.4) *given in Definition* 2.3 *is equivalent to the following one:*

$$\forall x \in \mathbb{R}^n \setminus \{0\},\ \forall \zeta \in \partial_L V(x), \quad \min_{u\in U}\langle \zeta, f(x,u)\rangle \le -W(x). \tag{3.3}$$

Finally, we derive the generalized gradient of Clarke as follows:

$$\partial V(x) := \text{co} \quad \partial_L V(x), \tag{3.4}$$

where co$A$ denotes the convex hull of the set A.

It is important to note that in our case of a locally Lipschitz function, the definition of the generalized gradient coincides with the following one based on Rademacher's theorem:

$$\partial V(x) := \text{co}\{\lim \nabla V(x_k) : x_k \to x, x_k \in D_f \setminus N\}, \tag{3.5}$$

where D_f denotes the set of differentiability of f and N is any set of Lebesgue measure zero in $\mathbb{R}^n$.

Moreover, we stress that there exist complete calculi of proximal subdifferentials and generalized gradients, ones that extend all theorems of the usual smooth calculus; our principal references for this theory are the books of Clarke [8] and Clarke et al. [12].

3.2. Results on semiconcave functions. We recall in this subsection some basic properties of the semiconcave functions. Let us first recall this definition; we assume in this section that Ω is a given open subset of $\mathbb{R}^n$.

DEFINITION 3.2. *Let $g : \Omega \longrightarrow \mathbb{R}$ be a continuous function on Ω; it is said to be semiconcave on Ω if for any point $x_0 \in \Omega$ there exist $\rho, C > 0$ such that*

$$g(x) + g(y) - 2g\left(\frac{x+y}{2}\right) \le C\|x-y\|^2 \tag{3.6}$$

for all $x, y \in x_0 + \rho B$.

The property (3.6) amounts to the concavity of $x \mapsto g(x) - 2C\|x\|^2$, as is easily checked. Hence, a semiconcave function g can be seen locally as the sum of a concave function and a smooth function. In particular, this implies that the semiconcave functions are locally Lipschitz. We know different examples of semiconcave functions. Concave functions are of course semiconcave. Another class of semiconcave functions is that of C^1 functions with locally Lipschitz gradient. We can in fact give a characterization of the semiconcavity property of a function g by using the proximal superdifferentials defined as follows:

$$\partial^P g(x) := -\partial_P g(x), \tag{3.7}$$

or, equivalently, $\zeta \in \partial^P g(x)$ if and only if there exists σ and $\eta > 0$ such that

$$g(y) - g(x) - \sigma\|y - x\|^2 \leq \langle \zeta, y - x\rangle \quad \forall y \in x + \eta B. \tag{3.8}$$

This analytic definition enables us to give a characterization of semiconcavity; we refer to [24] for the proof.

PROPOSITION 3.3. *A function $g : \Omega \to \mathbb{R}$ is semiconcave if and only if σ and η of* (3.8) *can be chosen uniform on the compact sets of Ω. Moreover, the superdifferential and the generalized gradients coincide on Ω.*

REMARK 3.4. *We can in fact relate the semiconcavity property of a given function to some geometric properties of its epigraph; we refer to* [24] *for such results. Furthermore we can define the semiconcavity property in a more general setting such that semiconcave functions keep the same behavior of concave functions; hence we can relate their differentiability properties to the ones of concave functions (see, for instance,* [32]*).*

Since the semiconcave functions are locally Lipschitz, we get by Rademacher's theorem that they are differentiable almost everywhere. Actually, since they are locally the sum of a semiconcave function and a smooth function, we can state positively by Alexandroff's theorem (see [2, 14]) that the semiconcave functions are twice differentiable almost everywhere.

A study has been devoted to the set of nondifferentiability of such functions. Alberti, Ambrosio, and Cannarsa [1] were able to provide some upper bounds on the dimension of singular sets of semiconcave functions.

Let $g : \Omega \to \mathbb{R}$ be a semiconcave function. Define

$$\Sigma^k(g) := \{x \in \Omega : \dim(\partial g(x)) = k\},$$

where $k \in [0, n]$ is an integer. Clearly, $\Sigma^0(g)$ represents the set of differentiability of u, and moreover

$$\Omega = \bigcup_{k=0}^{n} \Sigma^k(g). \tag{3.9}$$

We can evaluate the size of these sets.

PROPOSITION 3.5. *For any integer $k \in [0, n]$, the set $\Sigma^k(g)$ has Hausdorff dimension $\leq n - k$.*

We refer to [1] (see also [3]) for the proof and again to the book of Morgan [19] for a serious survey of the Hausdorff dimension.

Finally, Alberti, Ambrosio, and Cannarsa made some useful links between the Bouligand tangent cones of some subsets of the $\Sigma^k(g)$'s and the generalized gradients of g. Let us define for any $\alpha > 0$

$$\Sigma^k_\alpha(f) := \big\{x \in \Omega : \exists B^k_\alpha \subset \partial f(x) \text{ with } \operatorname{diam}(B^k_\alpha) = 2\alpha\big\},$$

where B^k_α denotes a ball of dimension k with diameter α. For any set $S \subset \mathbb{R}^n$, we shall denote by $S^\perp$ the set defined as follows:

$$S^\perp := \{p \in \mathbb{R}^n : q \mapsto \langle q, p\rangle \text{ is constant on } S\}.$$

We have the following result.

PROPOSITION 3.6. *The sets $\Sigma^k_\alpha(g)$ are closed sets, and*

$$T^B_{\Sigma^k_\alpha(g)}(x) \subset [\partial f(x)]^\perp \quad \forall x \in \Sigma^k_\alpha(f) \setminus \Sigma^{k+1}_\alpha(f).$$

We again refer to the paper of Alberti, Ambrosio, and Cannarsa for the proof.

3.3. Discontinuous stabilizing feedbacks. As it has been explained before, there do not exist robust stabilizing feedbacks in general. To overcome this difficulty, we describe a concept of solution of the general Cauchy problem

$$\dot{x} = f(x, u(x)), \quad x(0) = x_0, \tag{3.10}$$

where the feedback $u : \mathbb{R}^n \to U$ is not assumed to be continuous. This concept of solutions for differential equations with discontinuous right-hand side, inspired by the theory of differential games, has been used in the fundamental article of Clarke et al. [10] (see also [9]) to produce discontinuous stabilizing feedbacks; it provides an alternative approach to those developed by Sussmann [31] and Coron [13] (see also Pomet [20]).

Let $\pi = \{t_i\}_{i\geq 0}$ be a partition of $[0, \infty)$, by which we mean a countable, strictly increasing sequence t_i with $t_0 = 0$ such that $t_i \to \infty$ as $i \to \infty$. The *diameter* of π, denoted $\mathrm{diam}(\pi)$, is defined as $\sup_{i\geq 0}(t_{i+1} - t_i)$. Given an initial condition x_0, the π-trajectory $x(\cdot)$ corresponding to π is defined in a step-by-step fashion as follows. Between t_0 and t_1, $x(\cdot)$ is a classical solution of the differential equation

$$\dot{x}(t) = f(x(t), u(x_0)), \quad x(0) = x_0, \quad t_0 \leq t \leq t_1.$$

(Of course in general we do not have uniqueness of the solution, nor is there necessarily even one solution.) We then set $x_1 := x(t_1)$ and restart the system with control value $u(x_1)$:

$$\dot{x}(t) = f(x(t), u(x_1)), \quad x(t_1) = x_1, \quad t_1 \leq t \leq t_2,$$

and so on in this fashion. This resulting trajectory x is a physically meaningful one that corresponds to a natural sampling procedure and piecewise constant controls; this kind of solution, called a system sampling solution, is due to Krasovskii and Subbotin (see [15]). We proceed now to give the definition of the global asymptotic stabilization associated to this concept.

DEFINITION 3.7. *The system* (2.1) *is globally asymptotically stable in the sense of π-trajectories if there exist a function $M : \mathbb{R}_{>0} \longrightarrow \mathbb{R}_{>0}$ such that $\lim_{R\downarrow 0} M(R) = 0$ and two functions $T, \delta : \mathbb{R}_{>0} \times \mathbb{R}_{>0} \longrightarrow \mathbb{R}_{\geq 0}$ with the following property:*

For any $0 < r < R$, for any partition π with $\mathrm{diam}(\pi) \leq \delta(r, R)$, and for each initial state x_0 such that $\|x_0\| \leq R$, the corresponding π-trajectory $x(\cdot)$ is well-defined and satisfies the following:

(1) *for all $t \geq 0, \|x(t)\| \leq M(R)$;*

(2) *for all $t \geq T(r, R), \|x(t)\| \leq r$.*

REMARK 3.8. *This definition is equivalent to another one given by Sontag in* [27]. *In that paper, it was required that there exist a function $\beta \in \mathcal{KL}$ so that the following property held: For each $0 < \epsilon < K$, there exists a $\delta = \delta(\epsilon, K) > 0$ such that, for every sampling schedule π with $\mathrm{diam}(\pi) < \delta$, and for each initial state x_0 with $\|x_0\| \leq K$, the corresponding π-trajectory $x(\cdot)$ of* (2.1) *is well-defined and satisfies*

$$\|x_\pi(t)\| \leq \max\{\beta(K, t), \epsilon\} \quad \forall t \geq 0.$$

We can define from the concept of π-trajectories the notion of Euler trajectories. As presented in [25], we call an Euler solution of (2.2) any uniform limit of π-trajectories of this system with $\mathrm{diam}(\pi) \to 0$. Moreover, we will say that the closed-loop system (2.2) is globally asymptotically stable in the Euler sense (or that

the feedback u stabilizes in the Euler sense) if the two properties given in Definition 2.1 are satisfied for any Euler solutions.

We also recall briefly for the convenience of the reader that a function $x(\cdot) : [0,\infty) \to \mathbb{R}^n$ is called a Carathéodory solution (or trajectory) of our closed-loop system if it satisfies

$$\dot{x}(t) = f(x(t), u(x(t))) \quad \text{a.e. } \forall t \geq 0.$$

We will say in that case that the closed-loop system is stabilizing in the sense of Carathéodory.

4. Proof of Theorem 2.4. We can invoke the main result of [23] to get a control-Lyapunov function V_0 which is semiconcave on $\mathbb{R}^n \setminus \{0\}$. We begin by showing that there exists a smooth function $\gamma : (0,\infty) \to (0,\infty)$ which satisfies

$$\min_{u\in U} \langle \zeta, f(x,u)\rangle \leq -\gamma(V_0(x)) \quad \forall x \in \mathbb{R}^n \setminus \{0\},\ \forall \zeta \in \partial_P V_0(x). \tag{4.1}$$

We use the method given by Clarke, Ledyaev, and Stern in [11]. We set for all $v > 0$,

$$\gamma(v) := \min\{W(x); x \in \Gamma(v)\},$$

where

$$\Gamma(v) := \{x \in \mathbb{R}^n; V_0(x) = v\}.$$

It is not difficult to show that the multifunction Γ is locally Lipschitz, which implies that the function γ is locally Lipschitz on $(0,\infty)$ and verifies (4.1). Moreover, we can approximate γ by a smooth function $\tilde{\gamma}$ such that

$$0 < \tilde{\gamma} \leq \gamma.$$

Finally, without loss of generality we can suppose that γ is smooth and verifies (4.1). Now, we set

$$\Psi(t) := \int_1^t \frac{1}{\gamma(s)} ds. \tag{4.2}$$

This new function from $(0,\infty)$ into $\mathbb{R}$ is increasing, smooth, and verifies the three following properties:

$$\Psi'(t) = \frac{1}{\gamma(t)} \quad \forall t > 0, \tag{4.3}$$

$$\limsup_{t\downarrow 0} \Psi(t) \leq 0, \tag{4.4}$$

and

$$\liminf_{t\to\infty} \Psi(t) \geq 0. \tag{4.5}$$

We are now able to define a new control-Lyapunov function V_1. We set

$$V_1(x) := \begin{cases} V_0(x) e^{c\Psi(V_0(x))} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases} \tag{4.6}$$

By (4.4) and (4.5) and the properties of V_0, this new function is obviously proper, continuous at the origin, and locally Lipschitz on $\mathbb{R}^n \setminus \{0\}$. We now want to make the link between the proximal subdifferentials of V_1 and the proximal subdifferentials of V_0; for that, we give the following lemma.

LEMMA 4.1. *Let there be given two functions $f : \Omega \to \mathbb{R}$ and $F : \mathbb{R} \to \mathbb{R}$. If we assume that f is positive and locally Lipschitz on the open set Ω and that F is a C^2, positive, and increasing ($F' > 0$) function, then for all $x \in \Omega$,*

$$\partial_P[fF(f)](x) = [F(f(x)) + f(x)F'(f(x))]\partial_P f(x).$$

The same formula holds for the proximal superdifferential. Moreover, if the function f is taken to be semiconcave, then the new function $fF(f)$ is semiconcave as well.

Proof. Let us consider $x \in \mathbb{R}^n$ and $\zeta \in \partial_P[fF(f)](x)$; then by (3.1), there exists $\sigma \geq 0$ such that

$$f(y)F(f(y)) - f(x)F(f(x)) + \sigma\|y - x\|^2 \geq \langle \zeta, y - x\rangle \tag{4.7}$$

whenever y is in a neighborhood of x. The function $X \to XF(X)$ is C^2, so we have by Taylor's formula that there exists a constant C such that for all Y in a neighborhood of X,

$$YF(Y) - XF(X) = F(X) + XF'(X)(Y - X) + \frac{C}{2}\|Y - X\|^2 + o(\|Y - X\|^2).$$

We get for $Y = f(y)$ and $X = f(x)$ that $f(y)F(f(y)) - f(x)F(f(x))$ is equal to

$$[F(f(x)) + f(x)F'(f(x))][f(y) - f(x)] + \frac{C}{2}\|f(y) - f(x)\|^2 + h,$$

where $h = o(\|f(y) - f(x)\|^2)$.

We set $D := F(f(x)) + f(x)F'(f(x))$; by the assumptions on f and F, $D > 0$, and so we can divide by D. On the other hand, f being locally Lipschitz, we deduce that there exists a constant $\overline{\sigma} \geq 0$ such that

$$\frac{f(y)}{D} - \frac{f(x)}{D} + \overline{\sigma}\|y - x\|^2 \leq \left\langle \frac{\zeta}{D}, y - x \right\rangle \tag{4.8}$$

whenever y is in a neighborhood of x; and then by the characterization (3.1) we get

$$\partial_P f[F(f)](x) \subset [F(f(x)) + f(x)F'(f(x))]\partial_P f(x).$$

This proves one inclusion; the other is left to the reader. Of course, for the case of the proximal superdifferential, a similar proof is valid.

It remains to show the conservation of semiconcavity. If we assume that f is semiconcave, then by using Proposition 3.3 and following the same proof as above, we show that the different σ remain uniform on the compact sets of Ω. □

We now turn back to the proof of Theorem 2.4; the lemma implies immediately that for all $x \in \mathbb{R}^n \setminus \{0\}$ and all $\zeta \in \partial_P V_1(x) \subset \partial_L V_1(x)$,

$$\begin{aligned}
\min_{u \in U}\langle \zeta, f(x,u)\rangle &\leq -\gamma(V_0(x))[e^{\Psi(V_0(x))} + c\Psi'(V_0(x))V_0(x)e^{\Psi(V_0(x))}] \\
&\leq -V_1(x)\left[\frac{\gamma(V_0(x))}{V_0(x)} + c\gamma(V_0(x))\psi'(V_0(x))\right] \\
&\leq -V_1(x)\left[\frac{\gamma(V_0(x))}{V_0(x)} + c\right] \quad \text{by (4.3)} \\
&\leq -cV_1(x).
\end{aligned}$$

On the other hand, as the initial function V_0 was semiconcave on $\mathbb{R}^n \setminus \{0\}$, we have by Lemma 4.1 that the new function V_1 is semiconcave; the proof of Theorem 2.4 is complete.

5. Proof of Theorems 2.5 and 2.7. We will treat only the case where the control-Lyapunov function is that given by Theorem 2.4. The general case of a control-Lyapunov function related to a function W is left to the reader.

Let V be the semiconcave Lyapunov function given by Theorem 2.4 and two positive constants $r < R$. We set

$$M_R := \max\{V(x) : \|x\| \leq R\} \quad \text{and} \quad M(R) := \max\{\|y\| : V(y) \leq M_R\}.$$

Obviously, the function $M(\cdot)$ is nondecreasing and verifies

$$\lim_{R \downarrow 0} M(R) = 0.$$

We also set two constants depending on r:

$$m_r := \min\{V(x) : \|x\| \geq r\} \quad \text{and} \quad m_{\frac{r}{2}} := \min\left\{V(x) : \|x\| \geq \frac{r}{2}\right\};$$

we can say by definition that if $V(x) \leq \frac{m_{\frac{r}{2}}}{2}$, then $x \in \frac{r}{2}\overline{B}$. On the other hand, the Proposition 3.3 allows us to consider $\sigma := \sigma(\frac{r}{2}, R)$ and δ uniform on the set $\mathcal{A} := \{x : \frac{m_{\frac{r}{2}}}{2} \leq V(x) \leq M_R\} \subset \mathbb{R}^n \setminus \{0\}$.

We get that for all $x \in \mathcal{A}$, all $y \in \mathcal{A}$, and all $\zeta \in \partial_L V(x)$,

$$-V(y) + V(x) + \sigma\|y - x\|^2 \geq \langle -\zeta, y - x\rangle. \tag{5.1}$$

From now on, we denote by M_f the upper bound of f on $R\overline{B} \times U$, by L_f the Lipschitz constant of f on the same set, by m_V the minimum of V, and by L_V the Lipschitz constant of V on $\mathcal{A}$. Let us consider a π-trajectory $x(\cdot)$ associated to a partition $\pi = \{0 = t_0 < t_1 < \cdots\}$ and to nodes $x_i := x(t_i)$ with $x_0 \in \mathcal{A}$. We pick ζ_0 belonging to $\partial_L V(x_0)$. For any $t \in [t_0, t_1]$, we can compute by (5.1)

$$\begin{aligned}
V(x(t)) - V(x_0) &\leq \langle \zeta_0, x(t) - x_0\rangle + \sigma\|x(t) - x_0\|^2 \\
&\leq \left\langle \zeta_0, \int_{t_0}^{t} f(x(s), u(x_0))ds \right\rangle + \sigma\|x(t) - x_0\|^2 \\
&\leq \langle \zeta_0, (t - t_0)f(x_0, u(x_0))\rangle \cdots \\
&\quad + \left\langle \zeta_0, \int_{t_0}^{t} [f(x(s), u(x_0)) - f(x_0, u(x_0))]ds \right\rangle + \sigma\|x(t) - x_0\|^2 \\
&\leq -(t - t_0)V(x_0) + \|\zeta_0\|L_f \max_{s \in [t_0, t_1]} \|x(s) - x_0\|(t - t_0) \cdots \\
&\quad + \sigma\|x(t) - x_0\|^2 \\
&\leq -(t - t_0)V(x_0) + L_V L_f M_f (t - t_0)^2 + \sigma M_f^2 (t - t_0)^2 \\
&\leq (t - t_0)\left[-m_V + (L_f L_V M_f + \sigma M_f^2)(t - t_0)\right].
\end{aligned}$$

More generally, we have for all $t \in [t_i, t_{i+1}]$,

$$V(x(t)) - V(x_i) \leq -(t - t_i)V(x_i) + [L_V L_f M_f + \sigma M_f^2](t - t_i)^2. \tag{5.2}$$

We get that for any n and for all $t \in [t_{n-1}, t_n]$,

$$\begin{aligned} V(x(t)) - V(x_0) &\leq \sum_{i=0}^{n-2} [-(t_{i+1} - t_i)V(x_i) + (L_V L_f M_f + \sigma M_f^2)(t_{i+1} - t_i)^2] \cdots \\ &\quad - (t - t_{n-1})V(t_{n-1}) + (L_f L_V M_f + \sigma M_f^2)(t - t_{n-1})^2 \\ &\leq (t - t_0)[-m_V + (L_f L_V M_f + \sigma M_f^2)\mathrm{diam}(\pi)]. \end{aligned} \tag{5.3}$$

We deduce that if we set

$$\delta(r, R) := \min \left\{ \frac{m_V}{2(L_f L_V M_f + \sigma M_f^2)}, \frac{m_r - m_{\frac{r}{2}}}{2L_V M_f} \right\},$$

we obtain from (5.3) that for every π-trajectory $x(\cdot)$ starting at x_0 and such that $\mathrm{diam}(\pi) \leq \delta(r, R)$, we have

$$\forall t \geq 0, \quad V(x(t)) - V(x_0) \leq -\frac{m_V}{2}(t - t_0). \tag{5.4}$$

That means that the π-trajectory remains in $\{x : V(x) \leq V(x_0)\}$, which is included in $M(R)\overline{B}$, and that for $t \geq T(r, R) := \frac{2M_R - m_{\frac{r}{2}}}{m_V}$,

$$V(x(t)) \leq V(x_0) - \frac{m_V}{2}t \leq \frac{m_{\frac{r}{2}}}{2};$$

that is, $x(t) \in \frac{r}{2}\overline{B}$. There is a possible danger! The work done above is valid only when we stay in the set $\mathcal{A}$. But as $\delta(r, R) \leq \frac{m_r - m_{\frac{r}{2}}}{2L_V M_f}$, there exists a first step i_0 for which $\frac{m_{\frac{r}{2}}}{2} \leq V(x_{i_0}) \leq \frac{m_r}{2}$, and by the same computation as above, the set $\{x : V(x) \leq V(x_{i_0})\}$ is invariant, that is,

$$\forall t \geq t_{i_0}, \quad x(t) \in \{x : V(x) \leq V(x_{i_0})\} \subset r\overline{B}.$$

This completes the proof for the case of π-trajectories.

We get from this proof (more especially from (5.2) and a convergence result of a Riemann's sums) that for any Euler trajectory of (2.2), we have that

$$V(x(t)) - V(x(s)) \leq -\int_s^t V(x(y))dy \quad \forall 0 \leq s \leq t. \tag{5.5}$$

Gronwall's lemma now brings a proof of the property (2.4).

We now make the proof of Theorem 2.7.

Proof. As in the statement of Theorem 2.5, the formula given above considers for all x a limiting subgradient $\zeta_V(x)$ and a function $u(\cdot)$ satisfying

$$\langle \zeta_V(x), f(x, u(x)) \rangle \leq -V(x) \quad \forall x \in \mathbb{R}^n \setminus \{0\}.$$

This construction agrees with the one given in the statement of Theorem 2.5; the result follows. ∎

6. Proof of Theorem 2.9. Theorem 2.9 requires a more subtle proof; we will need the following lemma and refer to the book of Clarke et al. [12] or to [24] for the proof.

LEMMA 6.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz function and $x \in \mathbb{R}^n$; if $\partial_P f(x)$ and $\partial^P f(x)$ are nonempty, then*

$$\partial_P f(x) = \partial^P f(x) = \partial f(x) = \{\nabla f(x)\}.$$

We know by Theorem 2.4 and Remark 3.1 that

$$\forall x \in \mathbb{R}^n \setminus \{0\}, \quad \max_{\zeta \in \partial_L V(x)} \min_{u \in U} \langle \zeta, f(x,u) \rangle \leq -V(x) < 0. \tag{6.1}$$

We set the function $\Psi : \mathbb{R}^n \longrightarrow \mathbb{R}$ as follows:

$$\forall x \in \mathbb{R}^n, \quad \Psi(x) := \min_{u \in U} \max_{\zeta \in \partial V(x)} \langle \zeta, f(x,u) \rangle.$$

LEMMA 6.2. *The function Ψ is upper semicontinuous.*

Proof. Since the function $\zeta \mapsto \langle \zeta, f(x,u) \rangle$ is upper continuous and the function f is continuous, we deduce that the function

$$x \mapsto \max_{\zeta \in \partial V(x)} \langle \zeta, f(x,u) \rangle$$

is upper semicontinuous. To conclude, we know that a minimum of upper semicontinuous functions is upper semicontinuous. □

We define now the following sets:

$$\mathcal{D} := \left\{ x \in \mathbb{R}^n \setminus \{0\} \text{ s.t. } \Psi(x) < -\frac{V(x)}{2} \right\} \quad \text{and} \quad S = \mathbb{R}^n \setminus \mathcal{D}.$$

LEMMA 6.3. *The set $\mathcal{D}$ is an open dense set of $\mathbb{R}^n$ and*

$$\mathcal{H} - \dim S \leq n - 1.$$

Proof. Since the multivalued mapping $x \mapsto \partial V(x)$ has a closed graph, the set $\mathcal{D}$ is obviously open. On the other hand, by the density theorem [12, Theorem 3.1], the proximal subdifferential $\partial_P V(x)$ is nonempty on a dense set. Consequently, by the semiconcavity of V, both proximal sub- and superdifferentials are nonempty on this set; it implies that (by Lemma 6.1) $\partial_P V(x) = \partial_L V(x) = \{\nabla V(x)\}$ on a dense subset of $\mathbb{R}^n$. So, we conclude that the min-max and the max-min of (6.1) and of the definition of the set $\mathcal{D}$ coincide on an open dense set of $\mathbb{R}^n$; that means that $\mathcal{D}$ contains this set. Consequently $\mathcal{D}$ is an open dense set of $\mathbb{R}^n$. Moreover, the complement S of $\mathcal{D}$ is included in $\cup_{k=1,\ldots,n} S^k(V)$; therefore we get the upper bound on the Hausdorff dimension of S by Proposition 3.5 given in section 3.2. □

We define the following multifunction on $\mathcal{D}$:

$$\forall x \in \mathcal{D}, \quad G(x) := \left\{ u \in U : \forall \zeta \in \partial V(x), \langle \zeta, f(x,u) \rangle \leq -\frac{V(x)}{2} \right\}.$$

LEMMA 6.4. *The multifunction G has nonempty closed convex values and is lower semicontinuous on $\mathcal{D}$.*

Proof. Since the system is affine in the control, the multifunction G has nonempty closed convex values. We show now that G is lower semicontinuous on $\mathcal{D}$ (we refer to

[6] for a task about semicontinuity of multivalued functions). We then have to prove that for any sequence $(x_n)_n$ of points in $\mathcal{D}$ converging to $x \in \mathcal{D}$, and for any $z \in G(x)$, there exists a sequence $(z_n)_n$ of points in $G(x_n)$ with limit z.

Let $(x_n)_n$ be a sequence in $\mathcal{D}$ converging to $x \in \mathcal{D}$, and let $y = f(x, u_0)$ in $G(x)$. We set, for all n,

$$z_n := f(x_n, u_0).$$

From now on, we denote by M_1 the Lipschitz constant of V, by M_2 the upper bound of $f(\cdot, u_0)$, and by M_3 the Lipschitz constant of $f(\cdot, u_0)$ in a neighborhood of x containing all the x_n (without loss of generality we can assume this condition), and on the other hand we denote by $\beta(A, B)$ the Hausdorff distance between the sets A and B (see [6]). Two cases appear.

1. $\max_{\zeta \in \partial V(x)} \langle \zeta, y \rangle < -\frac{V(x)}{2}$.
   We fix n and we choose $\zeta_n \in \partial V(x_n)$; so we have

$$\begin{aligned}
\langle \zeta_n, z_n \rangle &= \langle \zeta, z_n \rangle + \langle \zeta_n - \zeta, z_n \rangle \quad (\text{where } \zeta := \mathrm{proj}_{\partial V(x)}(\zeta_n)) \\
&\leq \langle \zeta, y \rangle + M_1 \|z_n - y\| + M_2 \beta(\partial V(x_n), \partial V(x)) \\
&< -\frac{V(x)}{2} + M_1 \|z_n - y\| + M_2 \beta(\partial V(x_n), \partial V(x)) \\
&< -\frac{V(x_n)}{2} + \frac{M_3}{2} \|x_n - x\| + M_3 \|x_n - x\| + M_2 \beta(\partial V(x_n), \partial V(x)).
\end{aligned}$$

   Hence, for n sufficiently high, $z_n \in G(x_n)$, and the sequence $(z_n)_n$ converges to $y = f(x, u_0)$ by continuity of f.
2. $\max_{\zeta \in \partial V(x)} \langle \zeta, y \rangle = -\frac{V(x)}{2}$.
   We know by assumption that since $x \in \mathcal{D}$ there exists $u_1 \in U$ such that

$$\forall \zeta \in \partial V(x), \quad \langle \zeta, f(x, u_1) \rangle < -\frac{V(x)}{2}.$$

   Consequently, we can express $y = f(x, u_0)$ as a limit of

$$y_p = t_p y + (1 - t_p) f(x, u_1)$$

   (when $t_p \uparrow 1$) such that $\max_{\zeta \in \partial V(x)} \langle \zeta, y_p \rangle < -\frac{V(x)}{2}$. On the other hand, by the first case each y_p is the limit of some sequence $(z_p^n)_n$; we conclude by a diagonal process.

We conclude that G is a lower semicontinuous multifunction on the set $\mathcal{D}$. ∎

Returning to the proof of our theorem, we can apply the well-known selection theorem of Michael [18, 6] to deduce the existence of a continuous selection u of G on $\mathcal{D}$.

Let us now set $\mathcal{E}$ as the set defined by

$$\mathcal{E} := \left\{ x \in \mathbb{R}^n \setminus \{0\} \text{ s.t. } \Psi(x) = -\frac{V(x)}{2} \right\}.$$

LEMMA 6.5. *For each x in $\mathcal{E}$, there exists a unique $f(x, u) \in U$ such that*

$$\max_{\zeta \in \partial_C V(x)} \langle \zeta, f(x, u) \rangle = -\frac{V(x)}{2}.$$

Proof. This is due to the assumption of strict convexity on the set of control U. If $f_0(x) \neq 0$ we leave it to the reader to prove that, modifying the dynamics if necessary, the lemma holds. □

We are now able to complete the construction of our feedback $u(\cdot)$. We set for each $x \in \mathcal{E}, u(x) := u$, where u is the u of Lemma 6.5. Moreover, for each $x \in S \setminus \mathcal{E} \setminus \{0\}$, we set $u(x) := u$, where u verifies

$$\exists \zeta \in \partial_L V(x), \quad \langle \zeta, f(x,u) \rangle \leq -V(x).$$

We have defined our feedback on all the space (of course, we set $u(0) = 0$). Let us now prove the rest of the theorem. Consider the closed-loop system

$$\dot{x} = f_0(x) + \sum_{i=1}^{m} u_i(x) f_i(x) \tag{6.2}$$

and show that it is globally asymptotically stable with respect to the Carathéodory solutions. Let us first show that the property (2.10) holds for Euler trajectories.

LEMMA 6.6. *For any Euler trajectory $x(\cdot)$ of (6.2), we have*

$$\Psi(x(t)) \leq -\frac{V(x(t))}{2} \quad \forall t > 0 \text{ s.t. } x(t) \neq 0. \tag{6.3}$$

Moreover

$$\dot{x}(t) = f(x(t), u(x(t))) \quad \text{a.e. } t > 0. \tag{6.4}$$

Proof. Let us consider $x_0 \in \mathbb{R}^n \setminus \{0\}$ and $x(\cdot)$ is an Euler trajectory of (6.2) with $x(0) = x_0$. Obviously, if $x_0 = 0$, then since $f(0, u(0)) = 0$ all the Euler solutions of (6.2) starting at $x_0 = 0$ will stay at the origin; then the property (6.4) holds. Let us now assume that $x_0 \neq 0$.

Let t_0 be fixed in $(0, \infty)$; there exists $\sigma > 0$ such that for any $\zeta \in \partial^P V(x(t_0))$, we have that

$$-V(y) + V(x(t_0)) + \sigma \|y - x(t_0)\|^2 \geq \langle -\zeta, y - x(t_0) \rangle$$

whenever y is in a neighborhood of $x(t_0)$. We deduce that for some $s < t_0$ and close to t_0, we have

$$V(x(t_0)) - V(x(s)) + \sigma \|x(s) - x(t_0)\|^2 \geq \langle \zeta, x(t_0) - x(s) \rangle.$$

This implies

$$\begin{aligned} \langle \zeta, x(t_0) - x(s) \rangle &\leq V(x(t_0)) - V(x(s)) + \sigma \|x(s) - x(t_0)\|^2 \\ &\leq -\int_s^{t_0} \frac{V(x(y))}{2} dy + \sigma \|x(s) - x(t_0)\|^2 \quad \text{by (5.5)}. \end{aligned} \tag{6.5}$$

Now, by convexity of $f(x(t), U)$ (since f is affine in the control) there exists a sequence $(s_n)_n$ and u_0 in U such that

$$\lim_{n \to \infty} \frac{x(t_0) - x(s_n)}{t_0 - s_n} = f(x(t_0), u_0). \tag{6.6}$$

Consequently, passing to the limit for the sequence $(s_n)_n$, we obtain

$$\langle \zeta, f(x(t_0), u_0)\rangle \leq -\frac{V(x(t_0))}{2}.$$

We can repeat this argument for all $\zeta \in \partial^P V(x(t_0))$, that is,

$$\forall \zeta \in \partial^P V(x(t_0)), \quad \langle \zeta, f(x(t_0), u_0)\rangle \leq -\frac{V(x(t_0))}{2}.$$

Since $\partial^P V(x(t_0) = \partial V(x(t_0))$, that means that

$$\Psi(x(t_0)) \leq -\frac{V(x(t_0))}{2}.$$

Hence, we deduce that for any $t > 0$, (6.3) is satisfied and

$$x(t) \in \mathcal{D} \cup \mathcal{E}.$$

Two cases appear. If $x(t) \in \mathcal{D}$, then by continuity of $u(\cdot)$ in a neighborhood of $x(t)$, we have $\dot{x}(t) = f(x(t)), u(x(t))$.

Otherwise, $x(t) \in \mathcal{E}$. In this case, Lemma 6.5 asserts that the set of limits of the form (6.6) is a singleton. Thus, we deduce that the function $x(\cdot)$ is left-derivable on $(0, \infty)$ with derivate $f(x(t), u(x(t)))$.

Now, since the trajectory $x(\cdot)$ is locally Lipschitz on $[0, \infty)$, Rademacher's theorem asserts that it is derivable almost everywhere. Then we conclude that this derivate coincides with $f(x(t), u(x(t)))$ almost everywhere; consequently, the Euler trajectories are solutions in the sense of Carathéodory. □

Consider now the case of solutions in the sense of Carathéodory. Let $x_0 \neq 0$ and let $x(\cdot)$ be a Carathéodory solution of (6.2) starting at x_0. Hence, we have a set N of measure zero on $[0, \infty)$ such that

$$\dot{x}(t) = f(x(t), u(x(t))) \quad \forall t \in [0, \infty) \setminus N. \tag{6.7}$$

We have by the mean value inequality (see [12, Exercise 2.7(d), p. 122]) that for any $0 \leq t < t'$, there exists $z_{t,t'} \in [x(t), x(t')]$ and $\zeta_{t,t'} \in \partial_L V(z_{t,t'})$ such that

$$V(x(t')) - V(x(t)) \leq \langle \zeta_{t,t'}, x(t') - x(t)\rangle. \tag{6.8}$$

Now by setting the function $\theta : \mathbb{R}_{\geq 0} \to \mathbb{R}, \theta(t) := V(x(t))$, it means that for any $t, t' \geq 0$

$$\theta(t') - \theta(t) \leq \langle \zeta_{t,t'}, x(t') - x(t)\rangle. \tag{6.9}$$

Since the function f is locally bounded, the function θ is locally Lipschitz and hence by Rademacher's theorem differentiable outside a set of measure zero N'. Therefore, for all $t \in [0, \infty) \setminus N \cup N'$, we obtain by passing to the limit in (6.9)

$$\theta'(t) \leq \langle \zeta, \dot{x}(t)\rangle = \langle \zeta, f(x(t), u(x(t)))\rangle,$$

where $\zeta \in \partial_L V(x(t))$.

LEMMA 6.7. *The Carathéodory trajectory $x(\cdot)$ does not belong to $S \setminus \mathcal{E}$ almost everywhere:*

$$x(t) \notin S \setminus \mathcal{E} \quad \textit{a.e. } t \geq 0.$$

The proof is based on the properties on the sets $S^k(V)$ given in section 3.2 and is postponed to the end of this section.

By Lemma 6.7 there exists a third set N^0 of measure zero such that

$$x(t) \in \mathcal{D} \quad \forall t \in [0,\infty) \setminus N^0.$$

Thus we get by construction of u that for all $t \in [0,\infty) \setminus N \cup N' \cup N^0$,

$$\theta'(t) \leq -\frac{V(x(t))}{2}.$$

Therefore, we deduce by the characterization given in section 3.1 that for any $t \geq 0$

$$\partial\theta(t) \subset \left(-\infty, -\frac{V(x(t))}{2}\right].$$

We deduce that the function $t \mapsto \theta(t) + \int_0^t \frac{V(x(s))}{2} ds$ is nonincreasing, and consequently

$$\forall 0 \leq s \leq t, \quad V(x(t)) - V(x(s)) \leq -\int_s^t \frac{V(x(y))}{2} dy. \tag{6.10}$$

Now if we fix $t_0 > 0$, and if we take $\zeta \in \partial^P V(x(t_0))$, we get by (6.5) and (6.10) that

$$\langle \zeta, x(t_0) - x(s) \rangle \leq -\int_s^{t_0} \frac{V(x(y))}{2} dy + \sigma \|x(s) - x(t_0)\|^2.$$

Then we deduce as in the case of Euler trajectory that for all $t > 0$,

$$x(t) \in \mathcal{D} \cup \mathcal{E}.$$

Now, Gronwall's lemma easily gives

$$\forall t > 0, \quad V(x(t)) \leq e^{-\frac{t}{2}} v(x_0)$$

for any Euler trajectory and any Carathéodory trajectory starting at x_0. We get that the closed-loop system (6.2) is globally asymptotically stable.

It remains to prove Lemma 6.7.

Proof. Assume that the conclusion is false. Then there would exist a subset H of $[0,\infty)$ of positive measure such that $x(\cdot)$ is differentiable in H and

$$x(t) \in S \setminus \mathcal{E} \quad \forall t \in H.$$

On the other hand by (3.9), we can write

$$\begin{aligned} S \setminus \mathcal{E} &= \bigcup_{k=1}^{n} \Sigma^k(V) \cap S \setminus \mathcal{E} \\ &= \bigcup_{k=1}^{n} \bigcup_{p \in \{1,2,\ldots\}} \Sigma^k_{\frac{1}{p}}(V) \cap S \setminus \mathcal{E}. \end{aligned}$$

Thus there exists a couple (k,p) for which

$$x(t) \in \Sigma^k_{\frac{1}{p}}(V) \cap S \setminus \mathcal{E} \subset \Sigma^k_{\frac{1}{p}}(V)$$

on a set of positive measure $H' \subset H$. This implies that there exists a $t_0 \in H'$ such that

$$\dot{x}(t_0) = f(x(t_0), u(x(t_0))) \in T^B_{\Sigma^k_{\frac{1}{p}}(V)}(x(t_0)).$$

Hence we deduce by Proposition 3.6 that

$$\forall \zeta \in \partial V(x), \quad \langle \zeta, f(x(t_0), u(x(t_0)))\rangle = -V(x(t_0)).$$

This last inequality implies that $x(t_0) \in \mathcal{D}$; we get a contradiction. □

7. Proof of Theorem 2.11. Let us recall that

$$\mathcal{D}^\delta_V := \{x \in \mathbb{R}^n \setminus \{0\} : \Psi_V(x) < -\delta(x)\}. \tag{7.1}$$

Since the function Ψ_V is upper semicontinuous, the set $\mathcal{D}^\delta_V$ is open, and since $\delta < V$ on $\mathbb{R}^n \setminus \{0\}$, by the same proof as before (see proof of Theorem 2.9) it is dense. Furthermore, by hypothesis (2.9), it is straightforward to show that there exists a continuous positive definite function $\epsilon : \mathbb{R}^n \to \mathbb{R}$ such that

$$\forall x \in \mathcal{D}^\delta_V, \quad \delta(x) + \epsilon(x) < V(x)$$

and

$$\forall x \in \mathcal{D}^\delta_V, \quad \Psi_V(x) \le -\delta(x) - \epsilon(x).$$

This implies, by the same proof as for Theorem 2.9 (replacing the term $\frac{V(x)}{2}$ by $\delta(x) + \epsilon(x)$), that $\mathcal{D}^\delta_V$ is open dense and that there exists a continuous function

$$u : \mathcal{D}^\delta_V \to U$$

such that for any $x \in \mathcal{D}^\delta_V$,

$$\forall \zeta \in \partial V(x), \quad \langle \zeta, f(x, u(x))\rangle \le -\delta(x) - \epsilon(x).$$

Now, we claim that there exists a function

$$\overline{u} : \mathcal{D}^\delta_V \to U,$$

which is smooth and such that

$$\forall x \in \mathcal{D}^\delta_V, \quad \|u(x) - \overline{u}(x)\| \le \frac{\epsilon(x)}{K_V(x)\sum_{i=1}^m M_i},$$

where $K_V(x)$ denotes the Lipschitz constant of the function V on the ball $\overline{B}(x, \frac{\|x\|}{2})$ and where the M_i's are the upper bounds of the functions f_i's on the ball $2\|x\|\overline{B}$. Such a function brings that for any $x \in \mathcal{D}^\delta_V$ and for any $\zeta \in \partial V(x)$,

$$\begin{aligned}\left\langle \zeta, f_0(x) + \sum_{i=1}^m \overline{u_i}(x) f_i(x) \right\rangle &\le -\delta(x) - \epsilon(x) + \|\zeta\|\|u(x) - \overline{u}(x)\| \sum_{i=1}^m \|f_i(x)\| \\ &\le -\delta(x) - \epsilon(x) + K_V(x) \sum_{i=1}^m M_i \|u(x) - \overline{u}(x)\| \\ &\le -\delta(x).\end{aligned}$$

Finally, considering a Carathéodory solution $x(\cdot)$ of

$$\dot{x} = f_0(x) + \sum_{i=1}^{m} \overline{u_i}(x) f_i(x),$$

starting at $x_0 \in \mathbb{R}^n$ we get that for any $t_0 > 0$ such that $x(t_0) \neq 0$,

$$\Psi_V(x(t_0)) \leq -\delta(x(t_0)).$$

This means by (2.9) that the trajectory $x(\cdot)$ stays in $\mathcal{D}_V^\delta$ for positive times. Theorem 2.11 is proved.

Acknowledgments. The author is grateful to Francis Clarke for comments and several corrections on a previous version of the paper. He is also indebted to Olivier Ley for his help in the writing of this article.

REFERENCES

[1] G. Alberti, L. Ambrosio, and P. Cannarsa, *On the singularities of convex functions*, Manuscripta Math., 76 (1992), pp. 421–435.
[2] A.D. Alexandroff, *Almost everywhere existence of the second differential of a convex function and some properties of convex surfaces connected with it*, Leningrad State Univ. Annals [Uchenye Zapiski] Math. Ser., 6 (1939), pp. 3–35.
[3] L. Ambrosio, P. Cannarsa, and H.M. Soner, *On the propagation of singularities of semi-convex functions*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 20 (1993), pp. 597–616.
[4] F. Ancona and A. Bressan, *Patchy vector fields and asymptotic stabilization*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 445–471.
[5] Z. Artstein, *Stabilization with relaxed controls*, Nonlinear Anal., 7 (1983), pp. 1163–1173.
[6] J.P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser Boston, Boston, 1990.
[7] R.W. Brockett, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, R.W. Brockett, R.S. Millman, and H.J. Sussmann, eds., Birkhäuser Boston, Boston, 1983, pp. 181–191.
[8] F.H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983; reprinted as Classics Appl. Math. 5, SIAM, Philadelphia, 1990.
[9] F.H. Clarke, Yu.S. Ledyaev, L. Rifford, and R.J. Stern, *Feedback stabilization and Lyapunov functions*, SIAM J. Control Optim., 39 (2000), pp. 25–48.
[10] F.H. Clarke, Yu.S. Ledyaev, E.D. Sontag, and A.I. Subbotin, *Asymptotic controllability implies feedback stabilization*, IEEE Trans. Automat. Control, 42 (1997), pp. 1394–1407.
[11] F.H. Clarke, Yu.S. Ledyaev, and R.J. Stern, *Asymptotic stability and smooth Lyapunov functions*, J. Differential Equations, 149 (1998), pp. 69–114.
[12] F.H. Clarke, Yu.S. Ledyaev, R.J. Stern, and P.R. Wolenski, *Nonsmooth Analysis and Control Theory*, Grad. Texts in Math. 178, Springer-Verlag, New York, 1998.
[13] J.-M. Coron, *Global asymptotic stabilization for controllable systems without drift*, Math. Control Signals Systems, 5 (1992), pp. 295–312.
[14] M.G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.
[15] N.N. Krasovskii and A.I. Subbotin, *Game-Theoretical Control Problems*, Springer-Verlag, New York, 1988.
[16] Yu.S. Ledyaev and L. Rifford, *Robust Stabilization of the Nonholonomic Integrator*, in preparation.
[17] Yu.S. Ledyaev and E.D. Sontag, *A Lyapunov characterization of robust stabilization*, Nonlinear Anal., 37 (1999), pp. 813–840.
[18] E. Michael, *Continuous selections.* I, Ann. of Math. (2), 63 (1956), pp. 361–382.
[19] F. Morgan, *Geometric Measure Theory. A Beginner's Guide*, Academic Press, Boston, 1988.
[20] J.-B. Pomet, *Explicit design of time-varying stabilizing control laws for a class of controllable systems without drift*, Systems Control Lett., 18 (1992), pp. 147–158.
[21] L. Rifford, *On the existence of nonsmooth control-Lyapunov functions in the sense of generalized gradients*, ESAIM Control Optim. Calc. Var., 6 (2001), pp. 593–61.
[22] L. Rifford, *Singularities of Some Viscosity Supersolutions and the Stabilization Problem in the Plane*, manuscript.

[23] L. Rifford, *Existence of Lipschitz and semiconcave control-Lyapunov functions*, SIAM J. Control Optim., 39 (2000), pp. 1043–1064.
[24] L. Rifford, *Problèmes de stabilisation en théorie du controle*, Ph.D. thesis, Université Claude Bernard Lyon I, Lyon, France, 2000.
[25] L. Rifford, *Stabilisation des systèmes globalement asymptotiquement commandables*, C. R. Acad. Sci. Paris Sér. I Math., 330 (2000), pp. 211–216.
[26] E.D. Sontag, *A "universal" construction of Artstein's theorem on nonlinear stabilization*, Systems Control Lett., 13 (1989), pp. 117–123.
[27] E.D. Sontag, *Stability and stabilization: Discontinuities and the effect of disturbances*, in Nonlinear Analysis, Differential Equations and Control (Montreal, QC, 1998), Kluwer Academic, Dordrecht, The Netherlands, 1999, pp. 551–598.
[28] E.D. Sontag and H.J. Sussmann, *Remarks on continuous feedback*, in Proceedings of the IEEE Conference on Decision and Control, Albuquerque, NM, 1980, IEEE, pp. 916–921.
[29] E.D. Sontag and H.J. Sussmann, *Nonsmooth control-Lyapunov functions*, in Proceedings of the IEEE Conference on Decision and Control, New Orleans, LA, 1995, IEEE, pp. 2799–2805.
[30] E.D. Sontag and H.J. Sussmann, *General classes of control-Lyapunov functions*, in Stability Theory (Ascona, 1995), Birkhäuser, Basel, 1996, pp. 87–96.
[31] H.J. Sussmann, *Subanalytic sets and feedback control*, J. Differential Equations, 31 (1979), pp. 31–52.
[32] L. Zajíček, *On the differentiation of convex functions in finite and infinite dimensional spaces*, Czechoslovak Math. J., 29 (1979), pp. 340–348.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 3, pp. 682–699

DYNKIN GAMES VIA DIRICHLET FORMS AND SINGULAR CONTROL OF ONE-DIMENSIONAL DIFFUSIONS*

MASATOSHI FUKUSHIMA† AND MICHAEL TAKSAR‡

Abstract. We consider a zero-sum game of optimal stopping in which each of the opponents has the right to stop a one-dimensional diffusion process. There are two types of costs. The first is accumulated continuously at the rate $H(X_t)$, where X_t is the current position of the process. The second is a cost associated with the stopping of the process. It is given by the function $f_1(x)$ for the first player and the function $f_2(x)$ for the second player, where x is the position of the process when the stopping option is exercised.

We study the solution of the free boundary problem associated with this game via Dirichlet forms on the appropriate functional space. Integrating the value function of the game, we get a solution to another free boundary problem which yields the optimal return function for a singular stochastic control problem.

Key words. Dynkin game, Dirichlet form, free boundary problem, singular stochastic control

AMS subject classifications. 31C25, 60G40, 60H30, 60J55, 93E20

PII. S0363012901387136

1. Introduction.

The reflecting diffusion processes are interesting objects to be studied from a variety of different points of view. In particular, the reflecting Brownian motion on a one-dimensional interval was characterized as a solution of a singular control problem [7, 16]. More specifically, let (w_t, P) be a one-dimensional standard Brownian motion starting at the origin and let

$$X_t = x + \sigma w_t + \mu t + A_t^{(1)} - A_t^{(2)}, \qquad x \in \mathbb{R}, \tag{1.1}$$

where $\sigma \neq 0$, μ are constants, and $\mathbb{S} = (A_t^{(1)}, A_t^{(2)})$ is a pair of nonanticipating increasing processes. $\mathbb{S}$ represents a strategy under which the cost function

$$k_x(\mathbb{S}) = E\left(\int_0^\infty e^{-\alpha t} h(X_t)dt + \int_0^\infty e^{-\alpha t}(rdA_t^{(1)} + \ell dA_t^{(2)})\right) \tag{1.2}$$

is to be minimized. Here, α, r, ℓ are preassigned positive constants and $h(x)$ is a given convex function taking its minimum at the origin. It was then shown by Taksar [16] that there exists an optimal strategy $\tilde{\mathbb{S}}$ such that

$$W(x) = \min_{\mathbb{S}} k_x(\mathbb{S}) = k_x(\tilde{\mathbb{S}})$$

and that $\tilde{\mathbb{S}}$ is actually equal to (ℓ_t^a, ℓ_t^b), where ℓ^a, ℓ^b are local times at points a, b for uniquely determined a, b, $a < 0 < b$. Thus the corresponding optimal process (1.1) is the reflecting diffusion on the closed interval $[a, b]$. The proof in [16] was carried out

*Received by the editors March 28, 2001; accepted for publication (in revised form) March 2, 2002; published electronically July 24, 2002. This work was supported by NSF grant DMS 0072388 and in part by the Kansai University Grant-in-Aid for the Faculty Joint Program, 2000.
http://www.siam.org/journals/sicon/41-3/38713.html

†Department of Mathematics, Faculty of Engineering, Kansai University, Suita, Osaka 564-8680, Japan (fuku@ipcku.kansai-u.ac.jp).

‡Department of Applied Mathematics and Statistics, University at Stony Brook, Stony Brook, NY 11794-3600 (taksar@ams.sunysb.edu).

by solving a related free boundary problem by making use of a solution of an optimal stopping game problem, which had been formulated by Gusein-Zade [6].

The purpose of the present paper is to extend those results in [16] by replacing the process $x + \sigma w_t + \mu t$ appearing in (1.1) on the one hand and constant costs r, ℓ appearing in (1.2) on the other, with a more general diffusion process governed by variable C^1-coefficients $\sigma(x), \mu(x)$ and with variable costs $f_1(x), f_2(x)$, respectively. To this end, we shall employ the Dynkin optimal stopping game and its Dirichlet form characterization due to Zabczyk [18]. As will be explained in section 2, the value function of the Dynkin game for a general symmetric Hunt process was identified in [18] with the solution of a certain variational inequality in a regular Dirichlet space setting. Such an identification had been established by Nagai [14] for a one-sided optimal stopping problem. This sort of an analytic characterization of the stopping game was missing in [6], making the usage of [6] less simple.

We can then proceed along almost the same lines as in [16] in getting the solution of our singular control problem. However, it is more useful to rewrite the infinitesimal generator $\frac{1}{2}\sigma(x)^2 \frac{d^2}{dx^2} + \mu(x)\frac{d}{dx}$ of the controlled diffusion in the Feller canonical form $\frac{d}{dm}\frac{d}{ds}$. The conditions on the data h, f_1, f_2 will be stated in terms of the intrinsic quantities s and m.

In section 3, we shall apply the Dynkin game description of the solution V of a variational inequality presented in [18] to a one-dimensional diffusion with generator $\frac{d}{ds}\frac{d}{dm}$ in showing that an integral function W of V with respect to ds is a solution of a certain free boundary problem involving the operator $\frac{d}{dm}\frac{d}{ds}$, which will then be identified in section 4 with the optimal return function of our singular control of the (σ, μ)-diffusion. The admissible processes X_t to be optimized will be formulated in section 4 by SDE variants of the identity (1.1), and the optimal process will be shown to be the reflecting (σ, μ)-diffusion on the interval specified in the free boundary problem.

We emphasize that our Dirichlet form approach automatically guarantees the quasi continuity (actually the absolute continuity in the present one-dimensional application) of the value function V, which, combined with the saddle point characterization of V, readily implies that its integral function W is the classical solution of the free boundary problem. As a result we get a classical solution to the one-dimensional singular stochastic control problem as opposed to the viscosity solution guaranteed by a general theory (see [3]).

A slight extension of [16] has been considered by Kawabata [11], where the costs r, ℓ were still kept constant, however, and the method of [18] was not utilized.

In a recent paper [10], Karatzas and Wang obtained the same relation as in our case between the value functions of a Dynkin game and a control problem of general bounded variation processes. The method in [10] is more direct and pathwise, but the admissible process to be optimized is purely of bounded variation and the leading martingale part as in our case is absent.

In what follows, $C^k(I)$ (resp., $C_0^k(I)$) will denote the space of k-times continuously differentiable functions (resp., with compact support) on an interval $I \subset \mathbb{R}$, $k = 1, 2$.

2. Dynkin games via Dirichlet forms. Let X be a locally compact separable metric space and m be a positive Radon measure on X with full support. $L^2(X; m)$ denotes the real L^2-space with inner product $(\cdot, \cdot)$. We consider a Dirichlet form $(\mathcal{E}, \mathcal{F})$ on $L^2(X; m)$. By definition, $\mathcal{E}$ is a closed symmetric form with domain $\mathcal{F}$ dense in $L^2(X; m)$ such that the unit contraction operates on it:

$$u \in \mathcal{F} \Longrightarrow v = 0 \vee u \wedge 1 \in \mathcal{F}, \qquad \mathcal{E}(v, v) \leq \mathcal{E}(u, u).$$

Recall that a closed symmetric form is a Dirichlet form if and only if the associated L^2-semigroup $\{T_t, t > 0\}$ is Markovian in the sense that

$$0 \le f \le 1, \qquad f \in L^2 \Longrightarrow 0 \le T_t f \le 1.$$

We let $\mathcal{E}_\alpha(u,v) = \mathcal{E}(u,v) + \alpha(u,v)$ for $\alpha > 0$. We assume that the Dirichlet form $(\mathcal{E}, \mathcal{F})$ is regular in the sense that $\mathcal{F} \cap C_0(X)$ is $\mathcal{E}_1$-dense in $\mathcal{F}$ and uniformly dense in $C_0(X)$, where $C_0(X)$ denotes the space of continuous functions on X with compact support. There exists then a Hunt process (a right continuous, quasi-left continuous strong Markov process) $\mathbf{M} = (X_t, P_x)$ on X such that

$$p_t f(x) = E_x(f(X_t)), \qquad x \in X,$$

is a version of $T_t f$ for all $f \in C_0(X)$ [5].

In what follows, basic notions and relations concerning the regular Dirichlet form $(\mathcal{E}, \mathcal{F})$ and the associated Hunt process $\mathbf{M}$ shall be taken from [5]. In particular, the L^2-resolvent $\{G_\alpha,\ \alpha > 0\}$ associated with the Dirichlet form $(\mathcal{E}, \mathcal{F})$ satisfies

$$G_\alpha f \in \mathcal{F}, \quad \mathcal{E}_\alpha(G_\alpha f, v) = (f, v) \qquad \forall f \in L^2(X; m),\ \forall v \in \mathcal{F},$$

and further the resolvent $\{R_\alpha\ \alpha > 0\}$ of the Hunt process $\mathbf{M}$ defined by

$$R_\alpha f(x) = E_x\left(\int_0^\infty e^{-\alpha t} f(X_t) dt\right), \qquad x \in X,$$

is a quasi-continuous modification of $G_\alpha f$ for any Borel function $f \in L^2(X; m)$. For $v \in \mathcal{F}$, $\tilde{v}$ will denote a quasi-continuous modification of v.

Given $\alpha > 0,\ H \in L^2(X; m)$, and $f_1, f_2 \in \mathcal{F}$ with $-f_1 \le f_2$, we let

$$K = \{u \in \mathcal{F} : -f_1 \le u \le f_2, \quad m\text{-a.e.}\}. \tag{2.1}$$

One looks for a solution $V \in K$ of the inequality

$$\mathcal{E}_\alpha(V, u - V) \ge (H, u - V) \qquad \forall u \in K. \tag{2.2}$$

Such a variational inequality arises in various contexts and goes back to Stampacchia [15].

Proposition 2.1. *There exists a unique function $V \in K$ satisfying* (2.2).

Proof. This is a well-known fact, but we reproduce a proof given by Nagai [14] in a way convenient for later use. First consider the special case that $H = 0$. We can then see the equivalence of the following inequalities holding for $V \in K$:

$$\mathcal{E}_\alpha(V, u - V) \ge 0 \qquad \forall u \in K, \tag{2.3}$$

$$\mathcal{E}_\alpha(V, V) \le \mathcal{E}_\alpha(u, u) \qquad \forall u \in K. \tag{2.4}$$

In fact, (2.3) readily implies (2.4) by the Schwarz inequality. Conversely, suppose (2.4). Take any $u \in K$ and put $w = u - V$. Since K is convex,

$$V + \epsilon w = (1 - \epsilon)V + \epsilon u \in K \qquad \forall \epsilon \in (0, 1).$$

Equation (2.4) then leads us to

$$\mathcal{E}_\alpha(V, V) \le \mathcal{E}_\alpha(V + \epsilon w, V + \epsilon w)$$

and $2\mathcal{E}_\alpha(V,w)+\epsilon\mathcal{E}_\alpha(w,w)\geq 0$. We get (2.3) by letting $\epsilon\downarrow 0$.

Now (2.4) (and equivalently (2.3)) has a unique solution $V\in K$ by virtue of the closedness of the convex set K and the parallelogram law (see, for instance, the proof of [5, Lemma 2.1.2]).

Next consider a general $H\in L^2(X;m)$. By making use of the L^2-resolvent G_α, we can rewrite the inequality (2.2) as

$$\mathcal{E}_\alpha(V-G_\alpha H,(u-G_\alpha H)-(V-G_\alpha H))\geq 0$$

in concluding that the solution V of (2.1) and (2.2) is related to the solution V^0 of

(2.5)
$$K^0=\{u\in\mathcal{F}:-h_1\leq u\leq h_2,\quad m\text{-a.e.}\},\quad h_1=f_1+G_\alpha H,\quad h_2=f_2-G_\alpha H,$$

$$V^0\in K^0,\quad \mathcal{E}_\alpha(V^0,u-V^0)\geq 0\qquad \forall u\in K^0 \tag{2.6}$$

by the relation

$$V=V^0+G_\alpha H. \qquad \square \tag{2.7}$$

Zabczyk has related the solution of the variational inequality (2.2) to the value function of an optimal stopping game (called a *Dynkin game* after [2]) for the associated Hunt process $\mathbf{M}=(X_t,P_x)$ in the following manner [18, Theorem 1].

THEOREM 2.1 (Zabczyk). *For any Borel function $H\in L^2(X;m)$ and for any $f_1,f_2\in\mathcal{F}$ with $-f_1\leq f_2$, we put*

$$\begin{aligned}J_x(\tau,\sigma)&=E_x\left(\int_0^{\tau\wedge\sigma}e^{-\alpha t}H(X_t)dt\right)\\&\quad+E_x\left(e^{-\alpha(\tau\wedge\sigma)}(-I_{\sigma\leq\tau}\tilde{f}_1(X_\sigma)+I_{\tau<\sigma}\tilde{f}_2(X_\tau))\right)\end{aligned} \tag{2.8}$$

for $x\in X$ and for finite stopping times τ,σ. Then the solution of (2.1) *and* (2.2) *admits as its quasi-continuous version the value function of the game*

$$V(x)=\inf_\tau\sup_\sigma J_x(\tau,\sigma)=\sup_\sigma\inf_\tau J_x(\tau,\sigma),\qquad x\in X\setminus N, \tag{2.9}$$

where N is some properly exceptional set with respect to $\mathbf{M}$.

Furthermore if we let

$$E_1=\{x\in X-N:V(x)=-\tilde{f}_1(x)\},\quad E_2=\{x\in X-N:V(x)=\tilde{f}_2(x)\},$$

then the hitting times $\hat{\tau}=\sigma_{E_2}$, $\hat{\sigma}=\sigma_{E_1}$ are the saddle point of the game:

$$J_x(\hat{\tau},\sigma)\leq J_x(\hat{\tau},\hat{\sigma})\leq J_x(\tau,\hat{\sigma}) \tag{2.10}$$

for any $x\in X-N$ and for any stopping times τ, σ. In particular

$$V(x)=J_x(\hat{\tau},\hat{\sigma})\qquad\forall x\in X\setminus N. \tag{2.11}$$

Actually this theorem was shown in [18] only when $H=0$. However, on account of the proof of Proposition 2.1, the statements of Theorem 2.1 for a general Borel function $H\in L^2(X;m)$ can be reduced to this special case. In fact, by what was

proved in [18], the solution of (2.5) and (2.6) admits a quasi-continuous version given by

$$V^0(x) = \inf_\tau \sup_\sigma J_x^0(\tau,\sigma) = \sup_\sigma \inf_\tau J_x^0(\tau,\sigma), \qquad x \in X \setminus N,$$

where N is some properly exceptional set and

$$J_x^0(\tau,\sigma) = E_x\left(e^{-\alpha(\tau\wedge\sigma)}(-I_{\sigma\le\tau}\tilde{h}_1(X_\sigma) + I_{\tau<\sigma}\tilde{h}_2(X_\tau))\right),$$
$$\tilde{h}_1 = \tilde{f}_1 + R_\alpha H, \qquad \tilde{h}_2 = \tilde{f}_2 - R_\alpha H.$$

In view of (2.7), the solution of (2.1) and (2.2) then admits a quasi-continuous version

$$V(x) = V^0(x) + R_\alpha H(x), \qquad x \in X \setminus N,$$

which in turn can be seen to satisfy the identity (2.9), because the Dynkin formula

$$R_\alpha H(x) - E_x\left(e^{-\alpha(\tau\wedge\sigma)}R_\alpha H(X_{\tau\wedge\sigma})\right) = E_x\left(\int_0^{\tau\wedge\sigma} e^{-\alpha t}H(X_t)dt\right)$$

leads us to

$$J_x^0(\tau,\sigma) + R_\alpha H(x) = J_x(\tau,\sigma).$$

The second statement of Theorem 2.1 is also an immediate consequence of that for V^0 and J^0.

We refer to [18] for related literature prior to [18].

3. One-dimensional Dynkin game and free boundary problems. When the underlying space X is one-dimensional, the solution V of the variational inequality (2.1), (2.2) can be described as a solution of a certain free boundary problem. The proof can be carried out using primarily its Dynkin game description (2.9) and (2.10).

More specifically, let $\dot{s}(x)$ and $\dot{m}(x)$ be strictly positive C^1-functions on $\mathbb{R}$. Denote the one-dimensional Lebesgue measure by dx and the measures $\dot{s}(x)dx$, $\dot{m}(x)dx$ by ds, dm, respectively. We assume that both $-\infty$ and ∞ are natural (neither exit nor entrance) boundaries of $\mathbb{R}$ with respect to s, m in Feller's sense [9]:

$$\int_{-\infty<y<x<-1} ds(x)dm(y) = \infty, \qquad \int_{-\infty<y<x<-1} dm(x)ds(y) = \infty,$$
$$\int_{1<x<y<\infty} ds(x)dm(y) = \infty, \qquad \int_{1<x<y<\infty} dm(x)ds(y) = \infty. \tag{3.1}$$

For $A > 0$, we let

$$\mathcal{F} = H^1((-A,A);dx)$$
$$= \{u \in L^2((-A,A);dx) : u \text{ is absolutely continuous, } u' \in L^2((-A,A);dx)\}, \tag{3.2}$$

$$\mathcal{E}(u,v) = \int_{-A}^{A} u'(x)v'(x)\,\frac{1}{\dot{m}(x)}\,dx, \qquad u,v \in \mathcal{F}. \tag{3.3}$$

We can and shall regard $(\mathcal{E},\mathcal{F})$ as a regular local Dirichlet form on $L^2([-A,A];ds)$. The associated Hunt process $\mathbf{M} = (X_t, P_x)$ on the closed interval $[-A,A]$ is a conservative diffusion process, namely, a strong Markov process with continuous sample

paths and infinite life time, and actually it is a reflecting barrier diffusion on $[-A, A]$ with infinitesimal generator $\frac{d}{ds}\frac{d}{dm}$. Since $\mathcal{F}$ is the ordinary Sobolev space $H^1(-A, A)$ on the one-dimensional interval $(-A, A)$ and the metric $\mathcal{E}_1$ on it is equivalent to the square root of the Dirichlet integral plus L^2-norm, we see that each one point set $\{x\} \subset [-A, A]$ has a positive capacity, the quasi continuity reduces to the ordinary continuity, and $\mathbf{M}$ admits no nonempty exceptional set [5, Example 2.1.2].

We now let $V(x)$, $x \in [-A, A]$, be the solution of the variational inequality (2.1), (2.2) for the present Dirichlet form $(\mathcal{E}, \mathcal{F})$ on $L^2([-A, A]; ds)$ under the following assumptions on the data H, f_1, f_2.

ASSUMPTION 3.1. *$H(x)$ is a continuous function on $\mathbb{R}$ such that*

$$H(0) = 0, \quad H(x) \text{ is strictly increasing}, \quad H(x) \to \pm\infty \text{ as } x \to \pm\infty.$$

f_1, f_2 are C^2-functions with $0 < f_1, f_2 \le M$ for some $M > 0$ and

$$f_1'(x) \ge 0, \quad x \in \mathbb{R}, \qquad \frac{d}{ds}\frac{d}{dm} f_1 - H \text{ is strictly decreasing on } (-\infty, 0),$$

$$f_2'(x) \le 0, \quad x \in \mathbb{R}, \qquad \frac{d}{ds}\frac{d}{dm} f_2 + H \text{ is strictly increasing on } (0, \infty).$$

Remark 3.1. (i) The assumptions for f_1, f_2 are trivially satisfied by $f_1 = r, f_2 = \ell$ positive constant functions.

(ii) In the next section, we shall be concerned with controls of a diffusion with generator $\frac{d}{dm}\frac{d}{ds}$, a diffusion with scale ds, and speed measure dm in the sense of Feller [9]. For that purpose, we need to consider in the first part of this section a diffusion with the roles of ds and dm being interchanged.

LEMMA 3.1. *There exists $A > 0$ such that the diffusion $\mathbf{M} = (X_t, P_x)$ on $[-A, A]$ associated with the Dirichlet form* (3.2), (3.3) *satisfies*

$$E_{\xi_1}\left(\int_0^{\sigma_0 \wedge \sigma_{-A}} e^{-\alpha t} H(X_t) dt\right) < -2M, \quad E_{\xi_2}\left(\int_0^{\sigma_0 \wedge \sigma_A} e^{-\alpha t} H(X_t) dt\right) > 2M \tag{3.4}$$

for some $\xi_1 \in (-A, 0)$ and $\xi_2 \in (0, A)$. Here σ_x denotes the hitting time of the one point set $\{x\}$.

Proof. For an open interval $I \subset \mathbb{R}$, we denote by $\mathbb{D}^I$ the absorbing diffusion on I with infinitesimal generator $\frac{d}{ds} \cdot \frac{d}{dm}$ and by R_α^I its resolvent operator. By virtue of the condition (3.1), $\mathbb{D}^{\mathbb{R}}$ is conservative and its α-order hitting probability $E_x\left(e^{-\alpha\sigma_c}\right)$ for any fixed point c tends to zero as $x \to \pm\infty$ [9]. Hence, by Dynkin's formula,

$$\lim_{x \to -\infty} R_\alpha^{(-\infty, c)} 1(x) = \frac{1}{\alpha}, \qquad \lim_{x \to \infty} R_\alpha^{(d, \infty)} 1(x) = \frac{1}{\alpha} \tag{3.5}$$

for any c and d. By Assumption 3.1, we can take $\xi < 0$ such that

$$H(x) < -4\alpha M \qquad \forall x \le \xi.$$

By (3.5), $R_\alpha^{(-\infty, \xi)} 1(\xi_1) > 1/(2\alpha)$ for some $\xi_1 < \xi$. Since $R_\alpha^{(-A, \xi)} 1(\xi_1)$ increases to $R_\alpha^{(-\infty, \xi)} 1(\xi_1)$ as $A \to \infty$, we have $R_\alpha^{(-A, \xi)} 1(\xi_1) > 1/(2\alpha)$ for a sufficiently large A with $-A < \xi_1$.

For such A, let $\mathbf{M}$ be the diffusion on $[-A, A]$ governed by the Dirichlet form (3.2), (3.3). Since the process obtained from $\mathbf{M}$ by killing at time $\sigma_0 \wedge \sigma_{-A}$ coincides

with $\mathbb{D}^{(-A,0)}$, the first expectation appearing in (3.4) equals $R_\alpha^{(-A,0)}H(\xi_1)$, which in turn is not greater than

$$R_\alpha^{(-A,\xi)}H(\xi_1) \le -4\alpha M \cdot R_\alpha^{(-A,\xi)}1(\xi_1) < -2M,$$

proving the first inequality in (3.4). The second one can be shown in the same way. □

In what follows, we shall work with $A > 0$ for which (3.4) is satisfied.

THEOREM 3.1. *There exist unique a, b such that $-A < a < 0 < b < A$ and*

$$-f_1(x) < V(x) < f_2(x), \qquad x \in (a,b), \tag{3.6}$$

$$V(x) = -f_1(x), \quad x \in [-A,a], \qquad V(x) = f_2(x), \quad x \in [b,A], \tag{3.7}$$

$$V'(a) = -f_1'(a), \qquad V'(b) = f_2'(b). \tag{3.8}$$

Furthermore V is C^1 on $(-A,A)$, C^2 on (a,b) and

$$\begin{aligned} \alpha V(x) - \frac{d}{ds}\frac{d}{dm}V(x) &= H(x) && \forall x \in [a,b] \\ &> H(x) && \forall x \in (-A,a) \\ &< H(x) && \forall x \in (b,A). \end{aligned} \tag{3.9}$$

The theorem is divided into three propositions.

PROPOSITION 3.1. (i) $-f_1(0) < V(0) < f_2(0)$.

(ii) $V(x) > -f_1(x)$ *for* $x > 0$ *and* $V(x) < f_2(x)$ *for* $x < 0$.

(iii) *Let*

$$E_1 = \{x \in [-A,A] : V(x) = -f_1(x)\}, \quad E_2 = \{x \in [-A,A] : V(x) = f_2(x)\} \tag{3.10}$$

and $a = \sup E_1$, $b = \inf E_2$. Then

$$-A < a < 0 < b < A.$$

(iv) *If*

$$-f_1(x) < V(x) < f_2(x), \qquad \beta < x < \gamma,$$

for some interval $(\beta,\gamma) \subset (-A,A)$, then V is C^2 on (β,γ) and

$$\left(\alpha - \frac{d}{ds}\frac{d}{dm}\right)V(x) = H(x) \tag{3.11}$$

for $x \in (\beta,\gamma)$. In particular, this equation holds for $x \in (a,b)$.

(v) *If, for some $\beta \in (-A,0)$,*

$$-f_1(x) < V(x), \qquad -A \le x < \beta,$$

then V is C^2 on $(-A,\beta)$, V satisfies (3.11) on $(-A,\beta)$, and $V'(-A) = 0$.

(vi) *If, for some $\gamma \in (0,A)$,*

$$V(x) < f_2(x), \qquad \gamma < x \le A,$$

then V is C^2 on (γ, A), V satisfies (3.11) on (γ, A), and $V'(A) = 0$.

Proof. Denote by σ_E the hitting time of the diffusion $\mathbf{M}$ for a set E. The hitting time for the one point set $\{x\}$ is simply denoted by σ_x. We let $\hat{\sigma} = \sigma_{E_1}$, $\hat{\tau} = \sigma_{E_2}$ be the hitting times for the sets E_1, E_2 defined by (3.10).

(i) We give the proof of the first inequality. The second one can be proved similarly. We have from (2.10) and (2.11) (the exceptional set N is now empty, as was explained in the paragraph below (3.3)) that, for any positive $\epsilon < A$,

$$\begin{aligned} V(0) \geq J_0(\hat{\tau}, \sigma_{-\epsilon}) &= E_0\left(\int_0^{\hat{\tau}\wedge\sigma_{-\epsilon}} e^{-\alpha t} H(X_t)dt\right) \\ &\quad - f_1(-\epsilon)E_0\left(e^{-\alpha\sigma_{-\epsilon}}; \sigma_{-\epsilon} < \hat{\tau}\right) + E_0\left(e^{-\alpha\hat{\tau}} f_2(X_{\hat{\tau}}); \sigma_{-\epsilon} \geq \hat{\tau}\right) \\ &\geq H(-\epsilon)E_0\left(\int_0^{\sigma_{-\epsilon}} e^{-\alpha t}dt\right) - f_1(0)E_0\left(e^{-\alpha\sigma_{-\epsilon}}\right) \\ &= -f_1(0) + \left(f_1(0) + \frac{H(-\epsilon)}{\alpha}\right)\left(1 - E_0(e^{-\alpha\sigma_{-\epsilon}})\right), \end{aligned}$$

which is greater than $-f_1(0)$ for sufficiently small $\epsilon > 0$.

(ii) For $x > 0$,

$$\begin{aligned} V(x) \geq J_x(\hat{\tau}, \sigma_0) &= E_x\left(\int_0^{\hat{\tau}\wedge\sigma_0} e^{-\alpha t} H(X_t)dt\right) \\ &\quad - f_1(0)E_x\left(e^{-\alpha\sigma_0}; \sigma_0 < \hat{\tau}\right) + E_x\left(e^{-\alpha\hat{\tau}} f_2(X_{\hat{\tau}}); \sigma_0 \geq \hat{\tau}\right) \\ &\geq -f_1(0)E_x\left(e^{-\alpha\sigma_0}\right) > -f_1(0) \geq -f_1(x). \end{aligned}$$

The second inequality can be proved similarly.

(iii) Suppose that $V(x) > -f_1(x)$ for any $x \in (-A, 0)$. Then, by (i) and (ii), $P_x(\hat{\sigma} \geq \sigma_{-A}) = 1$ for all x. Further $P_x(\hat{\tau} > \sigma_0) = 1$ for any $x < 0$. Hence

$$P_x(\hat{\sigma} \wedge \hat{\tau} \geq \sigma_{-A} \wedge \sigma_0) = 1 \qquad \forall x < 0,$$

which implies that the function $V(x) = J_x(\hat{\tau}, \hat{\sigma})$ is H-α-harmonic on $(-A, 0)$ in the sense that, for $x \in (-A, 0)$,

$$V(x) = E_x\left(\int_0^{\sigma_0\wedge\sigma_{-A}} e^{-\alpha t} H(X_t)dt\right) + E_x\left(e^{-\alpha(\sigma_0\wedge\sigma_{-A})} V(X_{\sigma_0\wedge\sigma_{-A}})\right). \tag{3.12}$$

Since $V(x) \leq M$ for any $x \in [-A, A]$, we get from the above and (3.4)

$$V(\xi_1) < -2M + M = -M,$$

a contradiction. Hence $-A < a < 0$. The second inequality can be proved similarly.

(iv) As in the proof of (iii), V is then H-α-harmonic on the interval (β, γ) in the sense that the identity (3.12) with $\sigma_0 \wedge \sigma_{-A}$ being replaced by $\sigma_\beta \wedge \sigma_\gamma$ holds for $x \in (\beta, \gamma)$, which is equivalent to the validity of the following equation [5, section 4.3, section 4.4]:

$$\mathcal{E}_\alpha(V, v) = (H, v) \qquad \forall v \in C_0^1((\beta, \gamma)). \tag{3.13}$$

Since H is continuous, this equation in turn implies that V is C^2 on (β, γ), and an integration by parts yields (3.11) on the same interval.

(v) In this case, the identity (3.12) with $\sigma_0 \wedge \sigma_{-A}$ being replaced by $\sigma_\beta \wedge \sigma_{-A}$ holds for $x \in [-A, \beta)$, which is equivalent to the validity of (3.13) for any $v \in C_0^1([-A, \beta))$. Again, an integration by parts gives the validity of (3.11) on $(-A, \beta)$ together with the stated boundary condition.

(vi) is analogous to (v). □

Before proceeding further, we prepare some notation. For $\xi \in (-A, A)$ and $\epsilon > 0$, we denote by $\tau_{\xi,\epsilon}$ the first exit time from the interval $I_{\xi,\epsilon} = (\xi - \epsilon, \xi + \epsilon)$, namely, $\tau_{\xi,\epsilon} = \sigma_{[-A,A]\setminus I_{\xi,\epsilon}}$. We then set

$$h_\alpha^-(\xi,\epsilon) = E_\xi\left(e^{-\alpha\tau_{\xi,\epsilon}}; \sigma_{\xi-\epsilon} < \sigma_{\xi+\epsilon}\right),$$
$$h_\alpha^+(\xi,\epsilon) = E_\xi\left(e^{-\alpha\tau_{\xi,\epsilon}}; \sigma_{\xi-\epsilon} \geq \sigma_{\xi+\epsilon}\right),$$
$$g_\alpha(\xi,\epsilon) = 1 - E_\xi\left(e^{-\alpha\tau_{\xi,\epsilon}}\right).$$

LEMMA 3.2.

$$\lim_{\epsilon \downarrow 0} h_\alpha^\pm(\xi,\epsilon) = \frac{1}{2}.$$

$$g_\alpha(\xi,\epsilon) = o(\epsilon) \quad as \quad \epsilon \downarrow 0.$$

Proof. The first identity for $\alpha = 0$ is evident because

$$h_0^-(\xi,\epsilon) = \frac{\int_\xi^{\xi+\epsilon} \dot{m}(x)dx}{\int_{\xi-\epsilon}^{\xi+\epsilon} \dot{m}(x)dx}. \tag{3.14}$$

Let u be a C^2-function vanishing at $-A$ and A such that

$$\frac{d}{ds}\frac{d}{dm}u(x) = -1, \quad x \in (\xi - \epsilon, \xi + \epsilon).$$

By Dynkin's formula applied to the 0-order resolvent of the process obtained from $\mathbf{M}$ by killing at time $\sigma_{-A} \wedge \sigma_A$,

$$E_\xi(\tau_{\xi,\epsilon}) = u(\xi) - h_0^-(\xi,\epsilon)u(\xi-\epsilon) - h_0^+(\xi,\epsilon)u(\xi+\epsilon),$$

which combined with (3.14) leads us to $E_\xi(\tau_{\xi,\epsilon}) = o(\epsilon)$.

The rest of the proof is obvious since

$$g_\alpha(\xi,\epsilon) = \alpha E_\xi\left(\int_0^{\tau_{\xi,\epsilon}} e^{-\alpha t}dt\right) \leq \alpha E_\xi(\tau_{\xi,\epsilon}). \qquad \square$$

PROPOSITION 3.2. (i) *$V'(a) = -f_1'(a)$ and $V'(x)$ is right continuous at a.*

(ii) *$V'(b) = f_2'(b)$ and $V'(x)$ is left continuous at b.*

Proof. We give the proof of (i) only. The proof of (ii) is analogous. Take any $\epsilon > 0$ with $(a - \epsilon, a + \epsilon) \subset (-A, 0)$. Let θ_t be the shift operator on the probability space Ω for $\mathbf{M}$, that is, $X_s(\theta_t\omega) = X_{s+t}(\omega)$ for all $\omega \in \Omega$ (cf. [5]). If we let $\sigma = \tau_{a,\epsilon} + \hat{\sigma} \circ \theta_{\tau_{a,\epsilon}}$, then

$$\hat{\tau} \wedge \sigma = \tau_{a,\epsilon} + (\hat{\tau} \wedge \hat{\sigma}) \circ \theta_{\tau_{a,\epsilon}},$$

because $\hat{\tau} = \tau_{a,\epsilon} + \hat{\tau} \circ \theta_{\tau_{a,\epsilon}}$. Hence we have

$$V(a) \geq J_a(\hat{\tau}, \sigma) = E_a\left(\int_0^{\tau_{a,\epsilon}} e^{-\alpha t}H(X_t)dt\right) + h_\alpha^-(a,\epsilon)V(a-\epsilon) + h_\alpha^+(a,\epsilon)V(a+\epsilon)$$

and

$$h_\alpha^-(a,\epsilon)V(a-\epsilon)+h_\alpha^+(a,\epsilon)V(a+\epsilon)-E_a\left(e^{-\alpha\tau_{a,\epsilon}}\right)V(a)$$
$$\le g_\alpha(a,\epsilon)V(a)-E_a\left(\int_0^{\tau_{a,\epsilon}}e^{-\alpha t}H(X_t)dt\right)\le\left(V(a)-\frac{H(a-\epsilon)}{\alpha}\right)g_\alpha(a,\epsilon).$$

Therefore

$$h_\alpha^+(a,\epsilon)(-f_1(a+\epsilon)+f_1(a))\le h_\alpha^+(a,\epsilon)(V(a+\epsilon)-V(a))$$
$$\le h_\alpha^-(a,\epsilon)(V(a)-V(a-\epsilon))+\left(V(a)-\frac{H(a-\epsilon)}{\alpha}\right)g_\alpha(a,\epsilon)$$
$$\le h_\alpha^-(a,\epsilon)(-f_1(a)+f_1(a-\epsilon))+\left(V(a)-\frac{H(a-\epsilon)}{\alpha}\right)g_\alpha(a,\epsilon).$$

By dividing each side of the above inequality by ϵ and letting $\epsilon\to 0$, we get from the previous lemma the desired inequality

$$-D_+f_1(a)\le D_+V(a)\le D_-V(a)\le -D_-f_1(a),$$

yielding the first half of (i). Since $V'(x)$ is easily seen to have the right limit at $x=a$ by virtue of Proposition 3.1(iv), it is right continuous at a as well. □

PROPOSITION 3.3. *Let E_1, E_2 be the sets defined by* (3.10).

(i) $E_1=[-A,a]$ *and*

$$\left(\alpha-\frac{d}{ds}\frac{d}{dm}\right)f_1(x)>H(x)\qquad\forall x\in[-A,a).$$

(ii) $E_2=[b,A]$ *and*

$$\left(\alpha-\frac{d}{ds}\frac{d}{dm}\right)f_2(x)<H(x)\qquad\forall x\in(b,A].$$

Proof. We give the proof of (i) only. (ii) can be proved similarly. Putting $x=a+\epsilon$ in (3.11) and letting $\epsilon\downarrow 0$, we get

$$\alpha V(a)-\frac{d+}{ds}\frac{dV}{dm}(a)=H(a), \tag{3.15}$$

where $\frac{d+}{ds}$ denotes the right derivative. On the other hand,

$$\frac{d+}{ds}\frac{dV}{dm}(a)\ge-\frac{d}{ds}\frac{d}{dm}f_1(a). \tag{3.16}$$

In fact, the function $F(x)=V(x)+f_1(x)$ satisfies $F(x)\ge 0, F(a)=0$, and further $F'(a)=0$, $F'(x)$ is right continuous at a by the preceding proposition. Taylor's theorem applies and

$$0\le\frac{F(a+\epsilon)}{\epsilon^2}=F''(a+\theta\epsilon)\to\frac{d+}{dx}F'(a)\qquad\text{as}\quad\epsilon\downarrow 0.$$

Hence

$$\frac{d+}{ds}\frac{dF}{dm}(a)=\frac{1}{\dot s(a)\dot m(a)}\frac{d+}{dx}F'(a)-\frac{\dot m'(a)}{\dot s(a)\dot m(a)^2}F'(a)=\frac{1}{\dot s(a)\dot m(a)}\frac{d+}{dx}F'(a)\ge 0.$$

Now (3.15), (3.16), and Assumption 3.1 lead us to the inequality

$$-\left(\alpha - \frac{d}{ds}\frac{d}{dm}\right) f_1(x) > H(x) \qquad \forall x \in [-A, a). \tag{3.17}$$

Turning to the proof of $E_1 = [-A, a]$ by reduction to a contradiction, we assume that there exists $x_0 \in [-A, a)$ such that $V(x_0) > -f_1(x_0)$. Then we have two possibilities:

(I) There exists $\beta, \gamma \in E_1$ such that $-A \le \beta < x_0 < \gamma \le a$ and $V(x) > -f_1(x)$ for all $x \in (\beta, \gamma)$.

(II) There exists $\beta \in E_1$ such that $-A < x_0 < \beta$ and $V(x) > -f_1(x)$ for all $x \in [-A, \beta)$.

Suppose case (I) occurs. By combining Proposition 3.1(iv) with (3.17), we then see that the function $F = -f_1 - V$ satisfies

$$\left(\alpha - \frac{d}{ds}\frac{d}{dm}\right) F(x) > 0 \tag{3.18}$$

for any $x \in (\beta, \gamma)$. Since $F(\beta) = F(\gamma) = 0$, an integration by parts yields

$$\mathcal{E}_\alpha(F, v) \ge 0 \tag{3.19}$$

for any $v \in C_0^1((\beta, \gamma))$ such that $v \ge 0$. This means that (the restriction to (β, γ) of) F is α-excessive with respect to the part of the Dirichlet form $(\mathcal{E}, \mathcal{F})$ on the interval (β, γ) (see [5, Lemma 2.2.1, Theorem 4.4.3]). In particular, $F(x_0) \ge 0$, a contradiction.

Suppose case (II) occurs. On account of Proposition 3.1(v), (3.17), and Assumption 3.1, we see then that the function F satisfies inequality (3.18) holding for any $x \in (-A, \beta)$ as well as the inequality $F'(-A) \le 0$. Therefore, an integration by parts leads us to the inequality (3.19) holding for any $v \in C_0^1([-A, \beta))$ such that $v \ge 0$. F is then α-excessive with respect to the part of $(\mathcal{E}, \mathcal{F})$ on the interval $[-A, \beta)$, again arriving at the contradiction $F(x_0) \ge 0$. □

By the preceding three propositions, the proof of Theorem 3.1 is complete.

The function V of Theorem 3.1 (the solution of (2.1), (2.2) for the Dirichlet form (3.3) on $L^2([-A, A], ds)$ under the Assumption 3.1 for the data (H, f_1, f_2)) gives rise to a solution of another type of free boundary problem stated below. Let us first extend the function V to whole $\mathbb{R}$ by setting

$$V(x) = -f_1(x), \quad x < -A, \qquad V(x) = f_2(x), \quad x > A. \tag{3.20}$$

In view of Assumption 3.1, we see that the extended function V still satisfies the first inequality of (3.9) on $(-\infty, a)$ and the second inequality on (b, ∞).

We then let, for $x \in \mathbb{R}$,

$$h(x) = \int_0^x H(y)\dot{s}(y)dy + C, \tag{3.21}$$

where C is an arbitrarily taken fixed constant. We further let

$$W(x) = \int_a^x V(y)\dot{s}(y)dy + \frac{1}{\alpha}\left(-\frac{f_1'(a)}{\dot{m}(a)} + h(a)\right). \tag{3.22}$$

THEOREM 3.2. *$W \in C^2(\mathbb{R})$ and there exist a, b with $a < 0 < b$ such that*

$$\begin{aligned} \alpha W(x) - \frac{d}{dm}\frac{d}{ds}W(x) &= h(x), \qquad a < x < b \\ &< h(x), \qquad x < a \ \text{ or } \ x > b, \end{aligned} \tag{3.23}$$

$$-f_1 < \frac{d}{ds}W < f_2 \qquad \text{on} \quad (a, b), \tag{3.24}$$

$$\frac{d}{ds}W = -f_1 \quad \text{on } (-\infty, a], \qquad \frac{d}{ds}W = f_2 \quad \text{on } [b, \infty), \tag{3.25}$$

$$\frac{d}{dx}\frac{d}{ds}W(a) = -f_1'(a), \qquad \frac{d}{dx}\frac{d}{ds}W(b) = f_2'(b). \tag{3.26}$$

Proof. For the function

$$U(x) = \alpha W(x) - \frac{d}{dm}\frac{d}{ds}W(x) - h(x),$$

we have

$$\frac{1}{\dot{s}(x)}U'(x) = \alpha V(x) - \frac{d}{ds}\frac{d}{dm}V(x) - H(x).$$

Consider a, b of Theorem 3.1. Then, by Theorem 3.1 and the remark made before the statement of Theorem 3.2,

$$U(a) = 0; \qquad U'(x) > 0,\ x < a; \qquad U'(x) = 0,\ x \in (a, b); \qquad U'(x) < 0,\ x > b,$$

which implies (3.23). The rest of the proof is obvious. □

4. A singular control of the (σ, μ)-diffusion. Let $\sigma(x)$ and $\mu(x)$ be C^1-functions on $\mathbb{R}$ with $\sigma(x) \neq 0$ for all $x \in \mathbb{R}$. We are concerned with a diffusion on $\mathbb{R}$ with infinitesimal generator

$$Lu(x) = \frac{1}{2}\sigma(x)^2 \frac{d^2u}{dx^2}(x) + \mu(x)\frac{du}{dx}(x), \tag{4.1}$$

which can be converted into the Feller canonical form $\frac{d}{dm}\frac{du}{ds}(x)$ by setting

$$\dot{s}(x) = \exp\left(-\int_0^x \frac{2\mu(y)}{\sigma(y)^2}dy\right), \qquad \dot{m}(x) = \frac{2}{\sigma(x)^2}\exp\left(\int_0^x \frac{2\mu(y)}{\sigma(y)^2}dy\right), \tag{4.2}$$

and $ds(x) = \dot{s}(x)dx$, $dm(x) = \dot{m}(x)dx$. We assume that $-\infty$ and ∞ are natural boundaries with respect to the operator (4.1) in the sense that condition (3.1) is satisfied by $\dot{s}$, $\dot{m}$ of (4.2). Since $\dot{s}$, $\dot{m}$ of (4.2) are strictly positive C^1-functions, all results of section 3 apply.

Throughout this section, we fix $\sigma(x), \mu(x)$ as above and $\dot{s}(x), \dot{m}(x)$ are understood to be defined by (4.2). We call a triplet (S, X, A) *admissible policy* or just *admissible* if the following conditions are satisfied:

(𝔸.1) S is a compact interval of $\mathbb{R}$.

(𝔸.2) There is a filtered measurable space $(\Omega, \{\mathcal{F}_t\}_{t\geq 0})$ subject to usual conditions and probability measures $\{P_x\}_{x\in S}$ on it such that

$X = \{X_t\}_{t\geq 0}$ is an $\{\mathcal{F}_t\}$-adapted right continuous process, and

$A = \{A_t\}_{t\geq 0}$ is an $\{\mathcal{F}_t\}$-adapted right continuous process of bounded variation satisfying

$$E_x\left(\int_{0-}^{\infty} e^{-\alpha t} dA_t^{(1)}\right) < \infty, \quad E_x\left(\int_{0-}^{\infty} e^{-\alpha t} dA_t^{(2)}\right) < \infty, \qquad \forall x \in S, \tag{4.3}$$

where $A^{(1)}$ and $A^{(2)}$ are two $\{\mathcal{F}_t\}$-adapted right continuous increasing processes for which $A_t = A_t^{(1)} - A_t^{(2)}$ is the *minimal* decomposition of the bounded variation process A into a difference of two increasing processes.

(𝔸.3) There is an $\{\mathcal{F}_t\}$-adapted standard Brownian motion $\{w_t\}_{t\geq 0}$ starting at the origin under P_x for any $x \in S$ such that the stochastic differential equation

$$X_t = x + \int_0^t \sigma(X_s) dw_s + \int_0^t \mu(X_s) ds + A_t^{(1)} - A_t^{(2)}, \qquad t \geq 0, \tag{4.4}$$

holds P_x-a.s. for each $x \in S$, and further

$$P_x(X_t \in S \ \forall t \geq 0) = 1 \qquad \forall x \in S. \tag{4.5}$$

We denote by $\mathbb{A}$ the totality of admissible triplets (S, X, A). In what follows we will always represent A in terms of $A^{(1)}$ and $A^{(2)}$ and thus write (S, X, A) and $(S, X, A^{(1)}, A^{(2)})$ interchangeably.

Remark 4.1. (i) The probability space Ω with the filtration $\{\mathcal{F}_t\}$ in (𝔸.2) is not fixed a priori. It is a part of an admissible policy. The filtration $\{\mathcal{F}_t\}$ is assumed to be right continuous and $\mathcal{F}_0$ is assumed to contain every Ω-set which is P_x-negligible for any $x \in S$.

(ii) We shall use the notation

$$\Delta A_t^{(i)} = A_t^{(i)} - A_{t-}^{(i)}, \qquad t \geq 0, \quad i = 1, 2,$$

$$\Delta X_t = X_t - X_{t-}, \quad \Delta u(X)_t = u(X_t) - u(X_{t-}), \qquad t \geq 0.$$

Note that, due to the fact that $A^{(1)}$ and $A^{(2)}$ represent the *minimal* decomposition of A into two increasing processes, $\Delta A_t^{(1)} \Delta A_t^{(2)} = 0$ for each $t \geq 0$. By convention, we let

$$w_t = 0, \quad A_t^{(i)} = 0 \qquad \forall t < 0, \quad i = 1, 2,$$

so that

$$\Delta A_0^{(i)} = A_0^{(i)}, \quad i = 1, 2, \qquad X_{0-} = x \quad P_x\text{-a.s.} \quad \forall x \in S.$$

Further we define the continuous part of $A^{(i)}$ by

$$A_t^{(i),c} = A_t^{(i)} - \sum_{0\leq s\leq t} \Delta A_s^{(i)}, \qquad t \geq 0, \quad i = 1, 2.$$

(iii) The integrals in t in (4.3) involve the possible jump at 0 so that they are the sum of the integrals over $(0, \infty)$ and $A_0^{(i)}$, $i = 1, 2$.

PROPOSITION 4.1. *Let* $(S, X, A^{(1)}, A^{(2)}) \in \mathbb{A}$. *Then, for any* $u \in C^2(\mathbb{R})$, *the following identity holds:*

$$\begin{aligned} u(x) = {} & E_x\left[\int_0^\infty e^{-\alpha t}\left(\alpha - \frac{d}{dm}\frac{d}{ds}\right)u(X_t)dt\right] \\ & + E_x\left[\int_0^\infty e^{-\alpha t}\left(-\frac{du}{ds}(X_t)\dot{s}(X_t)dA_t^{(1),c} + \frac{du}{ds}(X_t)\dot{s}(X_t)dA_t^{(2),c}\right)\right] \\ & - E_x\left[\sum_{0\le t<\infty} e^{-\alpha t}\Delta u(X)_t\right], \qquad x \in S. \end{aligned} \tag{4.6}$$

All expectations in the right side of (4.6) *exist and are finite.*

Proof. By a generalized Ito formula ([13, p. 278]; see also [7, section 4]) applied to the semimartingale (4.4), we have

$$\begin{aligned} e^{-\alpha t}u(X_t) = {} & u(X_0) - \alpha\int_0^t e^{-\alpha s}u(X_s)ds + \int_0^t e^{-\alpha s}u'(X_s)\sigma(X_s)dw_s \\ & + \int_0^t e^{-\alpha s}u'(X_s)\mu(X_s)ds + \int_0^t e^{-\alpha s}u'(X_s)(dA_s^{(1),c} - dA_s^{(2),c}) \\ & + \frac{1}{2}\int_0^t e^{-\alpha s}u''(X_s)\sigma(X_s)^2ds + \sum_{0<s\le t} e^{-\alpha s}\Delta u(X)_s. \end{aligned} \tag{4.7}$$

Rewrite the sum of two terms in the right side of (4.7) as

$$u(X_0) + \sum_{0<s\le t} e^{-\alpha s}\Delta u(X)_s = u(X_{0-}) + \sum_{0\le s\le t} e^{-\alpha s}\Delta u(X)_s,$$

then take the expectation of the both sides of (4.7) with respect to P_x and let $t \to \infty$ to get the identity (4.6). □

LEMMA 4.1. *If* $(S, X, A^{(1)}, A^{(2)}) \in \mathbb{A}$, *then both* $A^{(1)}$ *and* $A^{(2)}$ *are nontrivial in the sense that, for each* $T > 0$,

$$P_x(A_t^{(i)} = A_0^{(i)} \quad \forall t \in [0, T)) = 0 \qquad \forall x \in S, \quad i = 1, 2. \tag{4.8}$$

Proof. Since S is compact, the integrand of the first integral of the right-hand side of (4.4) is bounded and is bounded away from zero, while the integrand of the second is bounded. If both $A^{(1)}$, $A^{(2)}$ were trivial, the process X_t satisfying (4.4) hits therefore any point of $\mathbb{R}$ almost surely as the Brownian motion does [8, pp. 85, pp. 437], a contradiction. If either $A^{(1)}$ or $A^{(2)}$ is trivial, the path of X_t cannot be concentrated on a compact set, again a contradiction. □

PROPOSITION 4.2. *For any finite* $\beta_1 < \beta_2$, *there exists* $([\beta_1, \beta_2], X, A^{(1)}, A^{(2)}) \in \mathbb{A}$ *such that*

$$A_t^{(i)} = \int_0^t I_{\{\beta_i\}}(X_s)dA_s^{(i)} \quad \forall t \ge 0, \qquad P_x\text{-a.s.} \quad \forall x \in [\beta_1, \beta_2], \qquad i = 1, 2. \tag{4.9}$$

Such X_t *and* $A_t^{(i)}$, $i = 1, 2$, *are necessarily continuous in* $t \ge 0$, P_x-a.s. *for any* $x \in [\beta_1, \beta_2]$. *Furthermore, the* P_x*-law of such* $(X, A^{(1)}, A^{(2)})$ *is unique for any* $x \in [\beta_1, \beta_2]$.

Proof. Equation (4.4) subjected to the conditions (4.5) and (4.9) is called the *Skorohod equation* for $[\beta_1, \beta_2]$.

Since $\sigma,\ \mu$ are C^1-functions, the existence and uniqueness of $(X, A^{(1)}, A^{(2)})$ satisfying (4.9) and all admissibility conditions except for the integrability (4.3) follow from Tanaka [17, Theorem 4.1], where the unique existence of the strong solution of the Skorohod equation with Lipschitz continuous coefficients for a multidimensional convex domain was proved. It was also shown in [17] that the solution is necessarily continuous. The integrability (4.3) is then an automatic consequence of (4.7) applied to the C^2-function u such that $u'(\beta_1) = 1,\ u'(\beta_2) = 0$ (resp., $u'(\beta_1) = 0,\ u'(\beta_2) = 1$). □

The triple $(X, A^{(1)}, A^{(2)})$ of Proportion 4.2 is called a *reflecting* (σ, μ)*-diffusion* on the interval $[\beta_1, \beta_2]$.

We are now in the position to formulate our main theorem about a singular control problem for the admissible family $\mathbb{A}$.

Let h, f_1, f_2 be functions on $\mathbb{R}$ satisfying the following conditions.

ASSUMPTION 4.1. *$h(x)$ is a C^1-function on $\mathbb{R}$ such that*

$$h'(0) = 0, \quad \frac{dh}{ds}(x) \text{ is strictly increasing}, \qquad \frac{dh}{ds}(x) \to \pm\infty \quad \text{as } x \to \pm\infty.$$

$f_1,\ f_2$ are C^2-functions with $0 < f_1, f_2 \le M$ for some $M > 0$ and

$$f_1'(x) \ge 0, \quad x \in \mathbb{R}, \qquad \frac{d}{ds}\frac{d}{dm}f_1 - \frac{dh}{ds} \text{ is strictly decreasing on } (-\infty, 0),$$

$$f_2'(x) \le 0, \quad x \in \mathbb{R}, \qquad \frac{d}{ds}\frac{d}{dm}f_2 + \frac{dh}{ds} \text{ is strictly increasing on } (0, \infty).$$

We note that h can be then expressed as (3.21) via a function H satisfying the condition of Assumption 3.1, and further f_1, f_2 satisfy the condition of Assumption 3.1 for this function H. Therefore Theorem 3.2 applies to the present functions h, f_1, f_2.

For each $(S, X, A^{(1)}, A^{(2)}) \in \mathbb{A}$, the cost function k_x is defined, for $x \in S$, by

$$\begin{aligned} k_x(S, X, A^{(1)}, A^{(2)}) &= E_x\left(\int_0^\infty e^{-\alpha t} h(X_t)dt\right) \\ &\quad + E_x\left[\int_0^\infty e^{-\alpha t}\left(f_1(X_t)\dot{s}(X_t)dA_t^{(1),c} + f_2(X_t)\dot{s}(X_t)dA_t^{(2),c}\right)\right] \\ &\quad + E_x\left[\sum_{0\le t<\infty} e^{-\alpha t}\left(\int_{X_{t-}}^{X_{t-}+\Delta A_t^{(1)}} f_1(y)ds(y) + \int_{X_{t-}-\Delta A_t^{(2)}}^{X_{t-}} f_2(y)ds(y)\right)\right]. \end{aligned} \tag{4.10}$$

Some remarks about the cost structure are due at this point. The first integral $\int_0^\infty e^{-\alpha t}h(X_t)dt$ in (4.10) represents the so-called *holding cost* associated with the position of the controlled process X_t. Other integrals represent the *control cost,* which is associated with the "efforts" to change the position of the controlled process. The cost associated with each of the control functionals $A^{(i)}, i = 1, 2$, is proportional to the displacement caused by each of these functionals; however, the coefficient of the proportionality is a function of the position of the control process and is equal to $f_1(x)\dot{s}(x)$ if the controlled process is at the point x. Thus if $A^{(i)}$ is a continuous functional, we can write an approximation to the control cost as $\sum_j e^{-\alpha t_j} f_i(X_{t_j})\dot{s}(X_{t_j})\delta A_j^{(i)}$, where $\delta A_j^{(i)}$ is an increment of $A^{(i)}$ on the interval

$[t_j, t_{j+1}]$. In a limit one gets $\int_0^\infty e^{-\alpha t} f_i(X_t)\dot{s}(X_t)dA_t^{(i)}$. When the control functional has a discontinuity at the point t, which results in a jump of the control process, then we represent this jump as if the real clock is stopped while a new clock is turned on, and the controlled process is moving uniformly in the new time clock up or down from X_{t-} to $X_t = X_{t-} + \Delta A_t^{(i)}$. In such a representation the control cost of this displacement is equal to $\int_{X_{t-}}^{X_{t-}+\Delta A_t^{(i)}} f_i(y)\dot{s}(y)dy$, which corresponds to the last two terms in the right-hand side of (4.10). The same expression in the right-hand side of (4.10) would have been obtained as a limit if we had started with continuous functionals $A^{(i)}$ and then had approximated by them (via a monotone pointwise convergence) discontinuous control functionals.

Of course, when $f_i(x)\dot{s}(x)$ is equal to a constant r_i, the control cost associated with the functional $A^{(i)}$ can be written as $\int_0^\infty e^{-\alpha t} r_i dA_t^{(i)}$, without a need to have a special expression associated with the discontinuities of $A^{(i)}$. This was the case treated in [16]. We extend k_x outside the closed interval S denoted by $[\ell_1, \ell_2]$ as

$$\begin{aligned} k_x(S, X, A^{(1)}, A^{(2)}) &= k_{\ell_1}(S, X, A^{(1)}, A^{(2)}) + \int_x^{\ell_1} f_1(y)ds(y), \qquad x < \ell_1, \\ &= k_{\ell_2}(S, X, A^{(1)}, A^{(2)}) + \int_{\ell_2}^{x} f_2(y)ds(y), \qquad x > \ell_2. \end{aligned} \tag{4.11}$$

Our problem is to find the function

$$W^*(x) = \inf_{(S,X,A^{(1)},A^{(2)})\in\mathbb{A}} k_x(S, X, A^{(1)}, A^{(2)}), \qquad x \in \mathbb{R}, \tag{4.12}$$

called the *optimal return function*, and to find an optimal admissible quadruple $(S, X, A^{(1)}, A^{(2)}) \in \mathbb{A}$ such that

$$W^*(x) = k_x(S, X, A^{(1)}, A^{(2)}) \qquad \forall x \in \mathbb{R}.$$

The solution will be provided by the function W, the values a, b appearing in Theorem 3.2, and the reflecting (σ, μ)-diffusion on $[a, b]$ appearing in Proposition 4.2.

Here we introduce a subfamily $\mathbb{A}_0$ of $\mathbb{A}$ by

$$\mathbb{A}_0 = \{(S, X, A^{(1)}, A^{(2)}) \in \mathbb{A} : A_0^{(i)} = 0,\ P_x\text{-a.s. } \forall x \in S,\ i = 1, 2\}.$$

The reflecting (σ, μ)-diffusion on a compact interval appearing in Proposition 4.2 is a member of $\mathbb{A}_0$.

THEOREM 4.1. *Under Assumption* 4.1 *for functions* h, f_1, f_2, *let* W, a, b *be the function and values in Theorem* 3.2. *Then*

(i) $W(x) \le k_x(S, X, A^{(1)}, A^{(2)})$ *for all* $x \in \mathbb{R}$ *for any* $(S, X, A^{(1)}, A^{(2)}) \in \mathbb{A}$.

(ii) $W(x) = k_x(S, X, A^{(1)}, A^{(2)})$ *for all* $x \in \mathbb{R}$ *for* $(S, X, A^{(1)}, A^{(2)}) \in \mathbb{A}_0$ *if and only if*

$$S = [a, b], \quad (X, A^{(1)}, A^{(2)}) \text{ is the reflecting } (\sigma, \mu)\text{-diffusion on the interval } [a, b]. \tag{4.13}$$

Proof. (i) Take any $(S, X, A^{(1)}, A^{(2)}) \in \mathbb{A}$. Subtracting from (4.10) the identity (4.6) for $u = W$, we have

$$k_x(S, X, A^{(1)}, A^{(2)}) - W(x) = E_x(I_1 + I_2 + I_3 + I_4), \qquad x \in S, \tag{4.14}$$

where

$$I_1 = \int_0^\infty e^{-\alpha t} \left\{ \left(\frac{d}{dm} \frac{d}{ds} - \alpha \right) W(X_t) + h(X_t) \right\} dt,$$

$$I_2 = \int_0^\infty e^{-\alpha t} \left\{ \frac{dW}{ds}(X_t) + f_1(X_t) \right\} \dot{s}(X_t) dA_t^{(1),c},$$

$$I_3 = \int_0^\infty e^{-\alpha t} \left\{ -\frac{dW}{ds}(X_t) + f_2(X_t) \right\} \dot{s}(X_t) dA_t^{(2),c},$$

$$I_4 = \sum_{0 \le t < \infty} e^{-\alpha t} \left(\Delta W(X)_t + \int_{X_{t-}}^{X_{t-} + \Delta A_t^{(i)}} f_1(y) ds(y) + \int_{X_{t-} - \Delta A_t^{(2)}}^{X_{t-}} f_2(y) ds(y) \right).$$

The integrands I_1, I_2, I_3 are nonnegative by virtue of Theorem 3.2. To see that I_4 is nonnegative, let

$$\Gamma_+ = \{t \ge 0 : \Delta A_t^{(1)} > 0\}, \quad \Gamma_- = \{t \ge 0 : \Delta A_t^{(2)} > 0\}.$$

Since $\Gamma_+ \cap \Gamma_- = \phi$ by Remark 4.1, we have for $t \in \Gamma_+$,

$$\Delta X_t = \Delta A_t^{(1)}, \qquad \Delta W(X)_t = W(X_{t-} + \Delta A_t^{(1)}) - W(X_{t-}),$$

and consequently the sum in I_4, taken over all $t \in \Gamma_+$, equals

$$\int_{X_{t-}}^{X_{t-} + \Delta A_t^{(1)}} \left(\frac{dW}{ds}(y) + f_1(y) \right) ds(y).$$

We have a similar expression for $t \in \Gamma_-$ and eventually get

(4.15)
$$I_4 = \sum_{t \in \Gamma_+} e^{-\alpha t} \int_{X_{t-}}^{X_t} \left(\frac{dW}{ds}(y) + f_1(y) \right) ds(y) + \sum_{t \in \Gamma_-} e^{-\alpha t} \int_{X_t}^{X_{t-}} \left(-\frac{dW}{ds}(y) + f_2(y) \right) ds(y),$$

which is nonnegative by Theorem 3.2.

We have seen that $k_x \ge W(x)$, $x \in S$. This inequality extends to $\mathbb{R}$ by the definition (4.11) and Theorem 3.2.

(ii) Suppose $k_x(S, X, A^{(1)}, A^{(2)}) = W(x)$ for all $x \in S$ for some $(S, X, A^{(1)}, A^{(2)}) \in \mathbb{A}_0$. Then all P_x-expectations of I_1, I_2, I_3, I_4 must vanish for any $x \in S$. Notice further that $X_0 = x$, P_x-a.s. for all $x \in S$, because $A_0^{(i)} = 0$, P_x-a.s. for all $x \in S$, $i = 1, 2$. We let $S = [\beta, \gamma]$.

Suppose that $\beta < a$ (resp., $b < \gamma$). Then $E_x(I_1) > 0$ for $x \in (\beta, a)$ (resp., (b, γ)) by (3.23) and the right continuity of X_t. Therefore we have that $[\beta, \gamma] \subset [a, b]$.

In view of Lemma 4.1, both $A^{(1)}$, $A^{(2)}$ are nontrivial. If $a < \beta$ (resp., $\gamma < b$), then $\frac{dW}{ds} + f_1$ $\left(\text{resp., } -\frac{dW}{ds} + f_2\right)$ is strictly positive on S by (3.24), (3.25), and hence either I_2 or the first sum of (4.15) (resp., either I_3 or the second sum of (4.15)) has a positive P_x-expectation for any $x \in S$. We have proven that $S = [a, b]$.

Then, by virtue of (3.24), we see that X_t or, equivalently, $A_t^{(i)}$ $i = 1, 2$, must be continuous in $t \geq 0$, P_x-a.s. for any $x \in S$ in order to make the expectation of I_4 expressed as (4.15) to be zero. Finally, using (3.24) and (3.25), we see that $A^{(i)} = A^{(i),c}$, $i = 1, 2$, must satisfy the relations (4.9) for $\beta_1 = a$, $\beta_2 = b$ in order to make both expectations of I_2, I_3 to be zero. This means that $(X, A^{(1)}, A^{(2)})$ must be the reflecting (σ, μ)-diffusion on the interval $[a, b]$.

Conversely, the cost function k_x of the reflecting (σ, μ)-diffusion on the interval $[a, b]$ is obviously identical with $W(x)$ on $\mathbb{R}$ in view of (4.14). □

COROLLARY 4.1. *Under Assumption* 4.1 *for functions* h, f_1, f_2, *the solution* $W \in C^2(\mathbb{R})$, *and the values* a, b $(a < b)$ *of the free boundary problem* (3.23), (3.24), *and* (3.25) *are unique. The solution* $W(x)$, $x \in \mathbb{R}$, *coincides with the optimal return function* $W^*(x)$ *given by* (4.12).

Proof. In the proof of Theorem 4.1, we have seen that any function W satisfying (3.23), (3.24), and (3.25) for some a, b $(a < b)$ coincides with the function defined by (4.12). Further this function determines a, b uniquely according to (3.23). □

REFERENCES

[1] A. BENSUOUSSAN AND J.L. LIONS, *Applications of Variational Inequalities in Stochastic Control,* North-Holland, Amsterdam, New York, 1982.

[2] E.B. DYNKIN, *Game variant of a problem on optimal stopping,* Soviet Math. Dokl., 10 (1967), pp. 270–274.

[3] W.H. FLEMING AND H.M. SONER, *Controlled Markov Processes and Viscosity Solutions,* Springer-Verlag, New York, 1993.

[4] M. FUKUSHIMA, *On semi-martingale characterizations of functionals of symmetric Markov processes,* Electron. J. Probab., 4 (1999), pp. 1–32.

[5] M. FUKUSHIMA, Y. OSHIMA, AND M. TAKEDA, *Dirichlet Forms and Symmetric Markov Processes,* Walter de Gruyter, Berlin, New York, 1994.

[6] S.M. GUSEIN-ZADE, *On a game connected with the Wiener process,* Theory Probab. Appl., 14 (1969), pp. 701–704.

[7] M. HARRISON AND M. TAKSAR, *Instantaneous control of Brownian motion,* Math. Oper. Res., 8 (1983), pp. 439–453.

[8] N. IKEDA AND S. WATANABE, *Stochastic Differential Equations and Diffusion Processes,* 2nd ed., North-Holland, Amsterdam, Kodansha, Tokyo, 1989.

[9] K. ITO AND H.P. MCKEAN, *Diffusion Processes and Their Sample Paths,* Springer-Verlag, Berlin, Heidelberg, New York, 1974.

[10] I. KARATZAS AND H. WANG, *Connection between bounded-variation control and Dynkin games,* in Optimal Control and Partial Differential Equations, J.L. Menaldi, E. Rofman, and A. Sulem, eds., IOS Press, Amsterdam, 2001, pp. 363–373.

[11] T. KAWABATA, *On a Singular Control Problem for a Time Changed Distorted Brownian Motion,* doctoral thesis, Graduate School of Engineering Science, Osaka University, Japan, 1998.

[12] Y. KIFER, *Game options,* Finance Stoch., 4 (2000), pp. 443–463.

[13] P.A. MEYER, *Un cours sur les intégrales stochastiques,* in Séminaire de Probabilités X, Lecture Notes in Math. 511, Springer-Verlag, Berlin, Heidelberg, New York, 1976, pp. 245–400.

[14] H. NAGAI, *On an optimal stopping problem and a variational inequality,* J. Math. Soc. Japan, 30 (1978), pp. 303–312.

[15] G. STAMPACCHIA, *Forms bilinéaires coercitives sur les ensembles convexes,* C. R. Acad. Sci. Paris, 258 (1964), pp. 4413–4416.

[16] M. TAKSAR, *Average optimal singular control and a related stopping problem,* Math. Oper. Res., 10 (1985), pp. 63–81.

[17] H. TANAKA, *Stochastic differential equations with reflecting boundary condition in convex regions,* Hiroshima Math. J., 9 (1979), pp. 163–177.

[18] J. ZABCZYK, *Stopping games for symmetric Markov processes,* Probab. Math. Statist., 4 (1984), pp. 185–196.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 3, pp. 700–711

ROBUST FILTERING OF DISCRETE-TIME LINEAR SYSTEMS WITH PARAMETER DEPENDENT LYAPUNOV FUNCTIONS*

J. C. GEROMEL†, M. C. DE OLIVEIRA†, AND J. BERNUSSOU‡

Abstract. Robust filtering of linear time-invariant discrete-time uncertain systems is investigated through a new parameter dependent Lyapunov matrix procedure. Its main interest relies on the fact that the Lyapunov matrix used in stability checking does not appear in any multiplicative term with the uncertain matrices of the dynamic model. We show how to use such an approach to determine high performance H_2 robust filters by solving a linear problem constrained by linear matrix inequalities (LMIs). The results encompass the previous works in the quadratic Lyapunov setting. Numerical examples illustrate the theoretical results.

Key words. linear systems, discrete-time systems, parameter uncertainty, filtering, parameter dependent Lyapunov functions, linear matrix inequalities

AMS subject classifications. 93C05, 93C55, 93E11, 93E25

PII. S0363012999366308

1. Introduction.

Filtering is a very important issue in systems diagnosis, surveillance, and control. The problem, which amounts to extracting the information from the measured output to provide an estimate of the state (or a linear combination of the state), has been addressed in the stochastic as well as in the deterministic framework. Seminal works in this domain are the ones of Kalman and Luenberger (see [1, 10] for a complete discussion).

As a dual of the control design problem, the development of robust filters closely follows the same design steps. For linear uncertain systems, the problem can be stated as the minimization of an appropriate bound on a transfer function between an exogenous noise and the estimation error. There have been several contributions using H_2 and H_∞ norms as criteria for filter determination under parameter uncertainty. In the unstructured norm bounded uncertainty case, one can cite [9, 11, 13]. The structured case, which is a bit more complex, has also received some attention in both continuous-time [5, 8] and discrete-time [6, 7] contexts. These approaches are based on the quadratic stability concept, where a single Lyapunov matrix is used for the estimation error norm evaluation over the whole uncertainty domain. If we consider time-invariant uncertain systems, this assumption reveals to be, in fact, a hard constraint, implying a significant degree of sufficiency to these results. In fact, all results based on quadratic stability can also be applied to arbitrarily fast time-variant uncertainties.

In this paper, we use a new stability condition for discrete-time uncertain systems which enables the determination of parameter dependent Lyapunov matrices. This stability condition, which was first introduced in [3], provides results which go beyond the ones attainable by the quadratic approach for time-invariant parameter uncertainty. It is expressed as a linear matrix inequality (LMI) and exhibits a kind

*Received by the editors December 21, 1999; accepted for publication (in revised form) January 16, 2002; published electronically July 24, 2002. This work has been supported in part by grants from "Fundação de Amparo à Pesquisa do Estado de São Paulo–FAPESP" and "Conselho Nacional de Desenvolvimento Científico e Tecnológico–CNPq", Brazil.

http://www.siam.org/journals/sicon/41-3/36630.html

†DT, School of Electrical and Computer Engineering, UNICAMP, Av. Albert Einstein, 400, 13083-970 Campinas, SP, Brazil (geromel@dt.fee.unicamp.br, mcdeoliveira@ieee.org).

‡LAAS–CNRS, 7 Avenue du Colonel Roche, 31077, Cedex 4, Toulouse, France (bernusou@laas.fr).

of separation property between the Lyapunov matrices and the uncertain dynamic matrices. Here we show how the given condition can be extended to provide optimal performance in terms of an H_2 norm guaranteed cost problem. Furthermore, we show how to parametrize linear filters so that the synthesis of robust filters can be cast as an LMI optimization problem. We prove that our results encompass the results obtained on the quadratic stability framework [6, 7] and, consequently, reduce to the classical Kalman filter in the absence of uncertainty. The filtering results are generalized to cope with structure constraints such as decentralization.

The outline of the paper is as follows. In section 2, we formulate the problem to be solved. In section 3, we summarize the existent results on robust stability and robust filtering. Then we introduce the new stability and performance conditions in section 4 and illustrate its features with a numerical example. The robust H_2 filtering problem is developed in section 5 and extended to cope with structural constraints in section 6. Several numerical examples illustrate the results in section 7, enlightening the efficiency of the proposed approach by comparing the given results to the existent ones. The paper finishes with some concluding words.

The notation used throughout is as follows. Capital letters denote matrices, and small letters denote vectors. For scalars, we use small Greek letters. For matrices or vectors, $(')$ indicates transposition. For symmetric matrices, $X > 0$ (≥ 0) indicates that X is positive definite (nonnegative definite). For square matrices, trace(X) denotes the trace function of X being equal to the sum of its eigenvalues. For a transfer function $T(\zeta)$ analytic outside the unit circle, $\|T(\zeta)\|_2$ denotes the standard H_2 norm. Finally, for the sake of easing the notation of partitioned symmetric matrices, the symbol $(\bullet)'$ generically denotes each of its symmetric blocks.

2. Problem statement and definitions. Let us consider the linear time-invariant discrete-time system

$$x(k+1) = Ax(k) + Bw(k), \tag{1}$$

$$y(k) = Cx(k) + Dw(k), \tag{2}$$

$$z(k) = Lx(k), \tag{3}$$

where $x \in R^n$ is the state, $w \in R^m$ is a white noise input with zero mean and identity covariance matrix, $y \in R^r$ is the measured output, and $z \in R^s$ is the vector to be estimated. All matrices are of compatible dimension. We assume that matrix L is known and that the time-invariant parameters gathered in the matrix

$$M := \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{4}$$

are unknown but belong to the convex polyhedron

$$\mathcal{M} := \left\{ M(\xi) = \sum_{i=1}^{N} \xi_i M_i, \ \sum_{i=1}^{N} \xi_i = 1, \ \xi_i \geq 0 \right\}. \tag{5}$$

The robust H_2 filtering problem considered here is to design an estimate $\hat{z}$ of z given by $\hat{z} = \mathcal{F} \cdot y$. The filter $\mathcal{F}$ is supposed to be a linear, finite dimensional, and causal operator. We characterize $\mathcal{F}$ by the generic element given in the form of a linear time-invariant operator with minimum state space realization

$$\hat{x}(k+1) = A_f \hat{x}(k) + B_f y(k), \tag{6}$$

$$\hat{z}(k) = C_f \hat{x}(k), \tag{7}$$

where the matrices $A_f \in R^{n\times n}$, $B_f \in R^{n\times r}$, and $C_f \in R^{s\times n}$ are to be determined and define the filter transfer function

$$T_f(\zeta) = C_f\,(\zeta I - A_f)^{-1}\,B_f. \tag{8}$$

Moreover, it is considered that the initial condition of system (1)–(3) as well as the initial condition of the filter (6)–(7) are both zero.

The connection of the filter and the system yields, for each element in the set $\mathcal{M}$, a linear system described by the transfer function from the noise input w to the estimation error $e := z - \hat{z}$,

$$T_M(\zeta) := \tilde{C}\left(\zeta I - \tilde{A}\right)^{-1}\tilde{B}, \tag{9}$$

where matrices $\tilde{A}$, $\tilde{B}$, and $\tilde{C}$ of compatible dimensions are given by

$$\tilde{A} := \begin{bmatrix} A & 0 \\ B_f C & A_f \end{bmatrix},\ \tilde{B} := \begin{bmatrix} B \\ B_f D \end{bmatrix},\ \tilde{C} := \begin{bmatrix} L & -C_f \end{bmatrix}. \tag{10}$$

With respect to the transfer function $T_M(\zeta)$, it is possible to determine the quantity $\|T_M(\zeta)\|_2$, called the H_2 norm of $T_M(\zeta)$, that represents a measure of the energy appearing in the output due to the noisy input. Our aim is to solve the problem

$$\inf_{\mathcal{F}} \sup_{M\in\mathcal{M}} \|T_M(\zeta)\|_2^2. \tag{11}$$

Since this problem is very hard to solve, many authors proceed by minimizing an available upper bound to the indicated supremum. In [6, 7], this problem is solved using the concept of quadratic stability to be discussed in the next section.

3. Previous results on robust stability and filtering. As stated before, the robust filtering problem (11) is very difficult, and many authors address the filtering problem by replacing the supremum over $\mathcal{M}$ by an appropriate upper bound. One of the most used upper bounds is based on the concept of quadratic stability, which we briefly review in the next paragraphs.

Consider the following linear time-invariant discrete-time system defined by its transfer function

$$T(\zeta) := C(\zeta I - A)^{-1}B, \tag{12}$$

where the triplet (A, B, C) is composed by matrices of compatible and known dimensions. We are interested in the study of the stability of linear time-invariant systems in the form (12), where the matrices A and B are uncertain. More specifically, we are interested in systems whose uncertain parameters, gathered in the matrix

$$M := \begin{bmatrix} A & B \end{bmatrix}, \tag{13}$$

belong to the convex polyhedron $\mathcal{M}$ previously defined in (5). Our objective is to characterize whenever the set $\mathcal{M}$ defines only stable systems, in which case we say $\mathcal{M}$ is Schur, and to determine whether, for a given $\mu > 0$, it is true that the upper bound

$$\sup_{M\in\mathcal{M}} \|T_M(\zeta)\|_2^2 < \mu \tag{14}$$

holds. As in the previous section, the subscript included in the notation of the transfer function defined in (12) indicates the dependence of $T(\zeta)$ on $M \in \mathcal{M}$. The next lemma provides an answer to these questions expressed in terms of well-known sufficient conditions for robust (quadratic) stability and performance.

LEMMA 3.1. *The following statements are true:*

(a) *The set $\mathcal{M}$ is Schur if there exists a symmetric matrix P of compatible dimension satisfying the LMI*

$$\begin{bmatrix} P & A_i P \\ (\bullet)' & P \end{bmatrix} > 0 \tag{15}$$

for all $i = 1, \ldots, N$.

(b) *For any given $\mu > 0$, the inequality $\|T_M(\zeta)\|_2^2 < \mu$ holds for all $M \in \mathcal{M}$ if there exist symmetric matrices P and W of compatible dimensions satisfying the LMI*

$$\text{trace}(W) < \mu, \quad \begin{bmatrix} W & CP \\ (\bullet)' & P \end{bmatrix} > 0, \quad \begin{bmatrix} P & A_i P & B_i \\ (\bullet)' & P & 0 \\ (\bullet)' & (\bullet)' & I \end{bmatrix} > 0 \tag{16}$$

for all $i = 1, \ldots, N$.

Proof. See [7]. □

The above lemma deserves some comments. First, keeping P constant and independent of the index i is essential to obtain the results. This is on the origin of the *quadratic stability* [2] concept, largely used in robust stability studies of uncertain systems. The main drawback associated with this fact is that a single Lyapunov matrix P must work for all matrices in the uncertain domain $\mathcal{M}$, which ensures the stability of all time-variant systems in the domain. This condition is often too conservative if used with time-invariant systems. The same reasoning can be used for the robust performance provided in part (b). Indeed, it is expected that the use of a single matrix P introduces a significant degree of conservativeness to the estimation of the worst performance attained for some $M \in \mathcal{M}$. A measure of this gap may be obtained by calculating the minimum value of μ given by the optimal solution to the convex programming problem

$$\mu_q := \min\{\mu \ : \ \text{s.t. (16)}\} \tag{17}$$

as compared to $\sup_{M \in \mathcal{M}} \|T_M(\zeta)\|_2^2$. This fact will be illustrated in the examples.

Using Lemma 3.1, the following set of robust filters with guaranteed quadratic performance has been established in [6, 7] as the discrete-time counterpart of the continuous-time robust filter design introduced in [5].

LEMMA 3.2. *Let $\mu > 0$ be given. The estimation error transfer function satisfies the inequality $\|T_M(\zeta)\|_2^2 < \mu$ for all $M \in \mathcal{M}$ provided that the robust filter transfer function is given by*

$$T_f(\zeta) := HR^{-1}\left(\zeta I - QR^{-1}\right)^{-1} F, \tag{18}$$

where $R := Z - Y$ and matrices Q, H, and F and the symmetric matrices Y, Z, and

W satisfy the LMI

$$\text{trace}(W) < \mu, \tag{19}$$

$$\begin{bmatrix} W & L-H & L \\ (\bullet)' & Z & Z \\ (\bullet)' & (\bullet)' & Y \end{bmatrix} > 0, \tag{20}$$

$$\begin{bmatrix} Z & Z & ZA_i & ZA_i & ZB_i \\ (\bullet)' & Y & YA_i + FC_i + Q & YA_i + FC_i & YB_i + FD_i \\ (\bullet)' & (\bullet)' & Z & Z & 0 \\ (\bullet)' & (\bullet)' & (\bullet)' & Y & 0 \\ (\bullet)' & (\bullet)' & (\bullet)' & (\bullet)' & I \end{bmatrix} > 0 \tag{21}$$

for all $i = 1, \ldots, N$.

Proof. See [7]. □

From this theorem, it is clear that a near optimal solution to the design problem (11) is readily calculated from the optimal solution to the convex programming problem

$$\mu_Q := \min\{\mu \ : \ \text{s.t. } (19) - (21)\}, \tag{22}$$

which provides the best filter when a quadratic guaranteed upper bound to the worst error performance is adopted. It is worth mentioning that, for $N = 1$, the system under consideration is completely known, and, in this case, (22) generates the celebrated Kalman filter (see [7] for details).

4. Parameter dependent robust stability. This section presents the main results of this paper related to robust stability and performance of uncertain discrete-time systems. The following theorem constitutes an extension of the stability test recently introduced in [3] and will be used as a basis for the development of the new filter design procedure to be introduced in section 5.

THEOREM 4.1. *The following statements are true:*

(a) *The set $\mathcal{M}$ is Schur if there exist symmetric matrices P_i, $i = 1, \ldots, N$, and a matrix G of compatible dimensions satisfying the LMI*

$$\begin{bmatrix} P_i & A_iG \\ (\bullet)' & G + G' - P_i \end{bmatrix} > 0 \tag{23}$$

for all $i = 1, \ldots, N$.

(b) *For any given $\mu > 0$, the inequality $\|T_M(\zeta)\|_2^2 < \mu$ holds for all $M \in \mathcal{M}$ if there exist symmetric matrices P_i, W_i, $i = 1, \ldots, N$, and a matrix G of compatible dimensions satisfying the LMI*

$$\text{trace}(W_i) < \mu, \quad \begin{bmatrix} W_i & CG \\ (\bullet)' & G + G' - P_i \end{bmatrix} > 0, \quad \begin{bmatrix} P_i & A_iG & B_i \\ (\bullet)' & G + G' - P_i & 0 \\ (\bullet)' & (\bullet)' & I \end{bmatrix} > 0 \tag{24}$$

for all $i = 1, \ldots, N$.

Proof. We prove part (a) by assuming that (23) holds for all $i = 1, \ldots, N$ and calculating the convex combination of inequality (23). That is, we first multiply each

inequality in (23) by the uncertain parameter $\xi_i > 0$ and then evaluate the sum from $i = 1, \ldots, N$ so as to obtain

$$\begin{bmatrix} P(\xi) & A(\xi)G \\ (\bullet)' & G + G' - P(\xi) \end{bmatrix} > 0.$$

From this inequality, we first conclude that $G + G' > P(\xi) > 0$, where $P(\xi) := \sum_{i=1}^{N} \xi_i P_i$. Since $P(\xi) > 0$, the inequality $(P(\xi) - G)' P(\xi)^{-1} (P(\xi) - G) \geq 0$ is true for all values of the uncertain parameter ξ so that $G' P(\xi)^{-1} G \geq G + G' - P(\xi)$. Replacing this in the above inequality, we get

$$\begin{bmatrix} P(\xi) & A(\xi)G \\ (\bullet)' & G' P(\xi)^{-1} G \end{bmatrix} \geq \begin{bmatrix} P(\xi) & A(\xi)G \\ (\bullet)' & G + G' - P(\xi) \end{bmatrix} > 0.$$

Finally, if we multiply the first inequality above by $T(\xi) = \operatorname{diag}\left[I, G^{-1} P(\xi)\right]$ on the right and by $T(\xi)'$ on the left, we recover

$$\begin{bmatrix} P(\xi) & A(\xi)P(\xi) \\ (\bullet)' & P(\xi) \end{bmatrix} > 0,$$

which lets us conclude that the set $\mathcal{M}$ is Schur.

In order to prove part (b), we manipulate the third inequality in (24) following the same steps as in the proof of part (a) so as to obtain

$$\begin{bmatrix} P(\xi) & A(\xi)P(\xi) & B(\xi) \\ (\bullet)' & P(\xi) & 0 \\ (\bullet)' & (\bullet)' & I \end{bmatrix} > 0,$$

which lets us conclude that

$$\|T_M(\zeta)\|_2^2 \leq \operatorname{trace}\left(C P(\xi) C'\right) \quad \forall\ M \in \mathcal{M}. \tag{25}$$

Then, taking the convex combination of the first and the second inequalities in (24) with respect to the uncertain parameters, we get

$$\begin{aligned} \operatorname{trace}\left(C P(\xi) C'\right) &= \operatorname{trace}\left[CG\left(G' P(\xi)^{-1} G\right)^{-1} G' C'\right] \\ &\leq \operatorname{trace}\left[CG\left(G + G' - P(\xi)^{-1}\right) G' C'\right] \\ &\leq \operatorname{trace}\left(\sum_{i=1}^{N} \xi_i W_i\right) \\ &\leq \max_{i=1,\ldots,N} \operatorname{trace}\left(W_i\right) \\ &< \mu, \end{aligned}$$

which, together with (25), concludes the proof of part (b). □

Part (a) of the above theorem first appeared in [3]. Theorem 4.1 represents some important contributions. First, it contains the quadratic stability result as a particular case. Notice that, if we aggregate to the LMI (23) and (24) the additional linear constraints

$$G = G' = P,\ P_i = P,\ i = 1, \ldots, N, \tag{26}$$

then we exactly recover Lemma 3.1. Second, it generalizes the concept of quadratic stability and quadratic robust performance of uncertain systems to cope with parameter dependent Lyapunov functions. As indicated in the proof, the stability of the family of matrices $M(\xi) = \sum_{i=1}^{N} \xi_i M_i$ is tested by the parameter dependent Lyapunov function

$$v(x) = x' \left(\sum_{i=1}^{N} \xi_i P_i \right) x. \tag{27}$$

From part (b), it is possible to determine an upper bound to the H_2 norm of $T_M(\zeta)$ by solving the following convex programming problem:

$$\mu_p := \min \{\mu \ : \ \text{s.t. } (24)\}. \tag{28}$$

Notice that $\mu_p \leq \mu_q$ since the inequality (24) reduces to (16) with the additional constraints (26) and $W_i = W$, $i = 1, \ldots, N$. In practice, the value of μ_p is much less that μ_q since the number of free variables in the problem (28) is much bigger than the number of free variables in the quadratic robust performance design problem (17). This behavior will be illustrated by the numerical examples provided in section 7.

The fact that $\mu_p \leq \mu_q$ will guarantee that the robust filters to be designed in the next section always perform better (no worse) than the ones obtained by the filtering design procedures based on the concept of quadratic stability [7]. Unfortunately, it is hard to quantify how much improvement can be obtained. Nevertheless, the introduced analysis conditions can indeed coincide with the actual robust stability analysis in some examples (see [3] and [4]).

5. A new robust filtering procedure. At this point, after the analysis results presented in the last section, we turn to the following question: *Is it possible to provide a numerically attractive procedure to synthesize a filter that takes advantage of the new robust stability and performance conditions provided in Theorem* 4.1*?* This is the goal of this section. Applying the result given in Theorem 4.1 to the estimation error transfer function (9), we have that, for a given $\mu > 0$, the robust filter under consideration is such that $\|T_M(\zeta)\|_2^2 < \mu$, provided that the inequalities

$$\text{trace}(W_i) < \mu, \quad \begin{bmatrix} W_i & \tilde{C}_i \tilde{G} \\ (\bullet)' & \tilde{G} + \tilde{G}' - \tilde{P}_i \end{bmatrix} > 0, \quad \begin{bmatrix} \tilde{P}_i & \tilde{A}_i \tilde{G} & \tilde{B}_i \\ (\bullet)' & \tilde{G} + \tilde{G}' - \tilde{P}_i & 0 \\ (\bullet)' & (\bullet)' & I \end{bmatrix} > 0 \tag{29}$$

hold for all $i = 1, \ldots, N$. More precisely, we are interested in investigating whether it is possible to convert the nonlinear matrix inequality in terms of the filter parameters in (29) into an LMI. If this goal is accomplished, the robust H_2 filter design problem turns out to be a convex programming problem which can be solved by efficient numerical algorithms.

To this end, we proceed by partitioning $\tilde{G}$ and its inverse as

$$\tilde{G} := \begin{bmatrix} Z^{-1} & ? \\ U & ? \end{bmatrix}, \ \tilde{G}^{-1} := \begin{bmatrix} Y & ? \\ V & ? \end{bmatrix}, \tag{30}$$

where Z, U, Y, V, and "?" denote matrices in $R^{n \times n}$. Notice that, given the quadruple (Z, U, Y, V), we can always calculate blocks "?" in order to have $\tilde{G}\tilde{G}^{-1} = I$. Also

notice that no additional constraints like symmetry or definiteness are present. From this partition of matrix $\tilde{G}$, we introduce the one-to-one change of variables

$$\begin{bmatrix} A_f & B_f \\ C_f & 0 \end{bmatrix} := \begin{bmatrix} V' & 0 \\ 0 & I \end{bmatrix}^{-1} \begin{bmatrix} Q & F \\ H & 0 \end{bmatrix} \begin{bmatrix} UZ & 0 \\ 0 & I \end{bmatrix}^{-1}, \tag{31}$$

where the existence of the indicated inverses will be proven in what follows. Denoting $R := V'UZ$, the next theorem gives a solution, expressed in terms of an LMI, to the robust H_2 filtering problem previously stated.

THEOREM 5.1. *Let $\mu > 0$ be given. The estimation error transfer function satisfies the inequality $\|T_M(\zeta)\|_2^2 < \mu$ for all $M \in \mathcal{M}$, provided that the robust filter transfer function is given by*

$$T_f(\zeta) := HR^{-1}\left(\zeta I - QR^{-1}\right)^{-1} F, \tag{32}$$

where matrices Q, H, F, R, Z, Y and $W_i = W_i'$, $P_i = P_i'$, $S_i = S_i'$, J_i, $i = 1, \ldots, N$, satisfy the LMI

$$\operatorname{trace}(W_i) < \mu, \tag{33}$$

$$\begin{bmatrix} W_i & L - H & L \\ (\bullet)' & Z + Z' - P_i & Z' + Y + R' - J_i \\ (\bullet)' & (\bullet)' & Y + Y' - S_i \end{bmatrix} > 0, \tag{34}$$

$$\begin{bmatrix} P_i & J_i & Z'A_i & Z'A_i & Z'B_i \\ (\bullet)' & S_i & Y'A_i + FC_i + Q & Y'A_i + FC_i & Y'B_i + FD_i \\ (\bullet)' & (\bullet)' & Z + Z' - P_i & Z' + Y + R' - J_i & 0 \\ (\bullet)' & (\bullet)' & (\bullet)' & Y + Y' - S_i & 0 \\ (\bullet)' & (\bullet)' & (\bullet)' & (\bullet)' & I \end{bmatrix} > 0. \tag{35}$$

Furthermore, (29) *holds for some filter if and only if the inequalities* (33)–(35) *are feasible.*

Proof. Let us first suppose that (29) is feasible. We can partition matrix $\tilde{G}$ as indicated in (30) and assume that matrices U and V are nonsingular. We can do that because, given singular matrices U and V, we can always slightly perturb them, keeping feasibility due to the fact that all inequalities are strict. Furthermore, $\tilde{G} + \tilde{G}' > \tilde{P}_i > 0$ ensures that $\tilde{G}^{-1}$ and Z exist, and, consequently, relation (31) defines a one-to-one transformation. So, defining the square and nonsingular matrices

$$\tilde{T} := \begin{bmatrix} Z & Y \\ 0 & V \end{bmatrix}, \quad \tilde{T}'\tilde{P}_i\tilde{T} := \begin{bmatrix} P_i & J_i \\ (\bullet)' & S_i \end{bmatrix},$$

it can be verified that the second inequality in (29), multiplied on the left by the full rank matrix $T' := \operatorname{diag}[I, \tilde{T}']$ and to the right by T, provides the LMI (34). Furthermore, doing the same to the third inequality in (29) with matrix $T := \operatorname{diag}[\tilde{T}, \tilde{T}, I]$, we get the LMI (35), which, together with the first inequality in (29), implies that all inequalities (33)–(35) are feasible. In addition, since R is also a nonsingular matrix, we get

$$\begin{aligned} T_f(\zeta) &= C_f\left(\zeta I - A_f\right)^{-1} B_f \\ &= HZ^{-1}U^{-1}\left(\zeta I - V^{-T}QZ^{-1}U^{-1}\right)^{-1} V^{-T}F \\ &= HR^{-1}\left(\zeta I - QR^{-1}\right)^{-1} F. \end{aligned}$$

For the converse, let us suppose that the LMIs (33)–(35) are feasible. First, notice that

$$\left[\begin{array}{cc} Z+Z' & Z'+Y+R' \\ (\bullet)' & Y+Y' \end{array}\right] > \left[\begin{array}{cc} P_i & J_i \\ (\bullet)' & S_i \end{array}\right] > 0,$$

which, multiplied on the left by $T = \left[\begin{array}{cc} I & -I \end{array}\right]$ and on the right by T', implies that $R + R' > 0$ so that R is a nonsingular matrix. The same inequality implies that $Z + Z' > 0$ and, consequently, that Z is also nonsingular. From the definition $R = V'UZ$, the regularity of matrices U and V holds. Consequently, transformation (31) provides a filter satisfying (29). □

It is important to compare the result of Theorem 5.1 with that of Lemma 3.2. The optimal guaranteed H_2 cost robust filter, provided using a single Lyapunov function, i.e., the quadratic optimal filter, is recovered by imposing on the inequalities (33)–(35) the additional constraints

$$Z = Z',\ Y = Y',\ R = Z - Y, \tag{36}$$

$$W_i = W,\ \left[\begin{array}{cc} P_i & J_i \\ (\bullet)' & S_i \end{array}\right] = \left[\begin{array}{cc} Z & Z \\ (\bullet)' & Y \end{array}\right],\ i = 1, \ldots, N, \tag{37}$$

as a consequence, and, following the same steps as in [7], it is worth noticing that the previous result also contains as a particular case the celebrated Kalman filter when $N = 1$. It is important to remark (see the illustrative examples) that the main issue of this paper is to provide a way to relax the constraints (36)–(37). As illustrated in the previous section, this fact enables us to get smaller guaranteed costs when compared with all other available design procedures based on a single and hence parameter independent Lyapunov function.

Finally, a suboptimal robust filter is readily obtained from

$$\mu_P := \min \{\mu \ :\ \text{s.t.}\ \ (33) - (35)\}, \tag{38}$$

which is still an LMI optimal filtering problem. Notice that it is possible to show that $\mu_P \leq \mu_Q$ holds, where μ_Q is given by (22).

6. Decentralized filtering. As in [6, 7], another interesting point of the design procedure provided in this paper concerns the filter structure. In signal and systems estimation, when the overall system is described by a number of units coupled together by means of an interconnection network, it is of interest to know whether it is possible to connect local filters in order to estimate the local state variables [12]. The model is given by (1)–(3), where B, C, D, and L are block diagonal matrices. The goal is to determine a filter as (6)–(7) with a state space representation (see (32)), where

$$C_f = HR^{-1},\ A_f = QR^{-1},\ B_f = F \tag{39}$$

are block diagonal matrices of compatible dimensions. If possible, the filter can be split into a set of local filters acting on each subsystem level. Recalling that the inverse of a block diagonal matrix is also a block diagonal matrix and that the product of block diagonal matrices is a block diagonal matrix, (39) reveals that our goal is accomplished, provided that we include in the H_2 filtering design problem (34)–(35) the following additional constraints: *Matrices H, R, Q, and F are block diagonal.* Fortunately, this corresponds to constrain some entries of those matrices to be equal to zero, and so convexity is preserved. Also notice that, on the contrary to what

happens in [6, 7], the Lyapunov matrices P_i, S_i, and J_i must not present a block diagonal structure. Another surprising feature is that none of the submatrices of $\tilde{G}$ and its inverse given in (30) must be block diagonal—a direct consequence of the extra degrees of freedom introduced with the new stability condition.

7. Illustrative examples. In this section, we solve the proposed robust filter design problem for several systems in the form (1)–(3). This example is taken from [7], and the results are compared with the ones given by the design procedures in [11] and in [7]. We have a discrete-time system with matrices

$$B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},\ C = \begin{bmatrix} 1 & 0 \end{bmatrix},\ D = \begin{bmatrix} 0 & 0 & \sqrt{2} \end{bmatrix},\ L = \begin{bmatrix} 1 & 1 \end{bmatrix}$$

and a nominal matrix $A = A_0$ given by

$$A_0 = \begin{bmatrix} 0.9 & 0.1 \\ 0.01 & 0.9 \end{bmatrix}.$$

For this nominal system, the Kalman optimal filter $\mathcal{F}_{\mathrm{K}}$ is associated with the minimum H_2 cost equal to 8.0759 and is given by the minimal state space realization

$$A_{\mathrm{K}} = \begin{bmatrix} 0.4427 & 0.1000 \\ -0.1615 & 0.9000 \end{bmatrix},\ B_{\mathrm{K}} = \begin{bmatrix} 0.4573 \\ 0.1715 \end{bmatrix},\ C_{\mathrm{K}} = \begin{bmatrix} 1 & 1 \end{bmatrix}.$$

The first robust filter design we propose copes with the structured uncertainty defined through $A = A_0 + \Delta A$ for

$$\Delta A = \begin{bmatrix} 0 & 0.06\alpha \\ 0.05\beta & 0 \end{bmatrix} = \begin{bmatrix} 0.06 & 0 \\ 0 & 0.05 \end{bmatrix} \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

where $|\alpha| \leq 1$ and $|\beta| \leq 1$. This is a two-block structured uncertainty which can be exactly described by the set $\mathcal{M}$. Although the filter design procedure given in [11] cannot be directly applied to this problem without introducing some conservativeness, we take the best solution it provides without imposing the diagonal structure on the uncertainty parameters for the sake of comparison. This solution is obtained for a parameter $\varepsilon = 1.5264e-04$ and provides a suboptimal guaranteed cost H_2 filter $\mathcal{F}_{\mathrm{S}}$ with minimal state space realization

$$A_{\mathrm{S}} = \begin{bmatrix} 0.0335 & 0.1014 \\ -0.2551 & 0.9117 \end{bmatrix},\ B_{\mathrm{S}} = \begin{bmatrix} 0.8667 \\ 0.2652 \end{bmatrix},\ C_{\mathrm{S}} = \begin{bmatrix} 1 & 1 \end{bmatrix}.$$

Using the result of [7] (Lemma 3.2) with $N = 4$ matrices corresponding to the extreme points of the uncertain domain, we get the optimal quadratic guaranteed cost H_2 filter $\mathcal{F}_{\mathrm{Q}}$ given by

$$A_{\mathrm{Q}} = \begin{bmatrix} 0.0826 & -0.0768 \\ -0.0002 & 0.8543 \end{bmatrix},\ B_{\mathrm{Q}} = \begin{bmatrix} -0.0413 \\ 0.0001 \end{bmatrix},\ C_{\mathrm{Q}} = \begin{bmatrix} -29.8415 & -70.1868 \end{bmatrix}.$$

Finally, for the same set of vertices, Theorem 5.1 provides the optimal filter $\mathcal{F}_{\mathrm{P}}$:

$$A_{\mathrm{P}} = \begin{bmatrix} -0.1312 & 0.0842 \\ -0.0073 & 0.8352 \end{bmatrix},\ B_{\mathrm{P}} = \begin{bmatrix} -0.1151 \\ -0.0007 \end{bmatrix},\ C_{\mathrm{P}} = \begin{bmatrix} -14.7625 & -41.3592 \end{bmatrix}.$$

Table 1 shows, for each filter, the value of the H_2 guaranteed H_2 cost μ as well as the supremum of $\|T_M(\zeta)\|_2^2$ with respect to the matrix $M \in \mathcal{M}$, calculated by brute

TABLE 1
Filter performance: Multiblock uncertainty.

Filter	$\mathcal{F}_{\rm K}$	$\mathcal{F}_{\rm S}$	$\mathcal{F}_{\rm Q}$	$\mathcal{F}_{\rm P}$
μ	—	129.7915	100.0278	44.0039
$\sup_{M\in\mathcal{M}} \|T_M(\zeta)\|_2^2$	49.4994	38.2183	30.0664	15.4506

TABLE 2
Filter performance: One-block uncertainty.

Filter	$\mathcal{F}_{\rm K}$	$\mathcal{F}_{\rm O}$	$\mathcal{F}_{\rm Q}$	$\mathcal{F}_{\rm P}$
$N=2$	—	—	9.6796	8.8499
$N=4$	—	—	13.0219	11.5307
$N=8$	—	—	13.0446	11.6053
μ	—	13.0446	13.0446	11.6053
$\sup_{M\in\mathcal{M}} \|T_M(\zeta)\|_2^2$	13.0036	11.8655	11.8655	11.5980

force. As in [7], the Kalman filter, which is optimal for the nominal system, is the worst under parametric uncertainty. The filters of [11] and [7] are both suboptimal with respect to the guaranteed cost—the first one because of the structure of the uncertainty and the second one because of the quadratic stability assumption. It is interesting to observe that the filter determined from Theorem 5.1 is approximately 50% better than the best obtained by the existent procedures with respect to guaranteed H_2 cost as well as with respect to the true worst case value of the H_2 estimation cost.

As a second design, we consider the same uncertain system given before, but we change the uncertainty description to

$$\Delta A = \begin{bmatrix} 0 & 0.06\alpha \\ 0 & 0.05\beta \end{bmatrix} = \begin{bmatrix} 0.06 & 0 \\ 0 & 0.05 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \begin{bmatrix} 0 & 1 \end{bmatrix},$$

where the uncertain parameters are such that $\alpha^2 + \beta^2 \le 1$. With respect to this one-block unstructured uncertainty, the results of [11] provide the optimal quadratic guaranteed H_2 cost filter $\mathcal{F}_{\rm O}$:

$$A_{\rm O} = \begin{bmatrix} 0.3521 & 0.1069 \\ -0.2211 & 0.9400 \end{bmatrix},\ B_{\rm O} = \begin{bmatrix} 0.5479 \\ 0.2311 \end{bmatrix},\ C_{\rm O} = \begin{bmatrix} 1 & 1 \end{bmatrix}.$$

Although this uncertainty domain cannot be exactly represented by the polytopic domain $\mathcal{M}$, we proceed as in [7] by approximating the ellipsoidal uncertainty domain by the polyhedron associated with the extreme matrices

$$\begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} = \begin{bmatrix} \cos(2\pi i/N) \\ \sin(2\pi i/N) \end{bmatrix},\ i = 1, \dots, N.$$

Table 2 shows that, with $N = 8$, the quadratic filter $\mathcal{F}_{\rm Q}$ given by Lemma 3.2 is associated with the same guaranteed cost as $\mathcal{F}_{\rm O}$. Applying Theorem 5.1, it is possible to go even further. Notice that the optimal parameter dependent filter $\mathcal{F}_{\rm P}$,

$$A_{\rm P} = \begin{bmatrix} 0.4491 & 0.0758 \\ 0.0006 & 0.9008 \end{bmatrix},\ B_{\rm P} = \begin{bmatrix} -0.2360 \\ -0.0013 \end{bmatrix},\ C_{\rm P} = \begin{bmatrix} -3.2370 & -8.5027 \end{bmatrix},$$

is associated with a guaranteed cost which virtually matches the actual worst case performance.

8. Conclusion. The robust filtering problem for linear time-invariant discrete-time uncertain systems has been addressed in this paper using parameter dependent Lyapunov functions when convex polytopic uncertainty is present on the dynamic, input, and output matrices. The work is based on a new robust stability condition, which presents a separation between the Lyapunov matrix and the matrices of the dynamic model.

We have shown how to determine optimal H_2 guaranteed cost filters by solving a linear problem constrained by an LMI. The results encompass most of the results available in the literature to date which are based on the quadratic stability framework. We have also shown how to extend the results to cope with decentralized filtering without assuming a block diagonal Lyapunov matrix structure. Some numerical examples have been solved, illustrating the superiority of the results for the design of filters for time-invariant uncertain systems.

REFERENCES

[1] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Prentice Hall, Englewood Cliffs, NJ, 1979.

[2] B. R. Barmish, *Necessary and sufficient conditions for quadratic stabilizability of an uncertain system*, J. Optim. Theory Appl., 46 (1985), pp. 399–408.

[3] M. C. de Oliveira, J. Bernussou, and J. C. Geromel, *A new discrete-time robust stability condition*, Systems Control Lett., 37 (1999), pp. 261–265.

[4] M. C. de Oliveira, J. C. Geromel, and L. Hsu, *LMI characterization of structural and robust stability: The discrete-time case*, Linear Algebra Appl., 296 (1999), pp. 27–38.

[5] J. C. Geromel, *Optimal linear filtering under parameter uncertainty*, IEEE Trans. Signal Process., 47 (1999), pp. 168–175.

[6] J. C. Geromel, J. Bernussou, G. Garcia, and M. C. de Oliveira, *H_2 and H_∞ robust filtering for discrete-time linear systems*, in Proceedings of the 37th IEEE Conference on Decision and Control, Tampa, FL, 1998, pp. 632–637.

[7] J. C. Geromel, J. Bernussou, G. Garcia, and M. C. de Oliveira, *H_2 and H_∞ robust filtering for discrete-time linear systems*, SIAM J. Control Optim., 38 (2000), pp. 1353–1368.

[8] J. C. Geromel and M. C. de Oliveira, *H_2 and H_∞ robust filtering for convex bounded uncertain systems*, IEEE Trans. Automat. Control, 46 (2001), pp. 100–107.

[9] P. P. Khargonekar, M. A. Rotea, and E. Bayens, *Mixed H_2/H_∞ filtering*, Internat. J. Robust Nonlinear Control, 6 (1996), pp. 313–330.

[10] J. O'Reilly, *Observers for Linear Systems*, Academic Press, New York, 1983.

[11] I. R. Petersen and D. C. MacFarlane, *Optimal guaranteed cost control and filtering for uncertain linear systems*, IEEE Trans. Automat. Control, 39 (1994), pp. 1971–1977.

[12] D. D. Siljak, *Large Scale Dynamic Systems: Stability and Structure*, North–Holland, Amsterdam, 1979.

[13] L. Xie, C. E. de Souza, and M. Fu, *H_∞ estimation for discrete-time linear uncertain systems*, Internat. J. Robust Nonlinear Control, 1 (1991), pp. 11–23.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 3, pp. 712–722

GENERATION OF OPTIMAL PERIODIC OSCILLATIONS FOR THE CONTROL OF BOUNDARY LAYERS*

ABDERRAHMANE HABBAL†

Abstract. This paper is devoted to the optimal design problem of periodic surfaces. Solutions to elliptic partial differential equations occurring in oscillating domains exhibit boundary layer behavior, and we intend to control the first order correctors. Using a mathematical framework derived from the homogenization techniques, the existence of an optimal boundary layer control is proved.

Key words. optimal control, boundary layers, periodic structures

AMS subject classifications. 35B27, 35B37, 49J20, 49J45

PII. S0363012999354892

1. Introduction.

The importance of understanding the physics of nonhomogeneous media in view of strategic industrial applications is now well established, and the progress in both mathematical models and numerical simulation methods has led to their effective use in industry (see the pioneering works of [Tar86], [Koh86] among many others). These days, software which computes and optimizes microstructured materials is widely available for commercial use. In contrast, the problem of optimal design of coatings is less studied; yet there are many industrial fields where rough surfaces play a central role: acoustic shields, thermal radiators, shark-skin wrapping of planes, or, as reported by Friedman [Liu97], the epithaxial growth of VLSI chips, and the tiled surface covering the space shuttle, as reported by Achdou, Pironneau, and Valentin [AY98]. There is also an obvious advantage in dealing with rough surfaces when, for example, aesthetic or functionality criteria impose a prescribed macroscopic outline of the boundary, allowing only for microscopic engraving of the surface.

The present paper is concerned with the optimal design of the shape of the waving to obtain a prescribed gradient profile. The mathematical approaches developed here, namely, the *transport-homogenization* technique and *bounding* of the unit-cell, are motivated by two reasons. The first is that homogenization, just as in the "volume" case, avoids the meshing of the very small structure of the boundary, which would be very expensive from the computational viewpoint.

The second reason is that the transport technique (see, e.g., [MF76], [Sok]) avoids successive meshing of the unit-cell since the varying cell is the image of a fixed one, and the control variable is simply the underlying mapping, just as is widely used in the shell optimization theory; refer to [Che87], for example, for a comprehensive survey.

Let us notice that since the geometry of the oscillating boundary is intended to vary during the optimization process, *effective boundary conditions*, though usually considered, are not well suited for our purpose. Effective boundary conditions lead to a problem stated in the *whole* domain, with more accurate conditions which involve the shape of the oscillations. However, a change in the shape then implies that one must redo computations in the *whole* domain, which could be very expensive. From

*Received by the editors April 22, 1999; accepted for publication (in revised form) March 7, 2002; published electronically July 24, 2002.
http://www.siam.org/journals/sicon/41-3/35489.html
†Laboratoire Jean Alexandre Dieudonné, UMR CNRS 6621, Parc Valrose, Université de Nice-Sophia Antipolis, 06108 Nice, France (habbal@unice.fr).

the physical point of view, effective conditions are a macroscopic interpretation which could be achieved as soon as the boundary layer optimization process is finished.

The paper is organized as follows. In section 2 the potential flow in an oscillating domain is presented. Then the mathematical setting for an asymptotic analysis and convergence results are outlined. We set in section 3 the optimal design framework. Finally, in section 4 weak convergence of the states and the existence of an optimum are proved.

2. The potential fluid flow. The potential fluid flow model is a linear approximation for the general nonlinear fluid mechanics when the fluid is inviscid, incompressible, and assumed to be permanent and irrotational. Then one can define a potential U related to the flow velocity $\vec{\mathbf{V}}$ by $\vec{\mathbf{V}} = \vec{\nabla} U$. The potential function U solves the classical Laplace equation with suitable boundary conditions, generally slip conditions, and free-surface conditions.

Many industrial applications related to computational fluid dynamics efficiently use this approximate model, sometimes as a preamble to the simulation of the Navier–Stokes or Euler equations. Good examples of such industries are the shipbuilding, the offshore station designing, and, more generally, the marine technology industries.

Our intention is to use this model as an example to illustrate the general problem of finding *smart coatings* in order to control the boundary velocity profile. We exhibit boundary layer terms which depend on the shape of the waving and are correctors to the velocity profile. They are used in order to minimize the distance to a prescribed profile.

Let us consider a bounded area occupied by a potential fluid. A part of the boundary of this domain is waved with small periodic oscillations. Within the oscillating domain, an open bounded set with Lipschitz boundary, denoted by Ω_ϵ (see Figure 2.1), we consider the following Laplace problem:

$$\left\{\begin{array}{rcll} -\Delta U_\epsilon & = & 0 & \text{in } \Omega_\epsilon, \\ \partial_\nu U_\epsilon & = & 0 & \text{over } \Gamma_\epsilon, \\ \partial_\nu U_\epsilon & = & g & \text{over } \Gamma_N, \\ U_\epsilon & = & 0 & \text{over } \Gamma_D. \end{array}\right. \tag{2.1}$$

The function g describes the normal velocity of the incident flow through the part Γ_N of the boundary. It is assumed to be smooth enough, i.e., $g \in H^{\frac{1}{2}}(\Gamma_N)$ and fulfills the classical compatibility condition.

Setting $H(\Omega_\epsilon) = \left\{v \in H^1(\Omega_\epsilon),\ v = 0 \text{ over } \Gamma_D\right\}$, it is then well known that for any $\epsilon > 0$, there exists a unique $U_\epsilon \in H(\Omega_\epsilon)$ solution to the *variational problem* derived from (2.1):

$$\int_{\Omega_\epsilon} \nabla U_\epsilon . \nabla v \, d\Omega_\epsilon \ = \ \int_{\Gamma_N} g v \, d\Gamma \qquad \forall v \in H(\Omega_\epsilon). \tag{2.2}$$

2.1. The general setting for a two-scale boundary layer approach. In the following, $\Omega \subset \mathbb{R}^N$ is an open bounded set with a C^1-piecewise boundary. We focus our attention on a selected part on this boundary, denoted by Γ_0.

We assume without loss of generality that the boundary Γ_0 is plane, union of small cells homothetic (with a ratio ϵ) to a unit-cell period denoted Y'. Given a positive small enough real number ϵ and an arbitrary positive real $L > 1$, we define the strip B_ϵ as the one obtained by normal inward increasing of the boundary Γ_0: $B_\epsilon = \Gamma_0 \times]0, L\epsilon[$. The unit-cell is defined as $G = Y' \times]0, L[$.

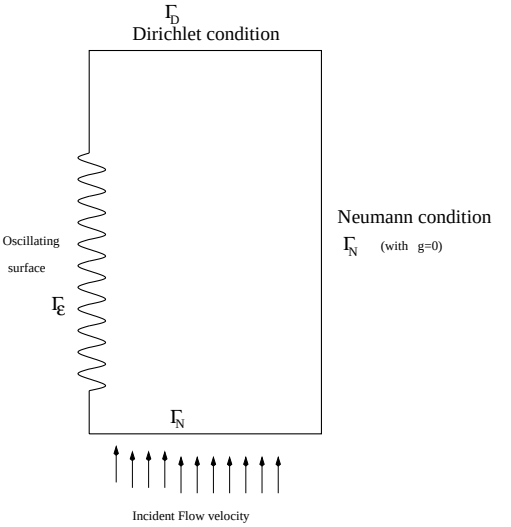


FIG. 2.1. *The oscillating domain and the physical configuration.*

Given a Y'-periodic function, $\vec{\psi} : Y' \to \mathbb{R}^N$, such that $\vec{\psi}_{|_{\partial Y'}} = 0$ (in order to avoid creating fissures) and a positive (small) real number ϵ, we define *an oscillating perturbation* $\vec{\psi}_\epsilon$ of the boundary Γ_0 by

$$\vec{\psi}_\epsilon : (x', 0) \in \Gamma_0 \to \vec{\psi}(x'/\epsilon) \in \mathbb{R}^N.$$

In order to define the oscillating domain Ω_ϵ, one usually applies a *harmonic extension* $\vec{V}_\epsilon$ of $\vec{\psi}_\epsilon$ on the reference domain Ω. The oscillating domain Ω_ϵ is then defined as the image of the reference domain Ω through the mapping $T_\epsilon = I + \epsilon\vec{V}_\epsilon$. It is clear from the definition of Ω_ϵ that the image of Γ_0 through T_ϵ is now an oscillating boundary, which we denote by Γ_ϵ. Moreover, in order to properly apply the two-scale boundary layer technique as presented in section 2.1, we choose perturbation fields $\vec{V}_\epsilon$ which vanish outside the strip $S_\gamma = \Gamma_0 \times\ ]0, \gamma\epsilon[$ with $\gamma \ll L$. Here the width L is intended to be large while γ is of order of the unit, e.g., $\gamma = 1$.

In the present paper, we explicitly consider as perturbations the mappings of the form

$$\vec{V}_\epsilon(x', x_n) = \vec{\psi}(x'/\epsilon)F(x_n/\epsilon), \tag{2.3}$$

where F is a smooth mollifier such as $F(t) = \exp(\frac{-t}{\gamma - t}),\ 0 \le t \le \gamma$.

We shall see in the third section of this paper (the setting of the optimal design framework) that the functions $\vec{\psi}$ should be Lipschitz, with derivatives not only bounded but also of *bounded variation*.

Now, we come back to the potential flow example. The perturbation field $\vec{V}_\epsilon$ waves the selected portion Γ_0, yielding Γ_ϵ, while it does not affect the remaining parts Γ_D and Γ_N of the boundary.

Our aim being to set the model problem in a domain which does not depend on the parameter ϵ, we define the reference Sobolev space by

$$H(\Omega) = \left\{ v \in H^1(\Omega), \quad v = 0 \quad \text{over } \Gamma_D \right\}.$$

Then, from the definition and regularity (algebraic and topological) properties of the mapping $T_\epsilon = I + \epsilon\vec{V}_\epsilon$, the natural norms over the spaces $H(\Omega)$ and $H(\Omega_\epsilon)$ are equivalent. Setting $u_\epsilon = U_\epsilon \circ T_\epsilon$, we get the following *transported* problem:

Find $u_\epsilon \in H(\Omega)$ such that $\forall v \in H(\Omega)$,

$$\int_\Omega (A_\epsilon \nabla u_\epsilon).\nabla v \, dx \;=\; \int_{\Gamma_N} g v \, d\Gamma, \tag{2.4}$$

where

$$A_\epsilon \;=\; (DT_\epsilon)^{-1}(DT_\epsilon)^{-T} |\det(DT_\epsilon)|. \tag{2.5}$$

It is an easy exercise to check that the operator associated to the diffusion matrix[1] A_ϵ is continuous, bounded, and $H(\Omega)$-elliptic (assuming that $\vec{\psi}$ belongs to the unit-ball of $W^{1,\infty}(Y')$). On the other hand, the use of the explicit fields given by (2.3) shows that A_ϵ depends only on the variable $y = x/\epsilon$. Then we define the y'-periodic operator $\overline{A}(y)$ as the restriction of $A_\epsilon(x/\epsilon)$ to the cell G, i.e.,

$$\overline{A}(y) = A_\epsilon(x/\epsilon). \tag{2.6}$$

2.2. The convergence results. Since the frequency and the amplitude of the oscillations are of the same order, only near-boundary effects occur (boundary layers). Far from the boundary, the solution behaves as if it doesn't see the oscillations.

Then the natural limit problem in our case is simply the one without oscillations at all ($\vec{\psi} = 0$), and it is expected that the difference between the nonoscillating solution and the oscillating one is a term which concentrates near the boundary. This fact is expressed by the strong convergence in the H^1-norm of the oscillating solution to the nonoscillating one (which implies that the difference terms are necessarily concentrating near the boundary).

Following [Con98] (in our case, the cell G is *bounded*), let us define the space

$$L^2(\Gamma_0; C_\#(\overline{G})) \;=\; \big\{ v(x', y);\ x' \in \Gamma_0,\ y = (y', y_n) \in G; \quad v(., y) \in L^2(\Gamma_0); \\ v(x', .) \in C(\overline{G})),\ \text{periodic w.r.t. } y' \big\}.$$

Definition 2.1. *Let $(u_\epsilon)_{\epsilon>0}$ be a sequence in $L^2(\Omega)$. It is said to two-scale converge in the sense of boundary layers on Γ_0 if there exists a function $u_0(x', y) \in L^2(\Gamma_0 \times G)$ such that*

$$\lim_{\epsilon\to 0} \frac{1}{\epsilon} \int_{B_\epsilon} u_\epsilon(x) v\left(x', \frac{x}{\epsilon}\right) dx \;=\; \frac{1}{|Y'|} \int_{\Gamma_0 \times G} u_0(x', y) v(x', y) dx' \, dy$$

for any $v \in L^2(\Gamma_0; C_\#(\overline{G}))$.

Let $C^\infty_\#(G)$ (respectively, $C^\infty_{0\#}(G)$) be the space of smooth functions in $\overline{G}$ which are Y'-periodic in y' and have a support in $y_N \in [0, 1]$ (respectively, in $y_N \in [0, 1[$). The space $H^1_\#(G)$ (respectively, $H^1_{0\#}(G)$) is the Sobolev space obtained by completion of $C^\infty_\#(G)$ (respectively, $C^\infty_{0\#}(G)$) with respect to the $H^1(G)$-norm.

Let then consider the nonoscillating solution u, which solves the following problem:

[1] We shall henceforth identify the second order linear elliptic operators and their associated diffusion matrices.

Find $u \in H(\Omega)$ such that $\forall v \in H(\Omega)$,

$$\int_\Omega \nabla u.\nabla v \, dx \;=\; \int_{\Gamma_N} gv \, d\Gamma. \tag{2.7}$$

The asymptotic analysis results presented below are quite classical (generally in the case of semi-infinite unit-strip G); see especially the works of [Lio81], [Lan77], [Cas96], [ole92], [Ami96], and [All99].

The limit equation for the two-scale boundary layer u_1 is obtained by application of the two-scale convergence to (2.4) using test functions *concentrating* on the strip B_ϵ (hence, vanishing over the Neumann boundary Γ_N). It is stated as follows: Find $u_1(x', y) \in L^2(\Gamma_0;\ H^1_\#(G)/\mathbb{R})$ such that for any $v \in L^2(\Gamma_0;\ H^1_{0\#}(G))$,

$$\frac{1}{|Y'|}\int_{\Gamma_0\times G} \overline{A}\nabla_y u_1 \nabla_y v \, dx'dy = -\frac{1}{|Y'|}\int_{\Gamma_0\times G} \overline{A}\nabla u_{|\Gamma_0} \nabla_y v \, dx'dy. \tag{2.8}$$

Note that the assumptions on Ω and on the flow inlet g are enough to allow us to consider the trace $\nabla u_{|\Gamma_0}$ in $L^2(\Gamma_0)$.

The equation above is ill-posed due to the lack in boundary conditions. Indeed, $u_1(x', y)$ is the sum of a unique first order corrector $c_1 \in H^1_{0\#}(G)$ and a first order "tail" term $t_1(x', y)$ whose gradient dies exponentially with respect to L. Let us pose $u_1(x', y) = c_1(x', y) + t_1(x', y)$ with c_1 and t_1 solutions to the following equations:

- Find $c_1(x', y) \in L^2(\Gamma_0;\ H^1_{0\#}(G))$ such that for any $v \in L^2(\Gamma_0;\ H^1_{0\#}(G))$,

$$\frac{1}{|Y'|}\int_{\Gamma_0\times G} \overline{A}\nabla_y c_1 \nabla_y v \, dx'dy = -\frac{1}{|Y'|}\int_{\Gamma_0\times G} \overline{A}\nabla u_{|\Gamma_0} \nabla_y v \, dx'dy. \tag{2.9}$$

- Find $t_1(x', y) \in L^2(\Gamma_0;\ H^1_\#(G)/\mathbb{R})$ such that for any $v \in L^2(\Gamma_0;\ H^1_{0\#}(G))$,

$$\frac{1}{|Y'|}\int_{\Gamma_0\times G} \overline{A}\nabla_y t_1 \nabla_y v \, dx'dy = 0. \tag{2.10}$$

THEOREM 2.2. *The two following statements hold:*

(i) *There exist two positive constants β and C_1 independent from the width L such that*

$$\|t_1\|_{L^2(\Gamma_0;\ H^1_\#(G)/\mathbb{R})} \le C_1 \exp(-\beta L). \tag{2.11}$$

(ii) *There exists a positive constant C_2 independent from L such that*

$$\lim_{\epsilon\to 0} \frac{1}{\sqrt{\epsilon}} \|u_\epsilon(x) - u(x) - \epsilon c_1(x', x/\epsilon)\|_{H^1(\Omega)} \le C_2 \exp(-\beta L/2). \tag{2.12}$$

Moreover, the constant $\beta > 0$ depends only on the operator $\overline{A}$ and on the boundary Γ_0.

Let us remark again that from the corrector equation (2.9) and Theorem 2.2, the trace of the gradient of u (i.e., the profile of u) over the boundary Γ_0 is a source for the generation of the boundary layer term c_1 when the boundary is waved. The gradient near the oscillating boundary is then approximated (in the two-scale boundary layer sense) by the initial profile of the nonoscillating u plus the correcting gradient of c_1. It is then very tempting to see c_1 and, more precisely, the periodic oscillation shape as *boundary layer controls* and pose the question, How can we design the waving in order to get a prescribed behavior of the solutions near the oscillating boundary? The remainder of this paper intends to answer this question.

3. The optimal design framework. To optimize the shape of the oscillations, there are mainly two methods. The first one consists of deriving an effective boundary condition (wall law), which depends on the roughness and on the cell functions but not explicitly on the boundary layer corrector. One obtains a new unknown defined in the whole domain Ω. If one seeks only the control of near-boundary effects with varying roughness, this approach is too expensive.

The second method is as follows: From the corrector convergence theorem of the previous section, it can be easily seen that if one computes the solution u of the nonoscillating domain and then fits the local behavior by controlling the corrector c_1, then the computational cost is slightly reduced. Effective law approach requires us to redo computations in the whole domain when the shape of oscillations varies, as well as the updating of the cell functions. Only the latter is needed in the case of control with c_1.

On the other hand, a drawback of rectification techniques such as the present homotopy one is that one obtains equations with the geometry described by a distributed parameter, which needs to be smoother than naturally expected. The need for midsurfaces of bounded third derivatives in the theory of shells is a good example. In the present case, it comes out from the proof of existence of minima that the shape of the oscillations should have a derivative with *bounded variations*, i.e., belong to the space $BV(Y';\mathbb{R}^N)$; see [Giu84] for a general presentation.

A simple way to avoid the use of such a space is to consider oscillations of bounded second derivatives, although then shapes with corners, such as saw-teeth ones, are forbidden. By working in the space with derivatives in BV, the latter shape is allowed; but again lower-scale oscillations (microstructures *in* the microstructure) are not allowed (since one obtains shapes with second derivatives which are infinite sums of Dirac measures, and this is not a Radon measure).

As usual in optimal design theory, the model which we intend to control is restated, carefully underlying the dependence on the control variable.

Henceforth, the following notation will be used:

- The Banach space of controls is

$$W = \left\{\phi \in W^{1,\infty}(Y';\ \mathbb{R}^N),\ \left(D_j\phi_i\right) \in BV(Y'),\ \phi|_{\partial Y'} = 0\right\},$$

  where

$$\left(D_j\phi_i\right) \ = \ \left(D_{y'}\phi\right)_{ij}, \quad 1 \le i \le N,\ 1 \le j \le N-1,$$

  is the Jacobian matrix of ϕ.
- The set of admissible shapes (vector-valued, with arrow notation omitted) is

$$\Phi_l = \{\phi \in W, \quad \|\phi\|_W \le l\},$$

  where the role of the upper bound $l > 0$ is twofold: from the mathematical viewpoint, it states that the norm of ϕ must be small enough to ensure that the waved boundary is still Lipschitz; from the practical viewpoint, it states that the amount of material used in the waving (or engraving) is limited.
- The Hilbert space of (corrector) state variables is

$$V = L^2(\Gamma_0, H^1_{0\#}(G)).$$

In the definition (2.6) of the y'-periodic diffusion operator, we underline the dependence on the control variable ψ and denote the operator by $\overline{A}(\psi)(y)$.

Then, the energy and source functionals are defined by

$$e : \Phi_l \times V \times V \to \mathbb{R}, \quad e(\psi;\ c,\ v) \ = \ \int_{\Gamma_0 \times G} \overline{A}(\psi) \nabla_y c \nabla_y v \, dx' dy, \tag{3.1}$$

$$S : \quad \Phi_l \times V \to \mathbb{R}, \qquad S(\psi;\ v) \ = \ -\int_{\Gamma_0 \times G} \overline{A}(\psi) \nabla u_{|\Gamma_0} \nabla_y v \, dx' dy. \tag{3.2}$$

Then the following is the state equation: for a given $\psi \in \Phi_l$, find $c(\psi) \in V$ such that

$$\forall v \in V \quad e(\psi;\ c(\psi),\ v) \ = S(\psi;\ v). \tag{3.3}$$

Then, in order to state the optimal design problem, the *observation* functional is defined by

$$J : \Phi_l \times V \to \mathbb{R}, \tag{3.4}$$

$$(\psi;\ v) \to J(\psi;\ v) = \int_{\Gamma_0 \times G} g(x', y, \psi, \xi(\psi;\ v)) \, dx' dy, \tag{3.5}$$

where the auxiliary variable ξ is defined by

$$\xi(\psi;\ v)(x', y) \ = (\xi_1,\ \xi_2) = (D_{y'}\psi(y'), \nabla_y v(x', y)), \tag{3.6}$$

where $D_{y'}\psi$ is the Jacobian matrix of the vector ψ, and the function g defined over $\Gamma_0 \times G \times \mathbb{R}^N \times \mathbb{R}^{N \times N}$ is a Carathéodory integrand (i.e., measurable with respect to (x', y) and continuous with respect to (ψ, ξ)).

We point out that in the present study, the integrand of interest to us is the one considered for the control of the velocity profile over Γ_0, i.e.,

$$g(x', y, \psi, \xi) \ = \ |\xi_2 \ - \ \tau|^2,$$

where $\tau \in L^2(\Gamma_0)$ is a fixed desirable profile. It is quite well known that the observation associated to this integrand is lower semicontinuous. The above presentation as well as the standard growth condition below (obviously fulfilled by g) are presented for the sake of generality.

We assume that the integrand g is *convex* with respect to ξ and satisfies the *standard growth condition*: $\forall r \in \mathbb{R},\ \exists a_r \in (L^1(\Gamma_0 \times G))^{N(N-1)} \times (L^2(\Gamma_0 \times G))^N$, and $b_r \in L^1(\Gamma_0 \times G)$ such that

$$\inf_{|\psi| \le r} g(x', y, \psi, \xi) \ \ge \ \langle a_r(x', y),\ \xi \rangle_{\mathbb{R}^{N(N-1)} \times \mathbb{R}^N} + b_r(x', y). \tag{3.7}$$

The assumptions above are made to ensure that the observation J is lower semicontinuous, a necessary preamble to the proof of existence of minima. These assumptions are quite general and are fulfilled by a very large class of integrands g classically used in the shape optimization problems.

The *cost function* is then defined over the set Φ_l by

$$j(\psi) \ = \ J(\psi;\ c(\psi)),$$

where $c(\psi) \in V$ is the solution to the state equation (3.3).

Finally, the *optimal design problem* is stated as follows:

$$\text{Find } \psi \in \Phi_l \text{ such that } j(\psi) \ = \inf_{\phi \in \Phi_l} j(\phi). \tag{3.8}$$

Now, after a complete setting of the optimization problem, we face in the next section the question of *existence* of a solution to the problem (3.8).

4. Existence of an optimum. Since the set of admissible shapes Φ_l is compact for the weak-$\star$ topology of $W^{1,\infty}(Y',\ \mathbb{R}^N)$, it is well known that a sufficient condition for the existence of a minimum is that the cost function j should satisfy the lower semicontinuity requirement with respect to the same topology, the latter being generally proved in two steps:

1. First, prove the lower semicontinuity of the observation $J(.,.)$ with respect to its two arguments.
2. Then prove the weak convergence of the state variable, i.e., $c(\psi_n) \rightharpoonup c(\psi)$ in V, whenever $\psi_n \overset{\star}{\rightharpoonup} \psi$ in Φ_l.

Thus, one has

$$j(\psi) = J(\psi;\ c(\psi)) \ \leq\ \liminf_{n\to+\infty} J(\psi_n;\ c(\psi_n)) \ =\ \liminf_{n\to+\infty} j(\psi_n). \tag{4.1}$$

4.1. Weak lower semicontinuity of the observation J. Mainly due to the assumptions made on the integrand g, the lower semicontinuity of J is proved by means of the following theorem; see, e.g., [Dac89, Theorem 3.4].

THEOREM 4.1. *Let Ω be an open bounded subset of $\mathbb{R}^{(N-1)} \times \mathbb{R}^N$, and let*

$$g\ :\ \Omega \times \mathbb{R}^N \times \mathbb{R}^{N\times N}\ \mapsto\ \mathbb{R} \cup \{+\infty\}$$

be a Carathéodory integrand satisfying the growth condition

$$g(x,s,\xi)\ \geq\ \langle a(x),\ \xi\rangle_{\mathbb{R}^{N(N-1)}\times\mathbb{R}^N} + b(x) \tag{4.2}$$

for almost every $x \in \Omega$ and $\forall (s,\xi) \in \mathbb{R}^N \times \mathbb{R}^{N\times N}$, with $a \in (L^{q'}(\Omega))^{N\times N}$ (q' is the conjugate of q) and $b \in L^1(\Omega)$.

Assume that $g(x,s,.)$ is convex and that

$$\begin{cases} s_k \to s & \text{in } (L^p(\Omega))^N, \\ \xi_k \rightharpoonup \xi & \text{in } (L^q(\Omega))^{N\times N}. \end{cases} \tag{4.3}$$

Then the functional

$$\mathcal{J}(s,\xi) = \int_\Omega g(x, s(x), \xi(x))dx$$

is lower semicontinuous, i.e.,

$$\mathcal{J}(s,\xi)\ \leq\ \liminf_{k\to\infty} \mathcal{J}(s_k,\xi_k).$$

The theorem above is immediately applicable to our problem if we set

$$\begin{cases} \Omega\ =\ \Gamma_0 \times G, \quad x = (x',y); \\ s(x)\ =\ \psi(y') \in \Phi_l, \quad \xi(\psi,v) = (D_{y'}\psi, \nabla_y v) \quad \text{(refer to (3.4)–(3.6))}; \\ J(\psi,\ v)\ =\ \mathcal{J}(s, \xi(\psi,v)); \end{cases} \tag{4.4}$$

and the convergences are to be understood *in the weak-$\star$ sense* whenever the L^∞-topology is involved. Indeed, if we consider a sequence $(\psi_k,\ v_k)$ such that

$$\begin{cases} \psi_k \overset{\star}{\rightharpoonup} \psi & \text{in } W, \\ v_k \rightharpoonup v & \text{in } V, \end{cases} \tag{4.5}$$

then the weak convergence $\xi_k \rightharpoonup \xi$ must be understood as a product of weak-$\star$ times weak convergences; more precisely,

$$\left\{ \begin{array}{ll} (\xi_k)_1 \overset{\star}{\rightharpoonup} (\xi)_1 & \text{in } L^\infty(Y'), \\ (\xi_k)_2 \rightharpoonup (\xi)_2 & \text{in } L^2(\Gamma_0; L^2(G)). \end{array} \right. \tag{4.6}$$

On the other hand, the embedding of W onto $C(\overline{Y'})$ is compact, so that

$$\psi_k \to \psi \text{ in } C(\overline{Y'}).$$

Hence, we have the strong convergence of the sequence $(s_k) = (\psi_k)$. Moreover, there exists a constant C such that for any $\psi \in \Phi_l$,

$$\sup_{y' \in \overline{Y'}} |\psi(y')| \leq C;$$

then, thanks to the standard growth condition (3.7), one can take $a = a_C$ and $b = b_C$ in order to fulfill the condition (4.2). Hence, one can apply Theorem 4.1, which yields

$$J(\psi,\ v) \ = \ \mathcal{J}(s, \xi(\psi, v)) \leq \ \liminf_{k\to\infty} \mathcal{J}(s_k, \xi_k) = \ \liminf_{k\to\infty} J(\psi_k, v_k),$$

i.e., the functional J is weakly lower semicontinuous over $\Phi_l \times V$.

4.2. Weak continuity of the state variable. Our aim now is to show that $c(\psi_n) \rightharpoonup c(\psi)$ in V for all sequences (ψ_n) such that $\psi_n \overset{\star}{\rightharpoonup} \psi$ in Φ_l.

The weak continuity holds with trivial proof if we restrict the space of controls (and the admissible set) to $W^{2,\infty}(Y';\ \mathbb{R}^N)$. In this case, for any kind of oscillations, we have a *strong* convergence of $\overline{A}(\psi_n)$ to $\overline{A}(\psi)$ in $L^\infty(Y')$, yielding immediately that $c(\psi_n) \rightharpoonup c(\psi)$ in V. In the general case, the weak convergence of states is proved by means of the G-convergence theory. Briefly, a sequence A_n of elliptic operators is said to G-converge to an elliptic operator A if, for any given right-hand side f, one has $A_n^- 1 f \rightharpoonup A^- 1 f$; see, e.g., [G.93].

Recall that the diffusion matrix $\overline{A}(\psi)$ is given by

$$\overline{A}(\psi) \ = \ (I + D_y V)^{-1} (I + D_y V)^{-T} |\det(I + D_y V)|,$$

where the perturbation field V that is considered is the following:

$$\vec{V}(y', y_N) = \psi(y') F(y_N) \tag{4.7}$$

with $F(t) = \exp(\frac{-t}{\gamma - t}),\ 0 \leq t \leq \gamma$.

On the other hand, the matrix norm of the Jacobian $D_y V$ is upperbounded by the norm of ψ in $W^{1,\infty}(Y')$. It is then a simple linear algebra exercise to show that the set

$$\left\{ \overline{A}(\psi), \quad \psi \in \Phi_l \quad \text{with} \ \ l < 1 \right\}$$

is uniformly bounded and equicoercive (in the usual sense of bounded elliptic operators) with respect to the design variable ψ.

All the ingredients are then ready for the application of the G-convergence of elliptic operators. It is straightforward to show that the sequence $(\overline{A}(\psi_n))$ G-converges to a limit $\overline{A}_G$, and that $c(\psi_n) \rightharpoonup c_{_G}$ in V. Then if one proves that $\overline{A}_G$ is equal to $(\overline{A}(\psi))$, this will immediately yield that $c_{_G} = c(\psi)$.

The equality $\overline{A}_G = (\overline{A}(\psi))$ is in fact a byproduct of the strong convergence (see [KO94, p. 150]) of the sequence $(\overline{A}(\psi_n))$ to $(\overline{A}(\psi))$ in $L^2(Y')$, as will be proved below.

Let us first give an explicit formula for the matrix $(\overline{A}(\psi))$, for example, in the two-dimensional case (here $\psi = (\psi_1, \psi_2)$):

(4.8)
$$\overline{A}(\psi) \;=\; \frac{1}{(1+\psi_1' F)(1+\psi_2' F)-\psi_1 F' \psi_2' F} \times \begin{pmatrix} (1+\psi_2 F')^2 + (\psi_1 F')^2 & -\psi_2' F(1+\psi_2 F') - \psi_1 F'(1+\psi_1' F) \\ -\psi_2' F(1+\psi_2 F') - \psi_1 F'(1+\psi_1' F) & (1+\psi_1' F)^2 + (\psi_2' F)^2 \end{pmatrix}.$$

The notation ψ_1' stands for the derivative with respect to $y' \in [0,1]$.

Let us then consider a minimizing sequence (ψ_n) such that $\psi_n \overset{\star}{\rightharpoonup} \psi$ in Φ_l. For the sake of clarity, we shall use scalar notation, although the involved functions and sequences are vector-valued. (The prime signed $(\psi)'$ stands of course for the Jacobian matrix of ψ.)

Due to the compact embedding of the space $W^{1,\infty}(Y';\ \mathbb{R}^N)$ onto the space of continuous functions $C(\overline{Y'})$, one has

$$\psi_n \to \psi \text{ in } C(\overline{Y'}). \tag{4.9}$$

Then, this time using the compactness of the embedding of $BV(Y')$ onto the space $L^1(Y')$, we get

$$(\psi_n)' \to (\psi)' \text{ in } L^1(Y').$$

As the derivatives of the ψ_n's belong also to $L^\infty(Y')$, it is straightforward to show that the strong convergence indeed holds in $L^p(Y')$ for any $1 \le p < +\infty$. A careful examination of the structure of $\overline{A}(\psi)$ easily shows that the whole $(\overline{A}(\psi_n))$ converges to $\overline{A}(\psi)$ in $L^2(Y')$. The conclusion holds of course in any dimension.

THEOREM 4.2. *Under the general setting and assumptions of section* 3, *the optimal design problem*

$$\textit{Find } \psi \in \Phi_l \textit{ such that } j(\psi) \;=\; \inf_{\phi \in \Phi_l} j(\phi)$$

has a solution $\phi \in \Phi_l$.

In order to implement a numerical simulation, the use of descent algorithms in order to achieve numerical computation of the optimum generally requires the user to supply a gradient subroutine. Classically, the fast computation of the gradient is done by means of the adjoint state method, provided that the quantities which depend on the control variable are differentiable. Since we are in a classical linear elliptic framework, it is well known that the state control, the energy, and the cost function are differentiable as soon as the user prescribed observation is. For reference, see [O.84], [Ven78], and [Has96]. Let us remark that the local nature of the considered controls allow for parallel processing of several oscillating parts of the domain, provided that the boundary layers are not sufficiently close to each other to interact.

As a short conclusion, it must be kept in mind that the techniques of boundary layer control introduced in this paper apply only to potential flows, which do not develop genuine boundary layers as they are inviscid by definition. While the techniques applied here could be adapted to the Stokes model with rather minimal effort, this

is not the case when studying the more realistic Navier–Stokes flows, which develop physical boundary layers directly related to the Reynolds number. The interaction between the Reynolds boundary layers and those due to the waving of the boundary is a complex physical phenomenon, and we refer to [AY98] for a good investigation of the subject.

REFERENCES

[AY98] Y. Achdou, O. Pironneau, and F. Valentin, *Effective boundary conditions for laminar flows overs periodic rough boundaries*, J. Comput. Phys., 147 (1998), pp. 187–218.

[All99] G. Allaire and M. Amar, *Boundary layer tails in periodic homogenization*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 209–243.

[Con98] G. Allaire and C. Conca, *Boundary layers in the homogenization of a spectral problem in fluid-solid structures*, SIAM J. Math. Anal., 29 (1998), pp. 343–379.

[Ami96] Y. Amirat and J. Simon, *Influence de la rugosite en hydrodynamique laminaire (Influence of rugosity in laminar hydrodynamics)*, C. R. Acad. Sci. Paris Sér. I Math., 323 (1996), pp. 313–318.

[Cas96] J. Casado-Diaz and I. Gayte, *A general compactness result and its application to the two-scale convergence of almost periodic functions*, C. R. Acad. Sci. Paris Sér. I Math., 323 (1996), pp. 329–334.

[Che87] D. Chenais, *Optimal design of midsurface of shells: Differentiability proof and sensitivity computation*, Appl. Math. Optim., 16 (1987), pp. 93–133.

[Dac89] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Springer-Verlag, Berlin, 1989.

[G.93] G. Del Maso, *An Introduction to* Γ*-Convergence*, Birkhäuser, Basel, 1993.

[Liu97] A. Friedman, B. Hu, and Y. Liu, *A boundary value problem for the Poisson equation with multi-scale oscillating boundary*, J. Differential Equations, 137 (1997), pp. 54–93.

[Giu84] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Monogr. Math. 80, Birkhäuser, Boston, Basel, Stuttgart, 1984.

[Has96] J. Haslinger and P. Neittaanmaki, *Finite Element Approximation for Optimal Shape, Material and Topology Design*, John Wiley, Chichester, UK, 1996.

[KO94] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, Berlin, Heidelberg, New York, 1994.

[Koh86] R. V. Kohn and G. Strang, *Optimal design and relaxation of variational problems.* I, Comm. Pure Appl. Math., 39, (1986), pp. 113–137.

[Lan77] E. M. Landis and G. P. Panasenko, *A theorem on the asymptotics of solutions of elliptic equations with coefficients periodic in all variables except one*, Soviet Math. Dokl., 18 (1977), pp. 1140–1143.

[Lio81] J. L. Lions, *Some Methods in the Mathematical Analysis of Systems and Their Control*, Gordon and Breach, New York, 1981.

[ole92] O. A. Oleinik and V. A. Kondratiev, *On asymptotic behaviour of solutions of some nonlinear elliptic equations in unbounded domains*, in Partial Differential Equations and Related Subjects, M. Miranda, ed., Trento, Italy, 1992.

[O.84] O. Pironneau, *Optimal Shape Design for Elliptic Systems*, Springer-Verlag, Berlin, Heidelberg, New York, 1984.

[MF76] J. Simon and F. Murat, *Sur le Contrôle Par un Domaine Géometrique*, Thèse d'Etat, Paris, 1976.

[Sok] J. Sokolowski, ed., *Shape optimization and scientific computations*, Appl. Math. Comput. Sci., 6 (1996), pp. 189–384.

[Tar86] L. Tartar, *Remarks on homogenization*, in Homogenization and Effective Moduli of Materials and Media (Minneapolis, MN, 1984/1985), IMA Vol. Math. Appl. 1, Springer-Verlag, New York, 1986, pp. 228–246.

[Ven78] V. Venkayya, *Structural optimization: A review and some recommendations*, Internat. J. Numer. Methods Engrg., 13 (1978), pp. 205–228.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 3, pp. 723–739

ZERO-SUM SEMI-MARKOV GAMES*

ANNA JAŚKIEWICZ†

Abstract. This paper deals with Borel state and action spaces zero-sum semi-Markov games under the expected long run average payoff criterion. The transition probabilities are assumed to satisfy some generalized geometric ergodicity conditions. The main result states that the optimality equation has a solution, which is approximated by the solutions of some ε-perturbed semi-Markov games. As a corollary, the existence of value and average optimal strategies for the players is established.

Key words. semi-Markov games, Borel state space, long run expected average payoff criterion, Bellman equation

AMS subject classifications. 90D10, 90D20, 90D05, 93E05

PII. S036301290038190X

1. Introduction and the model.

This paper deals with zero-sum undiscounted semi-Markov games in Borel spaces satisfying some stochastic stability conditions that imply so-called w-geometric ergodicity of the Markov chains induced by stationary strategies of the players [14, 15]. Assumptions of this type have recently been used in many papers on stochastic control [7, 8, 9] and Markov games [3, 10, 12, 20]. Such an approach enables us to consider unbounded payoffs, which is important from the point of view of many applications, e.g., to queueing models, networks [3], etc.

The main objective in this paper is to prove that the optimality equation for games under consideration has a solution. Our method of proof relies on considering some ε-perturbations of the transition structure of the game. Each ε-perturbed game can be solved by a natural iterative procedure. The solution to the optimality equation for the original game can be shown as a limit of some modifications of solutions for the ε-perturbed models as $\varepsilon \to 0$. This method of proof is essentially different from those used in the theory of Markov games [10, 12] based on the "vanishing discount factor approach."

There are few papers on Borel state space semi-Markov games [19]. The results, provided in [19] concern correlated equilibria in a class of strongly ergodic games with bounded payoffs and transition probabilities dominated by some probability measure on the state space. A predecessor of our paper is the article of Lal and Sinha [13]. They considered a countable state space model with bounded payoff function and much stronger ergodicity properties of the transition law. Their approach, by discounted games, allows only for bounded solutions to the optimality equation.

As noted by Cavazos-Cadena [5], under the same ergodicity assumptions as in [13], the players can restrict attention to some finite subset of the state space. This essentially reduces many interesting applications of stochastic games. Our paper is a considerable generalization of Lal and Sinha's paper [13] to Borel state spaces with unbounded costs. At the same time, our proof is based on different ideas, developed recently in [11].

*Received by the editors December 4, 2000; accepted for publication (in revised form) January 23, 2002; published electronically July 24, 2002. This work was supported by KBN grant 5 PO3A 014 20.

http://www.siam.org/journals/sicon/41-3/38190.html

†Institute of Mathematics, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland (ajaskiew@im.pwr.wroc.pl).

We consider zero-sum semi-Markov games (SMG) with Borel state and action spaces. By a Borel space we mean a nonempty Borel subset of a complete separable metric space, endowed with the σ-algebra $\mathcal{B}(X)$ of all its Borel subsets.

A *zero-sum SMG* is described by the following objects:

(i) X is the *set of states* for the game and is assumed to be a Borel space.

(ii) A and B are the *action spaces* for players 1 and 2, respectively, and are also assumed to be Borel spaces.

(iii) K_A and K_B are nonempty Borel subsets of $X \times A$ and $X \times B$, respectively. We assume that for each $x \in X$, the nonempty x-section

$$A(x) := \{a \in A : (x, a) \in K_A\}$$

of K_A represents the *set of actions available* to player 1 in the state x. Analogously, we define $B(x)$ for each $x \in X$. Define

$$K := \{(x, a, b) : x \in X,\ a \in A(x),\ \text{and } b \in B(x)\}.$$

It follows from [17] that K is a Borel subset of $X \times A \times B$.

(iv) q is a Borel measurable transition probability from K to X called the *law of motion among states*. If x is a state at some stage of the game and the players select actions $a \in A(x)$ and $b \in B(x)$, then $q(\cdot|x, a, b)$ is the probability distribution of the next state of the game.

(v) $Q(\cdot|x, a, b)$ is a distribution function of random variables $t_{n+1} := T_{n+1} - T_n$ and represents the *distribution of the holding (or sojourn) times*; T_n is the nth *jump time* of the process when it is in the state x_{n-1} and the actions are $a_{n-1} \in A(x_{n-1})$, $b_{n-1} \in B(x_{n-1})$ ($n = 0, 1, \ldots$ and $T_0 := 0$).

(vi) $r_1(x_n, a_n, b_n)$ is the *reward function* for player 1 (*cost function* for player 2) incurred at time T_n in state x_n by the control actions $a_n \in A(x_n)$ and $b_n \in B(x_n)$.

(vii) $r_2(x_n, a_n, b_n)$ denotes the *reward rate* for player 1 (*cost rate* for player 2) during the interval $[T_n, T_{n+1})$.

Throughout the remainder of this paper, $\mathbf{R}$ (resp., $\mathbf{R}_+$) stands for the set of real (resp., nonnegative real) numbers. By $P(D)$ we will denote the space of all probability measures on the Borel space D, equipped with the weak topology and the Borel σ-algebra.

Let H_n be the space of *admissible histories* up to the nth transition, i.e.,

$$H_n := (K \times \mathbf{R}_+)^n \times X, \quad \text{where} \quad H_0 := X.$$

An element of H_n is called a *partial history* of the game process and

$$h_n := (x_0, a_0, b_0, t_1, \ldots, x_{n-1}, a_{n-1}, b_{n-1}, t_n, x_n).$$

A randomized *strategy* π for player 1 is a sequence $\pi = (\pi_1, \pi_2, \ldots)$, where each π_n is a conditional probability $\pi_n(\cdot|h_n)$ on X, given the entire history h_n of the game up to its nth stage such that $\pi_n(A(x_n)|h_n) = 1$. (Of course, if $n = 0$, then $h_0 = x_0$.) The *class of all strategies* for player 1 will be denoted by Π. Let F be the set of all Borel measurable transition probabilities f from X to A such that $f(x) \in P(A(x))$ for each $x \in X$. It is well known that F is nonempty and every $f \in F$ can be identified with a Borel measurable mapping from X into $P(A)$ [4]. A *stationary strategy* for player 1 is a sequence $\pi = (f, f, \ldots)$ where $f \in F$. It can be identified with the mapping $f \in F$. Similarly, we define the set Γ (G) of all strategies (stationary strategies) for player 2.

Let $((X \times A \times B \times \mathbf{R}_+)^\infty, \mathcal{F})$ be the measurable space, where $\mathcal{F}$ denotes the corresponding product σ-algebra. Due to the theorem of C. Ionescu Tulcea (see Proposition V.1.1 in [16] or Chapter 7 in [4]), for each initial state $x \in X$ and strategies $\pi \in \Pi$, $\gamma \in \Gamma$ there exists a probability measure $P_x^{\pi\gamma}$ on $\mathcal{F}$ such that for all $A' \in \mathcal{B}(A)$, $B' \in \mathcal{B}(B)$, $X' \in \mathcal{B}(X)$, and $h_n = (x_0, a_0, b_0, t_1, \ldots, x_{n-1}, a_{n-1}, b_{n-1}, t_n, x_n)$ in H_n, $n = 0, 1, \ldots$,

$$P_x^{\pi\gamma}(x_0 = x) = 1,$$

$$P_x^{\pi\gamma}(a_n \in A'|h_n) = \pi_n(A'|h_n),$$

$$P_x^{\pi\gamma}(b_n \in B'|h_n) = \gamma_n(B'|h_n),$$

$$P_x^{\pi\gamma}(x_{n+1} \in X'|h_n, a_n, b_n, t_{n+1}) = q(X'|x_n, a_n, b_n),$$

and

$$P_x^{\pi\gamma}(t_{n+1} \le t|h_n, a_n, b_n) = Q(t|x_n, a_n, b_n), \quad t \ge 0.$$

By $E_x^{\pi\gamma}$ we denote the expectation operator with respect to the probability measure $P_x^{\pi\gamma}$.

Let $\tau(x, a, b)$ denote the *mean holding time*. When the process is in state x and the actions $a \in A(x)$, $b \in B(x)$ are chosen, then

$$\tau(x, a, b) := \int_0^\infty tQ(dt|x, a, b).$$

The *expected average reward per unit time* to player 1 is defined as

$$J(x, \pi, \gamma) := \liminf_{n\to\infty} \frac{E_x^{\pi\gamma}\left(\sum_{k=0}^{n-1} [r_1(x_k, a_k, b_k) + (T_{k+1} - T_k) r_2(x_k, a_k, b_k)]\right)}{E_x^{\pi\gamma}(T_n)}.$$

Making use of the properties of the conditional expectation, the last definition can be rewritten in the form

$$J(x, \pi, \gamma) = \liminf_{n\to\infty} \frac{E_x^{\pi\gamma}\left(\sum_{k=0}^{n-1} r(x_k, a_k, b_k)\right)}{E_x^{\pi\gamma}\left(\sum_{k=0}^{n-1} \tau(x_k, a_k, b_k)\right)},$$

where

$$r(x, a, b) := r_1(x, a, b) + \tau(x, a, b) r_2(x, a, b)$$

for every $(x, a, b) \in K$.

In section 2, we make assumptions under which the expected average reward considered in this paper is well defined.

For any initial state $x \in X$, we put

$$L(x) := \sup_{\pi\in\Pi} \inf_{\gamma\in\Gamma} J(x, \pi, \gamma) \quad \text{and} \quad U(x) := \inf_{\gamma\in\Gamma} \sup_{\pi\in\Pi} J(x, \pi, \gamma). \tag{1}$$

Then L (U) is called the *lower (upper) value* of the average payoff semi-Markov game. It is always true that $L(x) \le U(x)$ for $x \in X$. If $L(x) = U(x)$ for all $x \in X$, then this common function is called the *value* of the stochastic game and is denoted by V.

A strategy $\pi^* \in \Pi$ is called *optimal for player* 1 in the average payoff stochastic game if

$$\inf_{\gamma \in \Gamma} J(x, \pi^*, \gamma) = V(x)$$

for all $x \in X$. Similarly, a strategy $\gamma^* \in \Gamma$ is called *optimal for player* 2 in the average payoff stochastic game if

$$\sup_{\pi \in \Pi} J(x, \pi, \gamma^*) = V(x)$$

for all $x \in X$.

Let $x \in X$, $\nu \in P(A(x))$, and $\rho \in P(B(x))$. For any Borel measurable function $v : X \times A \times B \mapsto \mathbf{R}$, we put

$$v(x, \nu, \rho) := \int_{B(x)} \int_{A(x)} v(x, a, b) \nu(da) \rho(db),$$

provided that this integral exists. Further, for convenience we set

$$v(x, f, g) := v(x, f(x), g(x)),$$

where $f \in F$ and $g \in G$.

2. The assumptions. We are now ready to formulate our basic assumptions.

C1. For each $x \in X$, the sets $A(x)$ and $B(x)$ are nonempty and compact.

C2. For each $(x, a, b) \in K$, $r(x, \cdot, b)$ is upper semicontinuous on $A(x)$, and $r(x, a, \cdot)$ is lower semicontinuous on $B(x)$.

C3. For each $(x, a, b) \in K$ and every set $D \in \mathcal{B}(X)$, the function $q(D|x, \cdot, b)$ is continuous on $A(x)$, while $q(D|x, a, \cdot)$ is continuous on $B(x)$.

C4. For each $(x, a, b) \in K$, $\tau(x, \cdot, b)$ is continuous on $A(x)$, and $\tau(x, a, \cdot)$ is continuous on $B(x)$. Moreover, there exist positive constants m and M such that

$$m \le \tau(x, a, b) \le M$$

for all $(x, a, b) \in K$.

C5. (a) There exist a constant $L > 0$ and a Borel measurable function $w : X \mapsto \mathbf{R}$ such that $w(x) \ge 1$ for each $x \in X$ and $|r(x, a, b)| \le L w(x)$ for each $(x, a, b) \in K$.

(b) For each $(x, a, b) \in K$, the functions

$$\int w(y) q(dy|y, \cdot, b) \quad \text{and} \quad \int w(y) q(dy|x, a, \cdot)$$

are continuous on $A(x)$ and $B(x)$, respectively.

C6. (a) There exists a set $C \in \mathcal{B}(X)$ such that for some $\lambda \in (0, 1)$ and $\eta > 0$, we have

$$\int w(y) q(dy|x, a, b) \le \lambda w(x) + \eta 1_C(x)$$

for each $(x, a, b) \in K$. Here 1_C is the characteristic function of the set C and w is the function introduced in C5(a).

(b) The function w is bounded on C.

C7. There exist a $\delta \in (0,1)$ and a probability measure ζ_{fg}, concentrated on the Borel set C, associated with a pair of stationary strategies $(f,g) \in F \times G$ such that

$$q(D|x,f,g) \geq \delta \zeta_{fg}(D)$$

for each Borel set $D \subset C$ and $x \in C$.

REMARK 1. *Assumptions* C1–C4 *are standard, while* C5(a) *allows for unbounded payoffs. Conditions* C6 *and* C7 *imply that for every $f \in F$ and $g \in G$, the corresponding Markov chain is ψ_{fg}-irreducible for some σ-finite measure ψ_{fg} on X (see Theorem* 11.3.4 *and Chapter* 9 *in* [14]*). The ψ_{fg}-irreducibility means that if $\psi_{fg}(D) > 0$ for some set $D \in \mathcal{B}(X)$, then the chance that the Markov chain (starting at any $x \in X$ and induced by $f \in F$ and $g \in G$) ever enters D is positive. Lal and Sinha* [13] *employed much stronger ergodicity conditions for countable state space games, which correspond to* C6–C7 *with a bounded function w. The main disadvantage of such an approach is that the solution to the optimality (Bellman) equation is bounded. This fact excludes many interesting applications, even in the one-player case (i.e., dynamic programming); see* [1, 2, 3, 5, 21].

The set C in C7 *is called a* small set*, while* C6 *is called the* drift inequality [14]. *Such conditions and related ones are extensively used in the theory of Markov control processes and Markov games* [8, 9, 11, 12, 18, 20].

Assume that the function w in C5 is fixed. For a Borel measurable function $u : X \mapsto \mathbf{R}$, we define the weighted norm as

$$\|u\|_w := \sup_{x \in X} \frac{|u(x)|}{w(x)}.$$

We write L_w^∞ to denote the Banach space of all Borel measurable functions u for which $\|u\|_w$ is finite.

The following result is basic for this paper.

LEMMA 1. *Assume that* C6 *and* C7 *hold. Then for every $f \in F$ and $g \in G$, we have the following:*

(a) *The state process $\{x_n\}$ is a positive recurrent aperiodic Markov chain with the unique invariant probability measure, denoted by π_{fg}.*

(b) $\int w(x)\pi_{fg}(dx) < \infty$.

(c) $\{x_n\}$ *is w-uniformly ergodic, i.e., there exist $\theta > 0$ and $\alpha \in (0,1)$ such that*

$$\left| \int u(y) q^n(dy|x,f,g) - \int u(y)\pi_{fg}(dy) \right| \leq w(x)\|u\|_w \theta \alpha^n$$

for every $u \in L_w^\infty$ and $x \in X$, $n \geq 1$. Here $q^n(\cdot|x,f,g)$ denotes the n-stage transition probability induced by q and strategies f, g.

For a proof of (a) and (b) consult Theorem 11.3.4 and p. 116 in [14]. Part (c) follows from Theorem 2.3 in [15].

A conclusion to Lemma 1 is that for each $f \in F$ and $g \in G$ the expected average payoff is independent of the initial state:

$$J(f,g) := J(x,f,g) = \frac{\int r(y,f,g)\pi_{fg}(dy)}{\int \tau(y,f,g)\pi_{fg}(dy)}.$$

The following lemma is a consequence of our assumptions. It shows in particular how to obtain a solution to the Poisson equation associated with an arbitrary pair of strategies in $F \times G$. For a detailed discussion, see [8, 10].

LEMMA 2. *Suppose that assumptions* C5–C7 *are satisfied. Then for each pair of stationary strategies* $(f, g) \in F \times G$,

(a) *the function* h_{fg} *defined on* X *as*

$$h_{fg}(x) := E_x^{fg}\left(\sum_{n=0}^{\infty}[r(x_n, a_n, b_n) - J(f, g)\tau(x_n, a_n, b_n)]\right)$$

belongs to L_w^∞;

(b) *the pair* $(J(f, g), h_{fg})$ *is the unique solution to the equation*

$$J(f, g)\tau(x, f, g) + h_{fg}(x) = r(x, f, g) + \int h_{fg}(y)q(dy|x, f, g)$$

that satisfies the condition

$$\int h_{fg}(y)\pi_{fg}(dy) = 0.$$

The following auxiliary result follows easily from assumption C6(a) by induction.

LEMMA 3. *Let* C6(a) *hold and let* $\{x_n\}$ *denote the state space process under arbitrarily fixed strategies* $\pi \in \Pi$ *and* $\gamma \in \Gamma$. *Then for each initial state* $x \in X$, *any function* $u \in L_w^\infty$ *and* $n \geq 1$, *we obtain*

$$E_x^{\pi\gamma}|u(x_n)| \leq \|u\|_w\left(\lambda^n w(x) + \eta\sum_{k=0}^{n-1}\lambda^k\right) \leq \|u\|_w\left(w(x) + \frac{\eta}{1-\lambda}\right)$$

and

$$\limsup_{n\to\infty} E_x^{\pi\gamma}|u(x_n)| \leq \|u\|_w\frac{\eta}{1-\lambda}.$$

This lemma and conditions C4–C7 guarantee, for example, that the expected payoff introduced in section 1 is well defined.

We will also make use of the following fact, which follows easily from Proposition 10.1 in [22].

LEMMA 4. *Let* W_1 *be a compact metric space, and let* W_2 *be a nonempty set. Let* $\{v_n\}$ *be a nonincreasing sequence of functions* $v_n : W_1 \times W_2 \mapsto \mathbf{R}$ *such that* $v_n(\cdot, w_2)$ *is upper semicontinuous on* W_1 *for each* $n \geq 1$ *and* $w_2 \in W_2$. *Then*

$$\inf_{w_2\in W_2}\max_{w_1\in W_1}\lim_{n\to\infty} v_n(w_1, w_2) = \lim_{n\to\infty}\inf_{w_2\in W_2}\max_{w_1\in W_1} v_n(w_1, w_2).$$

3. The main results. Let $\varepsilon \in (0, 1)$ and δ_s be the probability measure concentrated at some fixed state $s \in X$. Put

$$q_\varepsilon(\cdot|x, a, b) := (1 - \varepsilon)q(\cdot|x, a, b) + \varepsilon\delta_s(\cdot)$$

for any $(x, a, b) \in K$.

The SMG with q replaced by q_ε is referred to as the ε-perturbed game (say, ε-perturbed SMG).

Let

$$\varepsilon_0 := \frac{1-\lambda}{2[w(s)-\lambda]}$$

and $0 < \varepsilon \le \varepsilon_0$. Then from C6(a), it follows that q_ε also satisfies the drift inequality with λ replaced by $(\lambda+1)/2 < 1$, $\eta > 0$ and the same small set C. Therefore from [15], we infer that for any $u \in L_w^\infty$ it holds that

$$\left| \int u(y) q_\varepsilon^n(dy|x,f,g) - \int u(y) \pi_{fg}^\varepsilon(dy) \right| \le w(x) \|u\|_w \theta_1 \alpha_1^n \tag{2}$$

and α_1 and θ_1 do not depend on ε. Thus a counterpart of Lemma 1 holds for the state process induced by any $f \in F$, $g \in G$, and q_ε. Consequently, for any $f \in F$, $g \in G$, we have

$$J^\varepsilon(f,g) := J^\varepsilon(x,f,g) = \frac{\int r(y,f,g)\pi_{fg}^\varepsilon(dy)}{\int \tau(y,f,g)\pi_{fg}^\varepsilon(dy)}$$

for all $x \in X$. Here $J^\varepsilon(x,f,g)$ stands for the expected average payoff in the ε-perturbed SMG.

PROPOSITION 1. *Under our assumptions* C1–C7*, there exists a positive constant d such that*

$$\sup_{g\in G} \sup_{f\in F} |J(f,g) - J^\varepsilon(f,g)| \le \varepsilon d \tag{3}$$

for $\varepsilon \in (0, \varepsilon_0)$. Constant d is expressed only by the constants used in assumptions C4–C7*.*

The above result has been established in our article [11]. It is worth pointing out that inequality (3) gives the rate of convergence between two expected payoffs: in the ε-perturbed and in the original SMG. Proposition 1 will play an important role in the proof of our main results.

Let $\varepsilon \in (0, \varepsilon_1)$, where $\varepsilon_1 := (1-\lambda)/(2w(s)+1-\lambda)$. Define

$$p(\cdot|x,a,b) := q_\varepsilon(\cdot|x,a,b) - \frac{\tau(x,a,b)\varepsilon}{M}\delta_s(\cdot)$$

for any $(x,a,b) \in K$. Clearly, p is a transition subprobability measure.

Let $T : L_w^\infty \mapsto L_w^\infty$ be the mapping defined by

$$(Tu)(x) = \min_{\rho\in P(B(x))} \max_{\nu\in P(A(x))} \left[r(x,\nu,\rho) + \int u(y) p(dy|x,\nu,\rho) \right], \tag{4}$$

where $x \in X$. Under our assumptions C1–C5 the operator T is well defined, and by Fan's theorem [6], we also have

$$(Tu)(x) = \max_{\nu\in P(A(x))} \min_{\rho\in P(B(x))} \left[r(x,\nu,\rho) + \int u(y) p(dy|x,\nu,\rho) \right] \tag{5}$$

for each $x \in X$.

Our first result deals with the optimality equation for the ε-perturbed SMGs.

THEOREM 1. *Assume* C1 *through* C7 *and that* $\varepsilon \in (0, \varepsilon_1)$. *Then there exists a fixed point* $l_\varepsilon \in L_w^\infty$ *of* T, *i.e.,*

$$l_\varepsilon(x) = (Tl_\varepsilon)(x) \tag{6}$$

for every $x \in X$. *Moreover,*

$$V_\varepsilon = \varepsilon l_\varepsilon(s)/M \tag{7}$$

is the value of the ε*-perturbed SMG.*

Let f and g be maxmin and minmax stationary strategies, respectively, obtained on the right-hand side in (6) (see also (4), (5)). From Lemma 2, Theorem 1, and the dynamic programming methods [8, 11], we conclude that

$$\begin{aligned} V_\varepsilon &= J^\varepsilon(f,g) \\ &= \sup_{\pi\in\Pi}\inf_{\gamma\in\Gamma} J^\varepsilon(x,\pi,\gamma) = \inf_{\gamma\in\Gamma}\sup_{\pi\in\Pi} J^\varepsilon(x,\pi,\gamma) \\ &= \sup_{\pi\in\Pi} J^\varepsilon(x,\pi,g) = \inf_{\gamma\in\Gamma} J^\varepsilon(x,f,\gamma) \end{aligned}$$

for every $x \in X$.

The following assertion concerns the optimality equation for the original SMG.

THEOREM 2. *Assume* C1–C7. *Then there exists a constant* V, *which is the value of the original SMG, and a function* $h \in L_w^\infty$ *unique up to an additive constant such that*

$$\begin{aligned} h(x) &= \min_{\rho\in P(B(x))}\max_{\nu\in P(A(x))}\left[r(x,\nu,\rho) + \int h(y)q(dy|x,\nu,\rho) - V\tau(x,\nu,\rho)\right] \\ &= \max_{\nu\in P(A(x))}\min_{\rho\in P(B(x))}\left[r(x,\nu,\rho) + \int h(y)q(dy|x,\nu,\rho) - V\tau(x,\nu,\rho)\right] \end{aligned}$$

for each $x \in X$. *Moreover, there exist* $f^0 \in F$ *and* $g^0 \in G$ *such that*

$$\begin{aligned} h(x) &= r(x,f^0,g^0) + \int h(y)q(dy|x,f^0,g^0) - V\tau(x,f^0,g^0) \\ &= \max_{\nu\in P(A(x))}\left[r(x,\nu,g^0) + \int h(y)q(dy|x,\nu,g^0) - V\tau(x,\nu,g^0)\right] \\ &= \min_{\rho\in P(B(x))}\left[r(x,f^0,\rho) + \int h(y)q(dy|x,f^0,\rho) - V\tau(x,f^0,\rho)\right] \end{aligned}$$

for each $x \in X$.

The above result implies, by the dynamic programming arguments [11], that

$$\begin{aligned} V &= J(f^0,g^0) \\ &= \sup_{\pi\in\Pi}\inf_{\gamma\in\Gamma} J(x,\pi,\gamma) = \inf_{\gamma\in\Gamma}\sup_{\pi\in\Pi} J(x,\pi,\gamma) \\ &= \sup_{\pi\in\Pi} J(x,\pi,g) = \inf_{\gamma\in\Gamma} J(x,f,\gamma) \end{aligned}$$

for every $x \in X$. Together with (3), the remarks lead to the following.

COROLLARY 1. *Let* $f \in F$, $g \in G$ *be the maxmin and minmax Borel measurable stationary strategies on the right side of* (6). *Then* f *and* g *are* $2d\varepsilon$*-optimal strategies in the original SMG.*

REMARK 2. *As already mentioned, the constant d can be computed using only the primitive data given in our assumptions (*C4–C6*) and an interesting result of Meyn and Tweedie* [15]. *In view of this, Corollary* 1 *is of some importance. In order to find almost optimal strategies for the players in the SMG, it is enough to solve the* ε*-perturbed SMG by iterative procedure. In this way we get the fixed point* l_ε *of operator* T *(see the proof of Theorem* 1*).*

COROLLARY 2. *From Proposition* 1*, it follows that*

$$V = \lim_{\varepsilon \to 0} V_\varepsilon = \lim_{\varepsilon \to 0} \frac{\varepsilon l_\varepsilon(s)}{M}$$

exists. Moreover, V *is a part of the solution to the optimality equation (see the proof of Theorem* 2*).*

4. The proofs.

Proof of Theorem 1. First we show that the operator T has a fixed point. Let $\varepsilon \in (0, \varepsilon_1)$. Using C6(a), it is easy to establish the drift inequality for p:

$$\int w(y)p(dy|x,a,b) \le (1-\varepsilon)\left(\frac{1+\lambda}{2}w(x) + \eta 1_C(x)\right). \tag{8}$$

Let $\mathcal{E}_x^{\pi\gamma}$ denote the expectation operator with respect to the measure induced by the transition subprobability measure p and any strategies $\pi \in \Pi$ and $\gamma \in \Gamma$. Likewise in Lemma 3, one can easily prove by induction using (8) that for any $u \in L_w^\infty$, it holds that

$$\mathcal{E}_x^{\pi\gamma}(|u(x_n)|) \le (1-\varepsilon)^n w(x) N_1$$

for some constant N_1.

Put $u_0(\cdot) \equiv 0$ and define

$$l_n(x) := \inf_{\gamma\in\Gamma} \sup_{\pi\in\Pi} \mathcal{E}_x^{\pi\gamma}\left(\sum_{k=0}^{n-1} r(x_k, a_k, b_k)\right).$$

Using minmax measurable selection theorems [17] and backward induction, one can show that

$$l_n(x) = (T^n u_0)(x)$$

for every $x \in X$. Now, note that under C5(a) there exists a constant N_2 such that for any natural n, N $(n > N)$, we have

$$\begin{aligned}
|l_n(x) - l_N(x)| &\le \sup_{\pi\in\Pi}\sup_{\gamma\in\Gamma}\left|\mathcal{E}_x^{\pi\gamma}\left(\sum_{k=0}^{n-1} r(x_k,a_k,b_k)\right) - \mathcal{E}_x^{\pi\gamma}\left(\sum_{k=0}^{N-1} r(x_k,a_k,b_k)\right)\right| \\
&\le \sup_{\pi\in\Pi}\sup_{\gamma\in\Gamma}\mathcal{E}_x^{\pi\gamma}\left(\sum_{k=N}^{n-1}|r(x_k,a_k,b_k)|\right) \\
&\le \sum_{k=N}^{n-1}(1-\varepsilon)^k w(x) N_2 \\
&\le \frac{(1-\varepsilon)^N}{\varepsilon} w(x) N_2.
\end{aligned}$$

This implies that $\{l_n\}$ is a Cauchy sequence in L_w^∞. Put

$$l_\varepsilon(x) := \lim_{n\to\infty} l_n(x), \quad x \in X.$$

Note that $l_\varepsilon \in L_w^\infty$ and

$$\sup_{x\in X} \frac{|l_n(x) - l_\varepsilon(x)|}{w(x)} \longrightarrow 0 \quad \text{when} \quad n \to \infty. \tag{9}$$

For $n \geq 2$ it holds that

$$l_n(x) = (T^n u_0)(x) = (Tl_{n-1})(x).$$

It only remains to prove that $Tl_{n-1} \to Tl_\varepsilon$ when $n \to \infty$. We have

$$\begin{aligned}|(Tl_n)(x) - (Tl_\varepsilon)(x)| \leq \varepsilon|l_n(s) - l_\varepsilon(s)| + \frac{\varepsilon}{M} \max_{a\in A(x)} \max_{b\in B(x)} \tau(x,a,b)|l_n(s) - l_\varepsilon(s)| \\ + (1-\varepsilon) \max_{a\in A(x)} \max_{b\in B(x)} \int \frac{|l_n(y) - l_\varepsilon(y)|}{w(y)} w(y) q(dy|x,a,b).\end{aligned}$$

Now from C1, C5(b), C4, and (9), we obtain that $Tl_n \to Tl_\varepsilon$ as $n \to \infty$. Hence l_ε is the fixed point of T.

The proof is completed by showing that V_ε (defined as in (7)) is the value of game for ε-perturbed SMG. In order to prove it, note that for every $\nu \in P(A(x))$ we have

$$l_\varepsilon(x) \geq r(x,\nu,g) + \int l_\varepsilon(y) q_\varepsilon(dy|x,\nu,g) - \tau(x,\nu,g)V_\varepsilon,$$

where g is a minmax stationary strategy for player 2 in (6) (see also (4)). If $k \geq 1$, then

$$\begin{aligned}\int l_\varepsilon(y) q_\varepsilon(dy|x_k,\nu,g) &= r(x_k,\nu,g) - \tau(x_k,\nu,g)V_\varepsilon \\ &\quad + \int l_\varepsilon(y) q_\varepsilon(dy|x_k,\nu,g) \\ &\quad - r(x_k,\nu,g) + \tau(x_k,\nu,g)V_\varepsilon \\ &\leq l_\varepsilon(x_k) - r(x_k,\nu,g) + \tau(x_k,\nu,g)V_\varepsilon.\end{aligned}$$

Iterating this inequality, we easily get

$$V_\varepsilon \mathbf{E}_x^{\pi g}\left(\sum_{k=0}^{n-1} \tau(x_k,a_k,b_k)\right) \geq \mathbf{E}_x^{\pi g} l_\varepsilon(x_{n-1}) - l_\varepsilon(x) + \mathbf{E}_x^{\pi g}\left(\sum_{k=0}^{n-1} r(x_k,a_k,b_k)\right). \tag{10}$$

Dividing both sides of (10) by $\mathbf{E}_x^{\pi g}\left(\sum_{k=0}^{n-1} \tau(x_k,a_k,b_k)\right)$, using condition C4 and Lemma 3, and taking $\liminf$ as $n \to \infty$, we obtain

$$V_\varepsilon \geq \sup_{\pi\in\Pi} J^\varepsilon(x,\pi,g) \geq U_\varepsilon(x), \tag{11}$$

where $U_\varepsilon(x)$ is the upper value for ε-perturbed SMGs. On the other hand, if we now take the maxmin stationary strategy $f \in F$ for player 1, it holds that

$$l_\varepsilon(x) \leq r(x,f,\rho) + \int l_\varepsilon(y) q_\varepsilon(dy|x,f,\rho) - \tau(x,f,\rho)V_\varepsilon$$

for every $\rho \in P(B(x))$. For $k \geq 1$, we get

$$\begin{aligned}\int l_\varepsilon(y) q_\varepsilon(dy|x_k, f, \rho) &= r(x_k, f, \rho) - \tau(x_k, f, \rho) V_\varepsilon \\ &\quad + \int l_\varepsilon(y) q_\varepsilon(dy|x_k, f, \rho) \\ &\quad - r(x_k, \nu, g) + \tau(x_k, f, \rho) V_\varepsilon \\ &\geq l_\varepsilon(x_k) - r(x_k, f, \rho) + \tau(x_k, f, \rho) V_\varepsilon.\end{aligned}$$

Along similar lines as above, we obtain

$$V_\varepsilon \geq \inf_{\gamma \in \Gamma} J_\varepsilon(x, f, \gamma) \geq L_\varepsilon(x), \tag{12}$$

where $L_\varepsilon(x)$ is the lower value for ε-perturbed SMGs. Finally, from (11) and (12) we infer

$$V_\varepsilon = \inf_{\gamma \in \Gamma} J_\varepsilon(x, f, \gamma) = L_\varepsilon(x) = \sup_{\pi \in \Pi} J_\varepsilon(x, \pi, g) = U_\varepsilon(x)$$

for every $x \in X$. This completes the proof. □

The proof of Theorem 2 contains a solution to the Bellman equation of the original SMG. Although V_ε tends to V, the sequence of the fixed points of T need not converge to any solution to the optimality equation, because the sequence $\{l_\varepsilon\}$ can be unbounded. By an appropriate modification of this sequence, we will obtain a new one, which is bounded in L_w^∞ and can be used for constructing the solution to the optimality equation. In order to emphasize the main points, the proof is divided into four steps.

Proof of Theorem 2.

Step 1. First we modify the sequence $\{l_\varepsilon\}$ in order to make it uniformly bounded in the space L_w^∞. Let $f \in F$, $g \in G$ be the maxmin and minmax strategies on the right side of (6), respectively. From (6) we have

$$\begin{aligned}l_\varepsilon(x) &= r(x, f, g) + \int l_\varepsilon(y) p(dy|x, f, g) \\ &= r(x, f, g) + \int l_\varepsilon(y) q_\varepsilon(dy|x, f, g) - \frac{\varepsilon}{M} \tau(x, f, g) l_\varepsilon(s).\end{aligned} \tag{13}$$

Define

$$z_\varepsilon(x, f, g) := r(x, f, g) - \frac{\varepsilon}{M} l_\varepsilon(s) \tau(x, f, g). \tag{14}$$

Put

$$h_\varepsilon(x) := \mathbf{E}_x^{fg} \left(\sum_{n=0}^{\infty} z_\varepsilon(x_n, a_n, b_n) \right), \tag{15}$$

where $\mathbf{E}_x^{fg}$ is the expectation operator with respect to the probability measure on $(\Omega, \mathcal{F})$ induced by f, g, and q_ε. Clearly,

$$|z_\varepsilon(x, f, g)| \leq w(x) N_3 \tag{16}$$

for some constant N_3, $f \in F$, $g \in G$, and $\varepsilon \in (0, \varepsilon_1)$. From (14) and (7), we conclude that

$$\int z_\varepsilon(x,f,g)\pi^\varepsilon_{fg}(dx) = 0. \tag{17}$$

Under our assumptions it follows from (15), (17), and the Markov property that

$$h_\varepsilon(x) = z_\varepsilon(x,f,g) + \int h_\varepsilon(y) q_\varepsilon(dy|x,f,g). \tag{18}$$

Recall that the drift inequality is satisfied by q_ε with λ replaced by $(\lambda+1)/2$. Therefore using (2), C5(a), (17), and (16), we infer that there exists some constant N_4 for which

$$|h_\varepsilon(x)| \le w(x) N_4 \tag{19}$$

for all $x \in X$ and $\varepsilon \in (0, \varepsilon_1)$.

Subtracting (13) from (18), it follows that

$$w_\varepsilon(x) = \int w_\varepsilon(y) q_\varepsilon(dy|x,f,g),$$

with $w_\varepsilon(x) = l_\varepsilon(x) - h_\varepsilon(x)$. Hence

$$w_\varepsilon(x) = \int w_\varepsilon(y)\pi^\varepsilon_{fg}(dy) = N_5,$$

where N_5 is some constant. Thus $h_\varepsilon(x) = l_\varepsilon(x) - N_5$ for all $x \in X$ and for every $\varepsilon \in (0, \varepsilon_1)$. Consequently, one can replace $l_\varepsilon(x)$ in (6) by $h_\varepsilon(x)$ and obtain

$$\begin{aligned} h_\varepsilon(x) &= \max_{\nu\in P(A(x))} \min_{\rho\in P(B(x))} \left[r(x,\nu,\rho) - \frac{\varepsilon}{M} l_\varepsilon(s)\tau(x,\nu,\rho) + \int h_\varepsilon(y) q_\varepsilon(dy|x,\nu,\rho) \right] \\ &= \min_{\rho\in P(B(x))} \max_{\nu\in P(A(x))} \left[r(x,\nu,\rho) - \frac{\varepsilon}{M} l_\varepsilon(s)\tau(x,\nu,\rho) + \int h_\varepsilon(y) q_\varepsilon(dy|x,\nu,\rho) \right] \\ &= r(x,f,g) - \frac{\varepsilon}{M} l_\varepsilon(s)\tau(x,f,g) + \int h_\varepsilon(y) q_\varepsilon(dy|x,f,g). \end{aligned} \tag{20}$$

Step 2. The second part of our proof starts with showing that V (see Corollary 2) is indeed the value of the original SMG and the players have optimal stationary strategies.

Fix $x \in X$ and choose any sequence $\{\varepsilon_n\}_{n\ge 2}$ converging to zero. For every $\nu \in P(A(x))$ there exists a sequence $\{\rho_{\varepsilon_n}\}$ (abbreviated $\{\rho_n\}$) (independent on ν) that attains minimum on the right side of inequality (20), i.e.,

$$\begin{aligned} h_{\varepsilon_n}(x) &= \min_{\rho\in P(B(x))} \max_{\nu\in P(A(x))} \left[r(x,\nu,\rho) + \int h_{\varepsilon_n}(y) q_{\varepsilon_n}(dy|x,\nu,\rho) - \tau(x,\nu,\rho) V_{\varepsilon_n} \right] \\ &= \max_{\nu\in P(A(x))} \left[r(x,\nu,\rho_n) + \int h_{\varepsilon_n}(y) q_{\varepsilon_n}(dy|x,\nu,\rho_n) - \tau(x,\nu,\rho_n) V_{\varepsilon_n} \right] \\ &\ge r(x,\nu,\rho_n) + \int h_{\varepsilon_n}(y) q_{\varepsilon_n}(dy|x,\nu,\rho_n) - \tau(x,\nu,\rho_n) V_{\varepsilon_n}. \end{aligned} \tag{21}$$

Since (19) holds, we can take the lim inf on both sides in (21) as $n \to \infty$. By Corollary 2, we put $V = \lim_{n\to\infty} V_{\varepsilon_n}$. The set $P(B(x))$ is compact when endowed with the weak

topology. Hence there exists a subsequence $\{n(k)\}$ of $\{n\}$ such that $\{\rho_{n(k)}\}$ converges weakly to some $\rho_x \in P(B(x))$, and simultaneously $h_*(x) := \liminf_{n\to\infty} h_{\varepsilon_n}(x) = \lim_{k\to\infty} h_{\varepsilon_{n(k)}}(x)$. From (21), Fatou's lemma for varying probability measures (see [8] or [18, Lemma 4]), we obtain

$$\begin{aligned}
h_*(x) &= \liminf_{n\to\infty} h_{\varepsilon_n}(x) \geq \liminf_{n\to\infty} \left(r(x,\nu,\rho_n) + \int h_{\varepsilon_n}(y) q_{\varepsilon_n}(dy|x,\nu,\rho_n) - \tau(x,\nu,\rho_n) V_{\varepsilon_n} \right) \\
&= \liminf_{k\to\infty} \left(r(x,\nu,\rho_{n(k)}) + \int h_{\varepsilon_{n(k)}}(y) q_{\varepsilon_{n(k)}}(dy|x,\nu,\rho_{n(k)}) - \tau(x,\nu,\rho_{n(k)}) V_{\varepsilon_{n(k)}} \right) \\
&\geq r(x,\nu,\rho_x) + \int \liminf_{k\to\infty} h_{\varepsilon_{n(k)}}(y) q(dy|x,\nu,\rho_x) - \tau(x,\nu,\rho_x) V \\
&\geq r(x,\nu,\rho_x) + \int h_*(y) q(dy|x,\nu,\rho_x) - \tau(x,\nu,\rho_x) V.
\end{aligned}$$

Because the last inequality holds for every $\nu \in P(A(x))$, we therefore get in the obvious manner

$$h_*(x) \geq \min_{\rho\in P(B(x))} \max_{\nu\in P(A(x))} \left[r(x,\nu,\rho) + \int h_*(y) q(dy|x,\nu,\rho) - \tau(x,\nu,\rho) V \right] \tag{22}$$

for each $x \in X$. By (22) and minmax measurable selection theorem [17], there exists some $g^* \in G$ such that

$$h_*(x) \geq r(x,\nu,g^*) + \int h_*(y) q(dy|x,\nu,g^*) - \tau(x,\nu,g^*) V.$$

From dynamic programming [11], it follows that

$$V \geq \sup_{\pi\in\Pi} J(x,\pi,g^*) \geq U(x). \tag{23}$$

On the other hand, it is easily seen from (20) that for each $\rho \in P(B(x))$ there exists a sequence $\{\nu_{\varepsilon_n}\}$ (abbreviated $\{\nu_n\}$) (independent on ρ) such that

$$h_{\varepsilon_n}(x) \leq r(x,\nu_n,\rho) + \int h_{\varepsilon_n}(y) q_{\varepsilon_n}(dy|x,\nu_n,\rho) - \tau(x,\nu_n,\rho) V_{\varepsilon_n}.$$

Taking lim sup as $n \to \infty$, we conclude

$$h^*(x) \leq \max_{\nu\in P(A(x))} \min_{\rho\in P(B(x))} \left[r(x,\nu,\rho) + \int h^*(y) q(dy|x,\nu,\rho) - \tau(x,\nu,\rho) V \right], \tag{24}$$

where $h^*(x) := \limsup_{n\to\infty} h_{\varepsilon_n}(x)$. Let $f^* \in F$ be the maxmin strategy in (24). Applying similar arguments, we infer

$$V \leq \inf_{\gamma\in\Gamma} J(x,f^*,\gamma) \leq L(x) \tag{25}$$

for each $x \in X$. Consequently, combining (23) and (25) we get

$$V = \inf_{\gamma\in\Gamma} J(x,f^*,\gamma) = \sup_{\pi\in\Pi} J(x,\pi,g^*) = L(x) = U(x),$$

which is the desired conclusion.

Step 3. Now we are left with the task of finding a solution to the optimality equation of the original SMG. Define

$$\hat{r}(x,a,b) := r(x,a,b) - V\tau(x,a,b)$$

for every $(x,a,b) \in K$. Because V is the value of game in the original SMG and (f^*,g^*) is the pair of the optimal stationary strategies, therefore

$$J_{\hat{r}}(f^*,g^*) := \int \hat{r}(y,f^*,g^*)\pi_{f^*g^*}(dy) = 0. \tag{26}$$

From Lemma 2 there exists a function $h_{f^*g^*} \in L_w^\infty$ such that

$$h_{f^*g^*}(x) = \hat{r}(x,f^*,g^*) + \int h_{f^*g^*}(y)q(dy|x,f^*,g^*) \tag{27}$$

and

$$h_{f^*g^*}(x) = E_x^{f^*g^*}\left(\sum_{k=0}^{\infty} \hat{r}(x_k,a_k,b_k)\right)$$

for all $x \in X$. It is evident that

$$h_*(x) \geq \hat{r}(x,f^*,g^*) + \int h_*(y)q(dy|x,f^*,g^*) \tag{28}$$

and

$$h^*(x) \leq \hat{r}(x,f^*,g^*) + \int h^*(y)q(dy|x,f^*,g^*) \tag{29}$$

for every $x \in X$. If we now subtract (28) from (27), (29) from (27), we get

$$h_*(x) - h_{f^*g^*}(x) \geq \int \left(h_*(y) - h_{f^*g^*}(y)\right) q(dy|x,f^*,g^*)$$

and

$$h^*(x) - h_{f^*g^*}(x) \leq \int \left(h^*(y) - h_{f^*g^*}(y)\right) q(dy|x,f^*,g^*).$$

Iterating these inequalities and taking the limit as $n \to \infty$, we obtain by Lemma 1

$$h_*(x) - h_{f^*g^*}(x) \geq \int \left(h_*(y) - h_{f^*g^*}(y)\right) \pi_{f^*g^*}(dy)$$

and

$$h^*(x) - h_{f^*g^*}(x) \leq \int \left(h^*(y) - h_{f^*g^*}(y)\right) \pi_{f^*g^*}(dy).$$

These inequalities hold for every $x \in X$; therefore

$$h_*(x) - h_{f^*g^*}(x) = \inf_{x\in X}\left(h_*(x) - h_{f^*g^*}(x)\right) = d_1 \quad \text{for a.e. } \pi_{f^*g^*}$$

and

$$h^*(x) - h_{f^*g^*}(x) = \sup_{x\in X}\left(h^*(x) - h_{f^*g^*}(x)\right) = d_2 \quad \text{for a.e. } \pi_{f^*g^*}.$$

Thus (see also [8, 10]) there exist Borel sets $X_1 \subset X$ and $X_2 \subset X$ such that $\pi_{f^*g^*}(X_1) = 1$ and $\pi_{f^*g^*}(X_2) = 1$, on which the above equalities are true, respectively. Moreover, $\pi_{f^*g^*}(X_1 \cap X_2) = 1$. By Proposition 4.2.3 and Theorem 10.4.9 in [14], the Markov chain induced by f^*, g^*, and q has an absorbing Borel set $Z \subset X_1 \cap X_2$, that is, $q(Z|x, f^*, g^*) = 1$ for all $x \in Z$.

Now we show that the optimality equation holds on the set Z. Let $x \in Z$. Substituting $h_{f^*g^*}(x) + d_1$ for $h_*(x)$ in (22) and $h_{f^*g^*}(x) + d_2$ for $h^*(x)$ in (24), we easily get

$$h_{f^*g^*}(x) \geq \max_{\nu \in P(A(x))} \left[\hat{r}(x, \nu, g^*) + \int h_{f^*g^*}(y) q(dy|x, \nu, g^* \right],$$

$$h_{f^*g^*}(x) \leq \min_{\rho \in P(B(x))} \left[\hat{r}(x, f^*, \rho) + \int h_{f^*g^*}(y) q(dy|x, f^*, \rho) \right].$$

Combining these inequalities with (27), we have

$$\begin{aligned} h_{f^*g^*}(x) &= \hat{r}(x, f^*, g^*) + \int h_{f^*g^*}(y) q(dy|x, f^*, g^*) \\ &= \max_{\nu \in P(A(x))} \left[\hat{r}(x, \nu, g^*) + \int h_{f^*g^*}(y) q(dy|x, \nu, g^* \right] \\ &= \min_{\rho \in P(B(x))} \left[\hat{r}(x, f^*, \rho) + \int h_{f^*g^*}(y) q(dy|x, f^*, \rho) \right] \end{aligned}$$

for all $x \in Z$, which gives the assertion on the Borel absorbing set Z.

In order to obtain this equation for each $x \in X$, we improve h_* on the set $X \setminus Z$. For this, we define inductively a sequence of Borel measurable functions $h_k \in L_w^\infty$ and some sequences $\{f_k\}$, $\{g_k\}$ of stationary strategies for players 1 and 2, respectively. Let $h_0 := h_*$, $f_0 := f^*$, and $g_0 := g^*$. Suppose that $h_0, \ldots, h_k$, $f_0, \ldots, f_k$, and $g_0, \ldots, g_k$ have been defined. Let

$$h_{k+1}(x) := \min_{\rho \in P(B(x))} \max_{\nu \in P(A(x))} \left[\hat{r}(x, \nu, \rho) + \int h_k(y) q(dy|x, \nu, \rho) \right] \tag{30}$$

for each $x \notin Z$ and $h_{k+1}(x) := h_*(x)$ for every $x \in Z$. By Fan's minmax theorem [6], we have

$$h_{k+1}(x) = \max_{\nu \in P(A(x))} \min_{\rho \in P(B(x))} \left[\hat{r}(x, \nu, \rho) + \int h_k(y) q(dy|x, \nu, \rho) \right]. \tag{31}$$

Put $f_{k+1}(x) := f^*(x)$ and $g_{k+1}(x) := g^*(x)$ for $x \in Z$. Let f_{k+1} and g_{k+1} be any maxmin and minmax Borel measurable stationary strategies of the players obtained for the right-hand sides in (31) and (30), defined on the set $X \setminus Z$ [17]. From our construction and the optimality equation on the set Z, it follows that

$$h_{k+1}(x) = \min_{\rho \in P(B(x))} \max_{\nu \in P(A(x))} \left[\hat{r}(x, \nu, \rho) + \int h_k(y) q(dy|x, \nu, \rho) \right] \tag{32}$$

for each $x \in X$. By C5(a) and C6(a), every h_k belongs to L_w^∞. Using (22), it is easy to see that the sequence $\{h_k\}$ is nonincreasing. Since Z is absorbing for the Markov chain induced by f^*, g^* this set is absorbing as well for the Markov chain induced by

any f_k, g_k constructed above. Consequently, we infer that $\pi_{f_k g_k} = \pi_{f^* g^*}$ for every k and thus by (26),

$$J_{\hat{r}}(f_k, g_k) = J_{\hat{r}}(f^*, g^*) = 0 \quad \text{and} \quad \int h_k(y)\pi_{f_k g_k}(dy) = \int h_*(y)\pi_{f^* g^*}(dy). \tag{33}$$

We shall prove that $\lim_{k\to\infty} h_k(x)$ exists and belongs to L_w^∞. For this, we show that there exists a constant ξ such that

$$h_k(x) \geq \xi w(x) \tag{34}$$

for every k and $x \in X$. Note that

$$h_k(x) \geq h_{k+1}(x) = \hat{r}(x, f_k, g_k) + \int h_k(y) q(dy|x, f_k, g_k)$$

for every k and $x \in X$. Iterating this inequality, we obtain

$$h_k(x) \geq E_x^{f_k g_k}\left(\sum_{m=0}^{n-1} \hat{r}(x_m, a_m, b_m)\right) + \int h_k(y) q^n(dy|x, f_k, g_k) \tag{35}$$

for every $n \geq 1$ and $x \in X$. Using (33), we conclude from (35) that

$$h_k(x) \geq E_x^{f_k g_k}\left(\sum_{m=0}^{\infty} [\hat{r}(x_m, a_m, b_m) - J_{\hat{r}}(f_k, g_k)]\right) + \int h_*(y)\pi_{f^* g^*}(dy) \tag{36}$$

for every $x \in X$. Now (34) follows from (36), Lemma 1 (with $f = f_k$ and $g = g_k$), and the fact that $h_* \in L_w^\infty$. Taking the limit in (32) as $k \to \infty$ on both sides, using the monotone convergence theorem, and using Lemma 4, we now easily obtain the optimality equation with the function $h(x) := \lim_{k\to\infty} h_k(x)$ and the constant V. The Borel measurable maxmin and minmax strategies that satisfy our assertion can be obtained by applying Fan's minmax theorem [6] and an appropriate measurable selection theorem [17].

Step 4. The uniqueness of the solution to the Bellman equation can be concluded from the proof in the Markov case [12] by taking the payoff function $r(x, a, b) - V\tau(x, a, b)$, where $(x, a, b) \in K$. □

REMARK 3. *In our proof we have taken into account the function h_* as our point of departure and have defined the nonincreasing sequence $\{h_k\}$ as in* (30). *But it is possible to define inductively the nondecreasing sequence $\{h_k\}$ as follows:*

$$h_{k+1}(x) := \max_{\nu \in P(A(x))} \min_{\rho \in P(B(x))} \left[\hat{r}(x, \nu, \rho) + \int h_k(y) q(dy|x, \nu, \rho)\right]$$

with $h_0 := h^*$ (recall (24)*). This sequence can also be used to construct the solution to the optimality equation. A simple modification of the proof of Theorem* 2 *(Step* 3*) yields an upper bound for the sequence $\{h_k\}$. More precisely, one can prove that there exists a positive constant ξ such that*

$$h_k(x) \leq \xi w(x)$$

for all $x \in X$. Then the solution to the optimality equation is the function $h(x) = \lim_{k\to\infty} h_k(x) := \sup_{k\geq 1} h_k(x)$.

Acknowledgments. I wish to express my thanks to Professor Andrzej S. Nowak for suggesting the problem and for several helpful conversations. I am also indebted to the referee for many useful remarks.

REFERENCES

[1] A. Arapostathis, V.S. Borkar, E. Fernández-Gaucherand, M.K. Ghosh, and S.I. Marcus, *Discrete-time controlled Markov processes with average cost criterion: A survey*, SIAM J. Control Optim., 31 (1993), pp. 282–344.

[2] E. Altman and A. Hordijk, *Zero-sum Markov games and worst-case optimal control of queueing systems*, Queueing Systems Theory Appl., 21 (1995), pp. 415–447.

[3] E. Altman, A. Hordijk, and F.M. Spieksma, *Contraction conditions for average and α-discount optimality in countable state Markov games with unbounded rewards*, Math. Oper. Res., 22 (1997), pp. 588–618.

[4] D.P. Bertsekas and S.E. Shreve, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.

[5] R. Cavazos-Cadena, *Recent results on conditions for the existence of average optimal stationary policies*, Ann. Oper. Res., 28 (1991), pp. 3–28.

[6] K. Fan, *Minimax theorems*, Proc. Natl. Acad. Sci. USA, 39 (1953), pp. 42–47.

[7] O. Hernández-Lerma and J.B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer-Verlag, New York, 1996.

[8] O. Hernández-Lerma and J.B. Lasserre, *Further Topics on Discrete-Time Markov Control Processes*, Springer-Verlag, New York, 1999.

[9] O. Hernández-Lerma, O. Vega-Amaya, and G. Carrasco, *Sample-path optimality and variance-minimization of average cost Markov control processes*, SIAM J. Control Optim., 38 (1999), pp. 79–93.

[10] O. Hernández-Lerma and J.B. Lasserre, *Zero-sum stochastic games in Borel spaces: Average payoff criteria*, SIAM J. Control Optim., 39 (2001), pp. 1520–1539.

[11] A. Jaśkiewicz, *An approximation approach to ergodic semi-Markov control processes*, Math. Methods Oper. Res., 54 (2001), pp. 1–19.

[12] A. Jaśkiewicz and A.S. Nowak, *On the optimality equation for zero-sum ergodic stochastic games*, Math. Methods Oper. Res., 54 (2001), pp. 291–301.

[13] A.K. Lal and S. Sinha, *Zero-sum two person semi-Markov games*, J. Appl. Probab., 29 (1992), pp. 56–72.

[14] S.P. Meyn and R.L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, New York, 1993.

[15] S.P. Meyn and R.L. Tweedie, *Computable bounds for geometric convergence rates of Markov chains*, Ann. Appl. Probab., 4 (1994), pp. 981–1011.

[16] J. Neveu, *Mathematical Foundations of the Calculus of Probability*, Holden–Day, San Francisco, 1965.

[17] A.S. Nowak, *Measurable selection theorems for minimax stochastic optimization problems*, SIAM J. Control Optim., 23 (1985), pp. 466–476.

[18] A.S. Nowak, *Optimal strategies in a class of zero-sum ergodic stochastic games*, Math. Methods Oper. Res., 50 (1999), pp. 399–420.

[19] A.S. Nowak, *Some remarks on equilibria in semi-Markov games*, Appl. Math. (Warsaw), 27 (2000), pp. 385–394.

[20] A.S. Nowak and E. Altman, *ε-equilibria for stochastic games with uncountable state space and unbounded cost*, SIAM J. Control Optim., 40 (2002), pp. 1821–1839.

[21] S.M. Ross, *Applied Probability Models with Optimization Applications*, Holden–Day, San Francisco, 1970.

[22] M. Schäl, *Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 32 (1975), pp. 179–196.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 3, pp. 740–797

A SENSITIVITY AND ADJOINT CALCULUS FOR DISCONTINUOUS SOLUTIONS OF HYPERBOLIC CONSERVATION LAWS WITH SOURCE TERMS*

STEFAN ULBRICH†

Abstract. We present a sensitivity and adjoint calculus for the control of entropy solutions of scalar conservation laws with controlled initial data and source term. The sensitivity analysis is based on shift-variations which are the sum of a standard variation and suitable corrections by weighted indicator functions approximating the movement of the shock locations. Based on a first order approximation by shift-variations in L^1 we introduce the concept of shift-differentiability, which is applicable to operators having functions with moving discontinuities as images and implies differentiability for a large class of tracking-type functionals. In the main part of the paper we show that entropy solutions are generically shift-differentiable at almost all times $t > 0$ with respect to the control. Hereby we admit shift-variations for the initial data which allows us to use the shift-differentiability result repeatedly over time slabs. This is useful for the design of optimization methods with time domain decomposition. Our analysis, especially of the shock sensitivity, combines structural results by using generalized characteristics and an adjoint argument. Our adjoint-based shock sensitivity analysis does not require us to restrict the richness of the shock structure a priori and admits shock generation points. The analysis is based on stability results for the adjoint transport equation with discontinuous coefficients satisfying a one-sided Lipschitz condition. As a further main result we derive and justify an adjoint representation for the derivative of a large class of tracking-type functionals.

Key words. optimal control, scalar conservation law, shock sensitivity, adjoint state, Fréchet differentiability

AMS subject classifications. 35L65, 49J50, 49K20, 35R05, 35B37

PII. S0363012900370764

1. Introduction.

This paper is concerned with the development of a sensitivity and adjoint calculus for the optimal control of entropy solutions of scalar conservation laws with source terms: Consider the Cauchy problem for an inhomogeneous conservation law

$$\begin{aligned} &\partial_t y + \partial_x f(y) = g(t,x,y,u_1), \quad (t,x) \in \Omega_T \stackrel{\text{def}}{=}]0,T[\times \mathbb{R}, \\ &y(0,x) = u_0(x), \quad x \in \mathbb{R}, \end{aligned} \tag{1.1}$$

where the flux $f : \mathbb{R} \longrightarrow \mathbb{R}$ is twice continuously differentiable, $u = (u_0, u_1) \in L^\infty(\mathbb{R}) \times L^\infty(\Omega_T)^m \stackrel{\text{def}}{=} U$ is the control, and $g : \Omega_T \times \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}$ satisfies a growth condition

(A1) $g \in L^\infty(\Omega_T; C^{0,1}_{loc}(\mathbb{R} \times \mathbb{R}^m))$ and $\forall\ M_u > 0$ there are $C_1, C_2 > 0$ with

$$g(t,x,y,u_1)\ \operatorname{sgn}(y) \le C_1 + C_2\,|y| \quad \forall\,(t,x,y,u_1) \in \Omega_T \times \mathbb{R} \times [-M_u, M_u]^m.$$

Then one can show (e.g., [18, 25]) that (1.1) admits for all $u \in U$ a unique entropy solution $y = y(u) \in L^\infty(\Omega_T) \cap C([0,T]; L^1_{loc}(\mathbb{R}))$. It is well known that in general weak solutions of (1.1) develop discontinuities after finite time and that uniqueness

*Received by the editors April 13, 2000; accepted for publication (in revised form) January 31, 2002; published electronically July 24, 2002. This work was supported by Deutsche Forschungsgemeinschaft under grant Ul158/2-1 and by CRPC grant CCR–9120008.
http://www.siam.org/journals/sicon/41-3/37076.html
†Zentrum Mathematik, Technische Universität München, 80290 München, Germany (sulbrich@ma.tum.de).

holds only in the class of entropy solutions. We recall that for given $u \in U$ a function $y = y(u) \in L^\infty(\Omega_T)$ is an entropy solution to (1.1) in the sense of Kružkov [18] if it satisfies the entropy inequality

$$\partial_t \eta(y) + \partial_x q(y) \le \eta'(y) g(t, x, y, u_1) \quad \text{in } \mathcal{D}'(\Omega_T)$$

for all convex functions (entropies) $\eta : \mathbb{R} \longrightarrow \mathbb{R}$ with corresponding entropy fluxes $q(y) = \int_0^y \eta'(s)\, f'(s)\, ds$ and the initial condition in the sense

$$\operatorname*{ess\,lim}_{t \to 0+} \|y(t, \cdot) - u_0\|_{1,K}\, d\tau = 0 \qquad \forall\ K \subset\subset \mathbb{R}.$$

The aim of this paper is to develop and justify—without a priori assumptions on the shock structure—a sensitivity calculus for the control-to-state mapping $u \longmapsto y(u)$ that yields in particular the differentiability and a formula for the derivative of objective functionals

$$J(y(u)) = \int_a^b \phi(y(\bar{t}, \cdot; u), y_d)\, dx \tag{1.2}$$

with data $y_d \in BV([a, b])$, $\phi \in C^{1,1}_{loc}(\mathbb{R}^2)$, and $\bar{t} \in\,]0, T]$. Moreover, we will derive an adjoint formula for the gradient of (1.2). The adjoint equation is a transport equation with source term and its coefficient is discontinuous along shock curves, which requires a careful definition of the adjoint state as a reversible solution to ensure uniqueness and stability. These results are useful for the design of gradient-based methods for the solution of control problems of the type

$$\text{(P)} \qquad \min \tilde{J}(y(u), u) \stackrel{\text{def}}{=} J(y(u)) + R(u) \quad \text{subject to} \quad u \in U_{ad}, \quad y(u) \text{ solves (1.1)}.$$

In [25] we have derived results on the existence of optimal solutions for (P) and the convergence of discretized approximations for the multidimensional case. For example, (P) has an optimal solution if U_{ad} is bounded in $L^\infty(\mathbb{R}) \times L^\infty(\Omega_T)^m$ and compact in $L^1_{loc}(\mathbb{R}) \times L^1_{loc}(\Omega_T)^m$ and if $\tilde{J} : C([0, T]; L^r_{loc}(\mathbb{R})) \times (U_{ad} \subset L^r_{loc}(\mathbb{R}) \times L^r_{loc}(\Omega_T)^m) \longrightarrow \mathbb{R}$ is sequentially lower semicontinuous for some $r \in [1, \infty[$. Moreover, for the present one-dimensional case existence results without compactness assumption on U_{ad} were obtained in [25] using compensated compactness.

The state equation (1.1) is a useful model for the study of control problems involving flows with shocks. In particular, it is shown in [10] that the steady flow of an inviscid fluid in a duct governed by the Euler equations can be reduced to determining the velocity y as a steady state solution of (1.1) for $f(y) = y + 2\bar{\gamma} H/y$, $g(x, y, u_1) = u_1 \bar{\gamma}(y - 2H/y)$, where H is the total enthalpy, $\bar{\gamma} = (\gamma - 1)/(\gamma + 1)$ with the gas constant $\gamma > 1$, and the design variable $u_1 = \partial_x A/A$, $A(x)$ being the cross-sectional area of the duct. Moreover, it is noted in [10] that the corresponding time-dependent problem (1.1) captures some essential features of the time-dependent Euler equations and is therefore a suitable model problem for the study of unsteady duct flows with shocks. Since the flow over a transonic airfoil is qualitatively similar to one-dimensional duct flows, the study of the differentiability properties of (1.1)–(1.2) is thus useful to gain insight into the optimal design of airfoils under unsteady flow conditions. In particular, the sensitivity of flows with shocks with respect to time-dependent changes of the geometry is of practical importance for the control of systems with fluid-structure coupling, e.g., the fluttering problem of airfoils.

In this work we give a rigorous sensitivity analysis and adjoint calculus for solutions of (1.1) with shocks and thereby provide an analytical framework for the study

and numerical solution of optimal control problems governed by hyperbolic balance laws (1.1). The following are main features of our approach:

- We derive a sensitivity calculus based on shift-variations that implies the Fréchet differentiability for objective functionals (1.2).
- We give and rigorously justify a gradient representation for objective functionals (1.2) by using an appropriately defined adjoint state.
- The shock structure does not have to be restricted a priori, and shock generation points are allowed.
- We admit nonentropy-admissible initial data and allow shift-variations of the initial data that move shock locations.

The crucial part is the analysis of the shock sensitivities. In our approach we derive the differentiability of a shock position x_s at time $\bar{t}$ by analyzing the smoothness properties of the functional $u \longmapsto \int_{x_s-\varepsilon}^{x_s+\varepsilon} y(\bar{t},x;u)\,dx$ with the help of an adjoint calculus for $\varepsilon \to 0$. The adjoint argument is mainly based on the stability of the adjoint equation with respect to its coefficients. Then the properties of $u \longmapsto x_s(u)$ follow from basic stability properties of the shock that we derive a priori by using the theory of generalized characteristics. An advantage of this method lies in the fact that the shock structure of the solution does not have to be restricted a priori. In particular, shock generation points are allowed. This approach can, at least in a formal manner, also be applied to hyperbolic systems and gives the correct shock sensitivity if the necessary stability properties of the shock and the adjoint equation actually hold.

It can be shown (see, e.g., [25] for the problem at hand) that the mapping $u \in U \longmapsto y(u) \in C([0,T];L^1_{loc}(\mathbb{R}))$ is locally Lipschitz, but very simple examples show that this mapping is in general not directionally differentiable if $y(u)$ contains shocks, even if L^∞ is replaced by C^∞ in the definition of U. This is caused by the fact that a variation of u results in a shift of the shocks, which cannot be approximated appropriately in the linear structure of L^1_{loc}. Consider, for example, the following family of Riemann problems for the inviscid Burgers equation:

$$\partial_t y_\varepsilon + \partial_x \frac{y_\varepsilon^2}{2} = 0, \quad y_\varepsilon(0,x) = u_0(x) + \varepsilon\delta u_0(x) \stackrel{\text{def}}{=} \begin{cases} 1+\varepsilon & \text{if } x \le 0, \\ -1 & \text{if } x > 0. \end{cases}$$

The solution has a shock η_ε emanating from $(0,0)$ with shock speed

$$\dot{\eta}_\varepsilon(t) = \frac{[(y_\varepsilon(t,\eta_\varepsilon(t)))^2/2]}{[y_\varepsilon(t,\eta_\varepsilon(t))]} = \frac{\varepsilon}{2},$$

according to the jump condition, where $[h(t,x)] = h(t,x-) - h(t,x+)$ denotes the jump of $h(t,\cdot)$ across x. Thus, we have $\eta_\varepsilon(t) = \frac{\varepsilon}{2}t$ and the corresponding entropy solution

$$y_\varepsilon = \begin{cases} 1+\varepsilon & \text{if } x \le t\varepsilon/2, \\ -1 & \text{if } x > t\varepsilon/2. \end{cases}$$

Of course, the Lipschitz continuous map $\varepsilon \longmapsto y_\varepsilon(t,\cdot) \in L^1_{loc}(\mathbb{R})$ is not differentiable, since the difference quotient only converges in a weaker topology—for example, weakly in the space $\mathcal{M}_{loc}(\mathbb{R})$ of locally bounded Borel measures. In fact, $\varepsilon \longmapsto y_\varepsilon(t,\cdot) \in \mathcal{M}_{loc}(\mathbb{R})$ is differentiable in the weak topology and we have, for example, at $\varepsilon = 0$

$$\frac{d}{d\varepsilon} y_\varepsilon(t,\cdot)|_{\varepsilon=0}\,\varepsilon = \mathbf{1}_{\{x<\eta_0(t)\}}\,\varepsilon + [y_0(t,\eta_0(t))]\delta_{\eta_0(t)}\,\frac{t}{2}\,\varepsilon, \tag{1.3}$$

where $\mathbf{1}_I$ denotes the indicator function of a set I, i.e., $\mathbf{1}_I(x) = 1$ if $x \in I$, $\mathbf{1}_I(x) = 0$ else, and δ_x denotes the Dirac measure located at x. Hereby, $\frac{t}{2}\,\varepsilon$ is a linear (in this case, exact) approximation of the actual shock shift $\eta_\varepsilon(t) - \eta_0(t)$. Note, however, that a differentiability result in the weak topology of $\mathcal{M}_{loc}(\mathbb{R})$ is not strong enough to derive the differentiability of the functional (1.2) without additional structural information. To get a first order approximation in L^1_{loc}, we have to leave the linear structure of L^1_{loc} in order to allow for an accurate approximation of the shock movement. A natural way to achieve this is to replace the singular (second) part of the measure (1.3) by the function $\operatorname{sgn}(\frac{t}{2}\varepsilon)[y_0(t,\eta_0(t))]\mathbf{1}_{I(\eta_0(t),\eta_0(t)+\frac{t}{2}\varepsilon)}$, where $I(a,b) \overset{\text{def}}{=} [\min\{a,b\}, \max\{a,b\}]$ and $\operatorname{sgn}(\cdot)$ is the sign function. We thereby obtain a first order approximation of $y_\varepsilon(t) - y_0(t)$ in L^1_{loc} by the *shift-variation*

$$S^{\eta_0(t)}_{y_0(t)}\left(\mathbf{1}_{\{x<\eta_0(t)\}}\varepsilon, \frac{t}{2}\varepsilon\right) \overset{\text{def}}{=} \mathbf{1}_{\{x<\eta_0(t)\}}\,\varepsilon + \operatorname{sgn}\left(\frac{t}{2}\varepsilon\right)[y_0(t,\eta_0(t))]\mathbf{1}_{I(\eta_0(t),\eta_0(t)+\frac{t}{2}\varepsilon)}.$$

In this paper we will develop a sensitivity calculus based on shift-variations for the mapping $u \longmapsto y(\bar{t},\cdot;u)$, $\bar{t} \in [0,T]$, defined in (1.1) in the case $f'' \geq m_{f''} > 0$. Let piecewise C^1 initial data $u_0 \in PC^1(\mathbb{R}; x_1, \ldots; x_N)$ and $u_1 \in L^\infty(0,T; C^1(\mathbb{R})^m)$ be given. Using the theory of generalized characteristics [8] we will show that for a given interval I and time $\bar{t} > 0$ the following situation is generic: $y(\bar{t},\cdot;u)$ has on I finitely many shocks at $\bar{x}_1 < \cdots < \bar{x}_K$, the shock locations depend differentiable on u, and the states connected by the shocks vary differentiable in the strong topology of L^1. From this we will deduce that the variation $y(\bar{t},\cdot;u+\delta u) - y(\bar{t},\cdot;u)$ allows a first order approximation by a shift-variation of the form

$$S^{(\bar{x}_k)}_{y(\bar{t},\cdot;u)}(\delta y, \bar{s}) \overset{\text{def}}{=} \delta y + \sum_k [y(\bar{t},\bar{x}_k;u)]\operatorname{sgn}(\bar{s}_k)\mathbf{1}_{I(\bar{x}_k,\bar{x}_k+\bar{s}_k)}, \tag{1.4}$$

where $(\delta y, \bar{s})$ depends linearly on δu, $\bar{x}_k$ are the shock locations,

$$[y(\bar{t},\bar{x}_k;u)] = y(\bar{t},\bar{x}_k-;u) - y(\bar{t},\bar{x}_k+;u)$$

denotes the jump across the shock, $I(\bar{x}_k, \bar{x}_k + \bar{s}_k)$ is the interval enclosed by the arguments, $\mathbf{1}_{I(\bar{x}_k,\bar{x}_k+\bar{s}_k)}$ is its indicator function, and $\bar{s}_k$ is a linear approximation of the shock shift. To mimic the behavior of $y(\cdot;u)$ at later times, it is natural to go one step further and to admit shift-variations already for the initial data. Roughly speaking, we will in particular show that for (u_0,u_1) as above, $W \overset{\text{def}}{=} PC^1(\mathbb{R}; x_1, \ldots, x_N) \times L^\infty(0,T; C^1(\mathbb{R})^m) \times \mathbb{R}^N$ and for a.a. $\bar{t}$ the mapping

$$(w_0, w_1, \sigma) \in W \longmapsto y(\bar{t},\cdot; u_0 + S^{(x_i)}_{u_0}(w_0,\sigma), u_1 + w_1) \in L^1(I) \tag{1.5}$$

is shift differentiable with respect to (w_0,w_1,σ) at 0 in the sense that its variation admits a first order approximation by a shift-variation of the form (1.4), where $(\delta y, \bar{s})$ depends linearly on $\delta w_0, \delta w_1$, and $s = \delta\sigma$. Hereby, δy can be obtained as the trace $\delta Y(\bar{t})$ of a function δY that is the piecewise solution of the linearization of (1.1) *outside* of the shock set, and $\bar{s}_k$ can be obtained by an adjoint formula. We admit shift-variations of the initial data since this allows us to use the shift-differentiability result repeatedly over time slabs. This is helpful for the design of optimization algorithms with time domain decomposition for the solution of (P). By introducing a general concept of shift-differentiability we will be able to derive results on the Fréchet differentiability of tracking-type functionals of the form (1.2) as long as the discontinuities of y_d and $y(\bar{t},\cdot;u)$ do not coincide. If y_d and $y(\bar{t},\cdot;u)$ share discontinuities

we will still obtain directional differentiability. For objective functionals of the form (1.2) we will derive a gradient representation via an adjoint state. The proper definition of the adjoint state requires an extension of the concept of reversible solutions of backward transport equations with discontinuous coefficients introduced in [1] to the case

$$\begin{aligned} &\partial_t p + f'(y)\partial_x p = -g_y(t,x,y,u_1)\,p, \quad (t,x) \in \Omega_{\bar t} \stackrel{\text{def}}{=}]0,\bar t[\times \mathbb{R}, \\ &p(\bar t, x) = p^{\bar t}(x), \quad x \in \mathbb{R}, \end{aligned} \tag{1.6}$$

with linear source term, where $p^{\bar t}$ are suitable end data.

The results of this paper can be straightforwardly extended to identification problems for the flux f, where f is the control. Identification problems of this type are considered by James and Sepúlveda [17]. The differentiability of the objective function for the hyperbolic case was left open. The techniques of the present paper can be used to obtain a sensitivity and adjoint calculus for flux identification as well.

In recent years several results on sensitivities and adjoints for hyperbolic conservation laws were obtained by other authors [1, 2, 3, 4, 5, 6, 12, 13], but most results assume a priori knowledge of the shock structure (usually one shock separating smooth states) or are restricted to the conservative case $g \equiv 0$. The conservative case admits special techniques, since the characteristics are straight lines and the solution can be represented by the integral formula of Lax [19]. Bouchut and James apply in [2] their existence and stability results of [1] for measure-valued duality solutions of linear conservation laws with discontinuous coefficients to derive for the case $g \equiv 0, f'' > 0$ that $u_0 \in L^\infty \longmapsto y(\cdot;u) \in C([0,T];\mathcal{M}_{loc}(\mathbb{R})\text{–w}^*)$ is directionally differentiable at an entropy-admissible u_0 where the space $\mathcal{M}_{loc}(\mathbb{R})$ of local Borel measures is equipped with the usual weak topology. Note that this topology is too weak to derive directly differentiability results for (1.2) without using additional structural information. Godlewski and Raviart study in [12] (see also [13]) the linearized stability of multidimensional hyperbolic systems of conservation laws for perturbations of the initial data of a base solution with a one-dimensional shock. They define measure solutions for the linearized equations with singular part along the shock and construct numerical schemes for the solution of the linearized problem. For the conservative scalar problem with Riemann initial data, it is shown that the linearization coincides with the first order expansion in $C([0,T];\mathcal{M}_{loc}(\mathbb{R})\text{–w}^*)$ of [2]. In this paper we give further justification of this linearization process for more general situations. Bouchut and James develop in [1] existence and stability results for transport equations with discontinuous coefficients satisfying a one-sided Lipschitz condition that will be extended in the present work for the analysis of the adjoint equation (1.6). Previous results on the adjoint equation were obtained in the context of uniqueness results in [7, 16, 20, 21] and of error estimates for approximate solutions in [24]. In [20] adjoint equations for a class of systems of conservation laws are considered. An extension of our approach to systems seems to be possible by building on this work. In [3] a new differential structure on the space BV obtained by horizontal shifts of the points of the graph is introduced, and it is shown that in the case $g \equiv 0, f'' > 0$ the flow $u_0 \in L^\infty \longmapsto y(t,\cdot;u)$ generated by (1.1) is generically differentiable with respect to (w.r.t.) this structure. The analysis uses the integral formula of Lax. Bressan and Marson [4] use generalized tangent vectors to develop a variational calculus for piecewise Lipschitz solutions of systems of conservation laws. Using our notation (1.4), they show that for piecewise Lipschitz initial data u_0^ε, $\varepsilon \geq 0$, such that $u_0^\varepsilon - u_0 = S_{u_0}^{(x_i)}(\varepsilon\delta u_0, \varepsilon s) + o(\varepsilon)$ in L^1_{loc}, the corresponding solutions y_ε

satisfy $y_\varepsilon(\bar{t},\cdot) - y(\bar{t},\cdot) = S^{(\bar{x}_k)}_{y(\bar{t},\cdot)}(\varepsilon\delta y, \varepsilon\bar{s}) + o(\varepsilon)$ in L^1_{loc} if $\bar{t}$ is so small that y_ε remains piecewise Lipschitz on $[0,\bar{t}]$. While this result applies to systems, it considers only directional variations and requires the structural assumption of piecewise Lipschitz solutions, which is not needed in the present paper. Especially the analysis of the shock sensitivity differs significantly from our approach, since in [4] the linearized Rankine–Hugoniot jump condition, together with the linearized state equation, is used to derive an ODE for the shock sensitivity, while we use an adjoint formula which reduces the necessary structural information on the history of the shock as far as possible. Moreover, we develop an adjoint calculus that gives a gradient representation for objective functionals (1.2). Cliff, Heinkenschloss, and Shenoy study in [5, 6] design problems for one-dimensional steady duct flow. By introducing the single shock location as additional state variable and transforming the space variable such that the shock location is fixed, Fréchet differentiability is shown. Optimality conditions are derived and an adjoint-based gradient representation of the objective function is given. Finally, numerical results for the application of an SQP method to the discretized problem are reported.

This paper is organized as follows. In section 2 we introduce the concept of shift-differentiability for operators having discontinuous functions with moving discontinuities as images, which is based on a first order approximation by shift-variations (1.4). Moreover, we will show that the superposition (1.2) of a shift-differentiable operator $u \longmapsto y(\bar{t},\cdot;u)$ and a tracking-type functional is Fréchet differentiable if $y_d, y(\bar{t},\cdot;u)$ do not share discontinuities and is otherwise directionally differentiable. In section 3 we state the main results of the paper. In section 3.1 we state in Theorem 3.2 a shift-differentiability result for entropy solutions of (1.1) w.r.t. the controls, more precisely of the control-state-mapping (1.5), if a nondegeneracy assumption holds for all shocks at the observation time $\bar{t}$. In Theorem 3.4 we give a formula for the corresponding shift-derivative. Moreover, we sketch the main line of the proofs. Theorem 3.8 shows that the required nondegeneracy assumption holds for all shocks at a.a. times $\bar{t} > 0$ if $u_0 \in PC^2$ and $u_1 \in L^\infty(0,T;C^2_{loc})$. In section 3.2 the results of section 3.1 and the general shift-differentiability calculus are combined to obtain in Theorem 3.9 and Corollary 3.10 the differentiability of tracking-type functionals $u \mapsto J(u)$ in (1.2). In section 3.3 we finally state a convenient adjoint-based gradient representation for the derivative of these objective functionals w.r.t. the inner product of L^2; see Theorem 3.11. The proofs of these main results are prepared in sections 4–8 and finally carried out in section 9. Section 4 provides the necessary stability results and collects structural results of [8] provided by the theory of generalized characteristics. We use this to derive basic differentiability results for the solution along generalized characteristics. In section 5 continuity points are analyzed that are not shock generation points. In sections 5.1 and 5.2 we study continuity points in the exterior and interior of rarefaction waves. In section 5.3 continuity points are analyzed that are located on characteristics emanating from points where discontinuities are produced under shift-variations, and section 5.4 studies points on the boundary of rarefaction waves. In section 6 the stability of shocks and the differentiability of the shock location at a time $\bar{t} > 0$ are shown under a nondegeneracy assumption. The proof of the latter is carried out in section 8, since it requires stability results for the adjoint equation, which are provided in section 7. In section 9 we prove the main results already stated in section 3 by combining the results of the previous sections. Conclusions and future work are presented in section 10. The appendix contains a proof of the results in section 7 on the adjoint equation.

Notations. For Lebesgue-measurable $S \subset \mathbb{R}^n$ the norm of the Lebesgue-spaces $L^r(S), 1 \le r \le \infty$, is denoted by $\|\cdot\|_{r,S}$. In the case $S = \mathbb{R}^n$ we write $\|\cdot\|_r$. By $(\cdot,\cdot)_{2,S}$ we denote the inner product on $L^2(S)$. For an interval $I \subset \mathbb{R}$ the space of functions $v \in L^1(I)$ with bounded variation $|v|_{var}$ is denoted by $BV(I)$. For open $S \subset \mathbb{R}^n$ we mean by $C^k(S)$, $k \in \mathbb{N}_0$, the space of functions with continuous, bounded derivatives on S up to order k equipped with the usual norm $\|v\|_{C^k(S)} = \sum_{|\beta|\le k} \|D^\beta v\|_{\infty,S}$. $C^k(S^{cl})$ is the subspace of functions in $C^k(S)$ that admit a continuous extension of the first k derivatives to S^{cl}. Moreover, we write $C(S)$ instead of $C^0(S)$. $C^{k,\alpha}(S^{cl})$, $0 < \alpha \le 1$, is the usual Hölder space. For closed $I \supset [a,b]$, $a < b$, we denote by $PC^k(I; x_1, \dots, x_N)$ the space of piecewise C^k-functions v with possible discontinuities at $a < x_1 < \cdots < x_N < b$; more precisely $v|_{I_i} \in C^k(I_i^{cl})$, $i = 0, \dots, N$, with $I_i =]x_i, x_{i+1}[$, $i = 1, \dots, N-1$, $I_0 = I \cap \{x < x_1\}$, $I_N = I \cap \{x > x_N\}$. It is endowed with the norm $\|v\|_{PC^k(I;x_1,\dots,x_n)} = \sum_{i=0}^N \|v\|_{C^k(I_i^{cl})}$. The indicator function of a set I is denoted by $\mathbf{1}_I$, i.e., $\mathbf{1}_I(x) = 1$ if $x \in I$ and $\mathbf{1}_I(x) = 0$ else.

2. Shift-differentiability. In this section we introduce a concept of shift-differentiability that yields an extension of classical differentiability to operators $u \longmapsto y(u) \in L^1_{loc}(\mathbb{R})$ having functions with moving discontinuities as images. It is based on shift-variations that are the sum of a standard variation and suitably scaled indicator functions approximating the actual shift of discontinuities. The interesting point is that the shift-differentiability of an operator implies under quite general circumstances the Fréchet differentiability of tracking-type functionals analogously to (1.2).

2.1. Shift-variations and shift-differentiability of operators. As motivated in the introduction we define shift-variations as follows.

Definition 2.1. *Let $I = [a,b]$, $a < b$, and let $w \in BV(I)$. Given $a < x_1 < x_2 < \cdots < x_N < b$, we call $(\delta w, s) \in L^1(I) \times \mathbb{R}^N$ a* generalized variation *of w and associate with it the* shift-variation $S_w^{(x_i)}(\delta w, s) \in L^1(\mathbb{R})$ *of w by*

$$S_w^{(x_i)}(\delta w, s)(x) \stackrel{\text{def}}{=} \delta w(x) + \sum_{i=1}^N [w(x_i)]_+ \operatorname{sgn}(s_i) \mathbf{1}_{I(x_i, x_i+s_i)}(x),$$

where $[w(x_i)]_+ \stackrel{\text{def}}{=} \max\{0, w(x_i-) - w(x_i+)\}$ and $I(\alpha,\beta) \stackrel{\text{def}}{=} [\min\{\alpha,\beta\}, \max\{\alpha,\beta\}]$ for $\alpha, \beta \in \mathbb{R}$.

The restriction that only down-jumps are shifted is motivated by the fact that entropy solutions of (1.1) satisfy Oleinik's entropy condition $y(t,x-) \ge y(t,x+)$ for all $x \in \mathbb{R}$ and a.a. $t \in]0,T]$; see section 4.1.

We call an operator shift-differentiable if its actual variation admits a first order approximation in L^1_{loc} by a shift-variation. The following definition states this more precisely.

Definition 2.2. *Let U be a real Banach space and $I = [a,b]$, $a < b$. For an open subset $D \subset U$ let $u \in D \longmapsto y(u) \in L^\infty(\mathbb{R})$ be locally bounded. Moreover, let $\bar u \in D$ with $y(\bar u) \in BV(I)$. We say that y is shift-differentiable at $\bar u$ if there are $a < \bar x_1 < \bar x_2 < \cdots < \bar x_K < b$ and a bounded linear operator $T_s(\bar u) = D_s y(\bar u) \in \mathcal{L}(U, L^r(I) \times \mathbb{R}^K)$, $r \in]1,\infty]$, such that*

$$\lim_{u \to \bar u} \frac{\|y(u) - y(\bar u) - S_{y(\bar u)}^{(\bar x_k)}(T_s(\bar u) \cdot (u - \bar u))\|_{1,I}}{\|u - \bar u\|_U} = 0.$$

We say that y is continuously shift-differentiable at $\bar u$ if y is shift-differentiable in a

neighborhood of $\bar{u}$ and if the corresponding $T_s(u)$, $\bar{x}_k(u)$, $k = 1, \dots, K$, as well as $y(u)(\bar{x}_k(u)\pm)$, are continuous at $\bar{u}$.

2.2. Differentiability after composition with cost functionals. The property of shift-differentiability is strong enough that it implies the Fréchet differentiability of functionals of the form

$$J(y(u)) \stackrel{\text{def}}{=} \int_a^b \phi(y(u)(x), y_d(x))\, dx \tag{2.1}$$

under quite moderate assumptions on ϕ and y_d.

LEMMA 2.3. *Let $u \longmapsto y(u)$ be shift-differentiable in $\bar{u}$ according to Definition 2.2. Moreover, let $y_d \in L^\infty(I)$ be approximately continuous at $\bar{x}_1, \dots, \bar{x}_K$. Then for any function $\phi \in C^{1,1}_{loc}(\mathbb{R}^2)$ the functional $u \in U \longmapsto J(y(u))$ given by (2.1) is Fréchet differentiable at $\bar{u}$, and with $(\delta y, \bar{s}) \stackrel{\text{def}}{=} T_s(\bar{u}) \cdot \delta u$, $\bar{y} \stackrel{\text{def}}{=} y(\bar{u})$ one has*

$$\begin{aligned} d_u J(\bar{y}) \cdot \delta u &= (\phi_y(\bar{y}, y_d), \delta y)_{2,I} \\ &\quad + \sum_{k=1}^{K} \int_0^1 \phi_y(\bar{y}(\bar{x}_k+) + \tau[\bar{y}(\bar{x}_k)], y_d(\bar{x}_k)) d\tau [\bar{y}(\bar{x}_k)]_+ \bar{s}_k. \end{aligned} \tag{2.2}$$

If y is continuously shift-differentiable at $\bar{u}$ and y_d is continuous in a neighborhood of $\bar{x}_1, \dots, \bar{x}_K$, then $u \longmapsto J(y(u))$ is continuously Fréchet differentiable at $\bar{u}$.

If at least one $\bar{x}_k$ is an approximate discontinuity of y_d, then $J(y(u))$ is still directionally differentiable, and with $(\delta y, \bar{s}) = T_s(\bar{u}) \cdot \delta u$ the directional derivative $\delta_u(J(y(\bar{u})); \delta u)$ is given by (2.2) if $y_d(\bar{x}_k)$ is replaced by $y_d(\bar{x}_k + 0 \cdot \operatorname{sgn}(\bar{s}_k))$.

Proof. Obviously, it is sufficient to consider the case $K = 1$. We set $\bar{x} = \bar{x}_1$. Let $B \subset U$ be a bounded neighborhood of 0 such that $\bar{u} + B \subset D$. Then $\|y(\bar{u} + \delta u)\|_\infty \le M_y$ for all $\delta u \in B$ by Definition 2.2. In what follows we will frequently use the abbreviation $u = \bar{u} + \delta u$ for $\delta u \in B$. Moreover, we will write $y, \bar{y}$ instead of $y(u), y(\bar{u})$ and set $(\delta y, \bar{s}) = T_s(\bar{u}) \cdot \delta u$. Finally, we reduce B such that $\bar{x} + \bar{s} \in I$ for $\delta u \in B$.

Since $y(u)$ is shift-differentiable in $\bar{u}$, we have for all $\delta u \in B$

$$\|y - \bar{y}\|_1 \le o(\|\delta u\|_U) + \|S^{\bar{x}}_{\bar{y}}(\delta y, \bar{s})\|_1 \le o(\|\delta u\|_U) + \|\delta y\|_1 + 2M_y|\bar{s}| \le C_1 \|\delta u\|_U. \tag{2.3}$$

Let L be a Lipschitz constant of ϕ_y w.r.t. y on $[-M_y, M_y] \times [-\|y_d\|_\infty, \|y_d\|_\infty]$. Set

$$\bar{\phi}_y(\delta u) \stackrel{\text{def}}{=} \int_0^1 \phi_y(\tau \bar{y} + (1-\tau)y, y_d)\, d\tau.$$

Then we have $\|\bar{\phi}_y(\delta u)\|_\infty \le C_2$ for all $\delta u \in B$ with a constant $C_2 > 0$ and

$$|J(y) - J(\bar{y}) - (\bar{\phi}_y(\delta u), S^{\bar{x}}_{\bar{y}}(\delta y, \bar{s}))_{2,I}| \le \|\bar{\phi}_y(\delta u)\|_\infty \|y - \bar{y} - S^{\bar{x}}_{\bar{y}}(\delta y, \bar{s})\|_1 = o(\|\delta u\|_U).$$

To compare the last term on the left-hand side with (2.2) we note that

$$(\bar{\phi}_y(\delta u), S^{\bar{x}}_{\bar{y}}(\delta y, \bar{s}))_{2,I} = (\bar{\phi}_y(\delta u), \delta y)_{2,I} + [\bar{y}(\bar{x})]_+ \operatorname{sgn}(\bar{s}) (\bar{\phi}_y(\delta u), \mathbf{1}_{I(\bar{x}, \bar{x}+\bar{s})})_2, \tag{2.4}$$

where we use $\bar{x} + \bar{s} \in I$. For the first term we have with $1/r + 1/r' = 1$ the estimate

$$\begin{aligned} \|\bar{\phi}_y(\delta u)\delta y - \phi_y(\bar{y}, y_d)\delta y\|_1 &\le \|\bar{\phi}_y(\delta u) - \phi_y(\bar{y}, y_d)\|_{r'} \|\delta y\|_r \\ &\le L \|y - \bar{y}\|_{r'} \|T_s(\bar{u})\| \, \|\delta u\|_U = o(\|\delta u\|_U), \end{aligned} \tag{2.5}$$

since the first factor tends to zero by interpolation using (2.3) and the L^∞-bound.

If $[\bar y(\bar x)]_+ = 0$, then the second term in (2.4) vanishes, and since $S^{\bar x}_{\bar y}(\delta y, \bar s) = \delta y$ in this case, the proof is complete. Otherwise, we have $[\bar y(\bar x)]_+ = [\bar y(\bar x)]$. To approximate the second term in (2.4) we observe that with $\delta\bar y \stackrel{\text{def}}{=} \max\{\min\{\delta y, 2M_y\}, -2M_y\}$ obviously

$$\|y - \bar y - S^{\bar x}_{\bar y}(\delta\bar y, \bar s)\|_1 = o(\|\delta u\|_U)$$

also holds. Since the function on the left-hand side has a uniform L^∞-bound, we obtain with the local Lipschitz continuity of ϕ_y

$$\left\| \bar\phi_y(\delta u) - \int_0^1 \phi_y(\bar y + \tau S^{\bar x}_{\bar y}(\delta\bar y, \bar s), y_d)\, d\tau \right\|_1 = o(\|\delta u\|_U). \tag{2.6}$$

This yields for the last factor of the last term in (2.4)

$$(\bar\phi_y(\delta u), \mathbf{1}_{I(\bar x, \bar x+\bar s)})_2 = \left(\textstyle\int_0^1 \phi_y(\bar y + \tau S^{\bar x}_{\bar y}(\delta\bar y, \bar s), y_d) d\tau, \mathbf{1}_{I(\bar x, \bar x+\bar s)}\right)_2 + o(\|\delta u\|_U).$$

Finally, we have

$$\begin{aligned} I_1 &\stackrel{\text{def}}{=} \left| \left(\textstyle\int_0^1 \phi_y(\bar y + \tau S^{\bar x}_{\bar y}(\delta\bar y, \bar s), y_d) d\tau, \mathbf{1}_{I(\bar x, \bar x+\bar s)}\right)_2 - |\bar s| \int_0^1 \phi_y(\bar y(\bar x+) + \tau[\bar y(\bar x)], y_d(\bar x)) d\tau \right| \\ &= \left| \int_{\bar x}^{\bar x+\bar s} \int_0^1 (\phi_y(\bar y + \tau[\bar y(\bar x)]\operatorname{sgn}(\bar s) + \tau\delta\bar y, y_d) - \phi_y(\bar y(\bar x\pm) \pm \tau[\bar y(\bar x)], y_d(\bar x)))\, d\tau dx \right|. \end{aligned}$$

Since the arguments of ϕ_y are bounded, we get with a Lipschitz constant L

$$\begin{aligned} I_1 &\le L\left(\|\delta\bar y\|_{1, I(\bar x, \bar x+\bar s)} + |\bar s| \left| \frac{1}{\bar s} \int_{\bar x}^{\bar x+\bar s} (|\bar y - \bar y(\bar x + 0\operatorname{sgn}(\bar s))| + |y_d - y_d(\bar x)|)\, dx \right| \right) \\ &\le L\|\delta y\|_r |\bar s|^{1/r'} + o(|\bar s|) = o(\|\delta u\|_U). \end{aligned} \tag{2.7}$$

Hereby we have used that y_d is approximately continuous in $\bar x$ and $y(\bar u) \in BV(I)$. Now the Fréchet differentiability of $J(y(u))$ and (2.2) follow by combining (2.4)–(2.7).

Now assume that y_d is continuous in a neighborhood of $\bar x_1, \dots, \bar x_K$ and that y is continuously shift-differentiable in $\bar u$. Then J is Fréchet differentiable in a neighborhood D' of $\bar u$ by the previous arguments. By assumption, $u \in D' \longmapsto \bar x_k(u)$ and $u \in D' \longmapsto y(u)(\bar x_k(u)\pm)$ are continuous at $\bar u$. Hence, the operator in the second term of (2.2) with $y, \bar x_k(u)$ instead of $\bar y, \bar x_k$ acting on $\bar s$ is obviously continuous at $\bar u$. Moreover, we have for all $\delta y \in L^r(I)$ and $u \in D'$ with a local Lipschitz constant L of ϕ_y

$$(\phi_y(y(u), y_d) - \phi_y(y(\bar u), y_d), \delta y)_{2,I} \le L\|y(u) - y(\bar u)\|_{r',I} \|\delta y\|_{r,I}.$$

The first factor tends to zero by interpolating (2.3) and the L^∞-bound. Combining the continuity in $(\delta y, \bar s) \in L^r(I) \times \mathbb{R}^K$ with the continuity of $u \longmapsto T_s(u) \in \mathcal{L}(U, L^r(I) \times \mathbb{R}^K)$ at $\bar u$ now yields the continuity of $d_u J(y(u))$ at $\bar u$.

Finally, if y_d has an approximate discontinuity in $\bar x$ and $y_d(\bar x)$ is replaced by $y_d(\bar x + 0\cdot\operatorname{sgn}(\bar s))$, then the directional differentiability can be shown exactly as above by fixing δu and taking $\varepsilon\delta u$ instead of δu. In fact, the only crucial point is the estimate for the resulting expression I_1. As in (2.7) we obtain $I_1 = o(\varepsilon)$. ∎

In the following sections we will analyze the control-to-state mapping $u \longmapsto y(t,\cdot;u)$ implicitly defined by (1.1). As a main result we will show that (1.5) is in general shift-differentiable. Then we obtain immediately the differentiability properties of objective functionals (1.2) by using Lemma 2.3.

3. Statement of the main results. In this section we state the main results of this paper. The proofs will be prepared and carried out in the remaining sections.

3.1. Shift-differentiability of entropy solutions. As outlined in the introduction our first main result is a shift-differentiability result for entropy solutions $y = y(\cdot;u)$ of

$$\begin{aligned} &\partial_t y + \partial_x f(y) = g(t,x,y,u_1), \quad (t,x) \in \Omega_T \stackrel{\text{def}}{=}]0,T[\times \mathbb{R}, \\ &y(0,x) = u_0(x), \quad x \in \mathbb{R}, \end{aligned} \tag{1.1}$$

w.r.t. the control $u = (u_0, u_1)$ if the initial data u_0 are varied by a shift-variation and u_1 by a conventional additive variation. After this result is shown, we obtain immediately differentiability properties of objective functionals (1.2) by using Lemma 2.3.

More precisely, let $u_0 \in PC^1(\mathbb{R}; x_1, \dots, x_N)$, $x_1 < x_2 < \cdots < x_N$, and $u_1 \in L^\infty(0,T; C^1(\mathbb{R})^m)$ be given and fix some $\bar{t} \in]0,T]$. For

$$W \stackrel{\text{def}}{=} PC^1(\mathbb{R}; x_1, \dots, x_N) \times L^\infty(0,T; C^1(\mathbb{R})^m) \times \mathbb{R}^N$$

consider the mapping

$$(w_0, w_1, \sigma) \in W \longmapsto y(\bar{t}, \cdot; u_0 + S_{u_0}^{(x_i)}(w_0, \sigma), u_1 + w_1) \in L^1(a,b) \tag{3.1}$$

for some $a < b$.

REMARK 3.1. *Of course,* (3.1) *includes the simpler case*

$$(\hat{u}_0, \hat{u}_1) \in PC^1(\mathbb{R}; x_1, \dots, x_N) \times L^\infty(0,T; C^1(\mathbb{R})^m) \longmapsto y(\bar{t}, \cdot; \hat{u}_0, \hat{u}_1),$$

where the initial shocks are not shifted.

We will show (see Theorem 3.2 below) that the mapping (3.1) is continuously shift-differentiable in a sufficiently small neighborhood of $(0,0,0)$ if at time $\bar{t}$ a nondegeneracy condition is satisfied at the shock locations. We will moreover show in Theorem 3.8 that this nondegeneracy condition for the shocks at time $\bar{t}$ holds at a.a. $\bar{t}$ if $u_0 \in PC^2(\mathbb{R}; x_1, \dots, x_N)$ and $u_1 \in L^\infty(0,T; C^2_{loc}(\mathbb{R})^m)$. For a precise statement of the theorems we need the following additional assumptions on f and g.

ASSUMPTIONS:

(A2) (A1) holds, f is twice continuously differentiable, $g \in L^\infty(0,T; C^1_{loc}(\mathbb{R} \times \mathbb{R} \times \mathbb{R}^m))$, and g is Lipschitz continuous w.r.t. x.

(A3) $f'' \geq m_{f''} > 0$ for some $m_{f''} > 0$.

(A4) g is affine linear w.r.t. y.

We have the following result on the shift-differentiability of (3.1).

THEOREM 3.2 (shift-differentiability of entropy solutions). *Let* (A2)–(A4) *hold. Let* $u_0 \in PC^1(\mathbb{R}; x_1, \dots, x_N)$ *with* $x_1 < \cdots < x_N$, *and let* $u_1 \in L^\infty(0,T; C^1(\mathbb{R})^m)$. *For* $u = (u_0, u_1)$ *let* $y(u) \in L^\infty(\Omega_T) \cap C([0,T]; L^1_{loc}(\mathbb{R}))$ *be the entropy solution of* (1.1). *Finally, let* $I = [a,b]$, $a < b$, *and* $\bar{t} \in]0,T]$ *such that* $y(\bar{t}, \cdot; u)$ *has no shock generation points on* I *and finitely many nondegenerate shocks on* I *at* $a < \bar{x}_1 < \cdots < \bar{x}_K < b$ *that are not shock interaction points; see Definition* 6.1. *Finally, let*

$$\begin{aligned} W &\stackrel{\text{def}}{=} PC^1(\mathbb{R}; x_1, \dots, x_N) \times L^\infty(0,T; C^1(\mathbb{R})^m) \times \mathbb{R}^N, \\ \|(w_0, w_1, \sigma)\|_W &\stackrel{\text{def}}{=} \|w_0\|_{PC^1(\mathbb{R}; x_1, \dots, x_N)} + \|w_1\|_{L^\infty(0,\bar{t}; C^1(\mathbb{R}))} + \|\sigma\|_2 \end{aligned}$$

and consider the mapping

$$(w_0, w_1, \sigma) \in W \longmapsto y(\bar{t}, \cdot; u_0 + S_{u_0}^{(x_i)}(w_0, \sigma), u_1 + w_1)|_I \in L^1(I). \tag{3.2}$$

Then the following hold:

(i) *If all x_i, $i = 1, \dots, N$, are discontinuities of u_0, i.e., $u_0(x_i-) \neq u_0(x_i+)$, then the mapping* (3.2) *is continuously shift-differentiable on a neighborhood $\{\|(w,\sigma)\|_W < \rho\}$ for $\rho > 0$ sufficiently small, and the shift-derivative $T_s(0) = D_s y(\bar{t}; u)$ satisfies $T_s(0) \in \mathcal{L}(W, PC(I; \bar{x}_1, \dots, \bar{x}_K) \times \mathbb{R}^K)$.*

(ii) *If some x_i, $i = 1, \dots, N$, are continuity points of u_0, then the mapping* (3.2) *is at least shift-differentiable w.r.t. (w_0, w_1, σ) at $(0,0,0)$, and the shift-derivative $T_s(0) = D_s y(\bar{t}; u)$ satisfies $T_s(0) \in \mathcal{L}(W, PC(I; \tilde{x}_1, \dots, \tilde{x}_{\tilde{K}}) \times \mathbb{R}^K)$, where $\tilde{x}_1 < \dots < \tilde{x}_{\tilde{K}}$ contain in addition to $\bar{x}_1, \dots, \bar{x}_K$ all continuity points in I for which the backward characteristic meets the line $t = 0$ in a continuity point x_i, $i = 1, \dots, N$, of u_0.*

REMARK 3.3. *Several comments on the used terminology are in order:*

- *By "characteristics" we mean generalized characteristics in the sense of Dafermos* [8]*; see section* 4.2.
- *The nondegeneracy of a shock at $(\bar{t}, \bar{x}_k)$ is a nondegeneracy condition on the dependence of forward characteristics on their starting point $(0, z)$ that start close to footpoints $(0, z_\mp)$ of the minimal and maximal backward characteristic through $(\bar{t}, \bar{x}_k)$. See Definition* 6.1 *below.*

The next theorem states a formula for the shift-derivative of entropy solutions ensured by Theorem 3.2.

THEOREM 3.4 (formula for the shift-derivative). *Let the assumptions of Theorem* 3.2 *hold. Then the shift derivative $(\delta y^{\bar{t}}, \bar{s}) = T_s(0) \cdot (\delta w_0, \delta w_1, s)$ ensured by Theorem* 3.2 *is given as follows: With $\bar{x}_0 \overset{\text{def}}{=} a$, $\bar{x}_{K+1} \overset{\text{def}}{=} b$, denote by $\xi_k^\mp$ the minimal/maximal backward characteristic through $(\bar{t}, \bar{x}_k)$, $k = 0, \dots, K+1$, by S_k the domain confined by ξ_k^+ and ξ_{k+1}^-, $k = 0, \dots, K$, and by D_k the domain confined by ξ_k^- and ξ_k^+, $k = 1, \dots, K$.*

Let δY on all S_k be the broad solution of the linearized equation

$$\partial_t \delta Y + \partial_x (f'(y)\delta Y) = g_y(t,x,y,u_1)\delta Y + g_{u_1}(t,x,y,u_1)\delta w_1 \tag{3.3}$$

with initial conditions

$$\delta Y(0,\cdot) = \begin{cases} \delta w_0 & \text{on } (S_k^{cl} \cap \{t=0\}) \setminus \{x_i : \ 1 \le i \le N\}, \\ 0 & \text{on } (S_k^{cl} \cap \{t=0\}) \cap \{x_i : \ 1 \le i \le N\}, \end{cases} \tag{3.4}$$

respectively; see Remarks 5.4 *and* 5.9 *below. Moreover, let p^k, $k = 1, \dots, K$, be the reversible solution $p^k = p$ (cf. Definition* 7.5*) of the adjoint equation*

$$\partial_t p + f'(y)\partial_x p = -g_y(t,x,y,u_1)p, \quad p(\bar{t},\cdot) = p^{\bar{t}} \tag{3.5}$$

for the end data (we recall that $[y(\bar{t}, \bar{x}_k; u)] \overset{\text{def}}{=} y(\bar{t}, \bar{x}_k-; u) - y(\bar{t}, \bar{x}_k+; u)$).

$$p^{\bar{t}} = \frac{1}{[y(\bar{t}, \bar{x}_k; u)]} \mathbf{1}_{\{x = \bar{x}_k\}}(\cdot). \tag{3.6}$$

Then $(\delta y^{\bar{t}}, \bar{s}) = T_s(0) \cdot (\delta w_0, \delta w_1, s)$ is given by

$$\delta y^{\bar{t}} = \delta Y(\bar{t}, \cdot), \tag{3.7}$$

$$\bar{s}_k = (p^k g_{u_1}(\cdot, y, u_1), \delta w_1)_{2, \Omega_{\bar{t}}} + (p^k(0,\cdot), \delta w_0)_2 + \sum_{i=1}^{N} p^k(0, x_i)[u_0(x_i)]_+ s_i. \tag{3.8}$$

Hereby, the adjoint states p^k *for the shock sensitivity* $\bar{s}_k$ *of* $\bar{x}_k$ *have the disjoint supports* D_k^{cl} *and coincide on* D_k^{cl} *with the reversible solution of* (3.5) *for the constant end data* $1/[y(\bar{t}, \bar{x}_k; u)]$.

REMARK 3.5.

- *Broad solutions of the linearized equation* (3.4) *are defined as the solution of the characteristic equation along each generalized characteristic. We discuss this issue in detail in section* 5*; see Lemmas* 5.1, 5.6 *and Remarks* 5.4, 5.9. *This leads to a piecewise definition outside the shock set. In general, these patches yield together* not *a weak solution of* (3.4) *on all of* $\Omega_{\bar{t}}$, *since the jump condition across the shocks, which must hold for weak solutions of the linear conservation law* (3.4), *is in general not satisfied. To compensate the incompatibility across shocks, a measure part has to be added on the shock set. Appropriate global weak solutions of* (3.4) *are measure-valued* duality solutions. *For details we refer to the author's habilitation thesis* [26]. *In* [26] *we extend results of Bouchut and James* [1] *in several respects that are of importance for weak sensitivities in optimal control. See also Remark* 3.7.
- *The adjoint equation* (3.5) *is a transport equation with discontinuous coefficient. Since solutions of* (3.5)–(3.6) *are in general not unique, we extend ideas of* [1] *and consider appropriate* reversible solutions *that enjoy uniqueness and stability. We will introduce the necessary facts on reversible solutions in section* 7. *For a detailed study we refer to our recent works* [27, 26].

For the proof of Theorems 3.2 and 3.4 we will use structural properties of entropy solutions obtained by a careful application of Dafermos' theory of generalized characteristics [8] together with an adjoint-based analysis of the shock sensitivities. The necessary results on the structure of entropy solutions will be developed in sections 4–6. The adjoint-based analysis of the shock sensitivities will be carried out in sections 7–8. Using these results, the proof of Theorems 3.2 and 3.4 is then given in section 9.

The main line of the proof is as follows: With the abbreviation

$$u^s(w, \sigma) = (u_0 + S_{u_0}^{(x_i)}(w_0, \sigma), u_1 + w_1)$$

we can write the control-to-state-mapping (3.2) as

$$(w, \sigma) \in W \longmapsto y(\bar{t}, \cdot; u^s(w, \sigma)). \tag{3.9}$$

Under the assumptions of Theorem 3.2 (which is a generic situation; see Theorem 3.8) we show that the mapping (3.9) has the following properties: on any compact subset $J \subset I \setminus \{\bar{x}_1, \dots, \bar{x}_K\}$ the mapping (3.2) is Fréchet differentiable in $(0, 0)$ as a map to $L^r(J)$, $r \in [1, \infty)$, and even to $L^\infty(J \setminus E)$ if E is an open neighborhood of points on the boundary of rarefaction waves (where we include forward characteristics emanating from a point $(0, x_i)$, $1 \le i \le N$, where u_0 is continuous); see section 5. Hereby the derivative is given by $d_{(w,\sigma)} y(\bar{t}, \cdot; u) \cdot (\delta w, s) = \delta y^{\bar{t}}|_J$, where $\delta y^{\bar{t}} = \delta Y(\bar{t}, \cdot)$ with the broad solution δY of the linearized equation (3.3)–(3.4). Moreover, we show in section 6 that there are neighborhoods J_k of any shock position $\bar{x}_k$, such that for a suitable neighborhood $W_\rho \overset{\text{def}}{=} \{\|(w, \sigma)\|_W < \rho\}$ the shock is stable in the sense that

$$(w, \sigma) \in W_\rho \longmapsto y(\bar{t}, \cdot; u^s(w, \sigma)) \in PC^{0,1}(J_k; \bar{x}_k(u^s(w, \sigma)))$$

is bounded, the shock location $\bar{x}_k(u^s(w, \sigma))$ depends Lipschitz continuously on (w, σ), and left and right states are stable on W_ρ in the sense that

$$\|y(\bar{t}, \cdot; u^s(w, \sigma)) - y(\bar{t}, \cdot; u)\|_{\infty, J_k \setminus I(\bar{x}_k, \bar{x}_k(u^s(w,\sigma)))} \le L_k \|(w, \sigma)\|_{w,\sigma};$$

see Lemma 6.2 and Corollary 6.5 in section 6. Moreover, we show by an adjoint argument in sections 6–8 that the shock location depends Fréchet differentiably on (w, σ), and the shock sensitivity is given by $d_{(w,\sigma)}\bar{x}_k(u)\cdot(\delta w, s) = \bar{s}_k$, with $\bar{s}_k$ according to the adjoint formula (3.8); see Lemma 6.4. If this is shown, then it is straightforward to show the shift-differentiability of (3.9); see section 9.

To show these properties of (3.9) we use the theory of generalized characteristics [8] together with an adjoint argument for the analysis of the shock sensitivity as follows: the theory of generalized characteristics yields that entropy solutions have an at most countable number of shock curves and are continuous on the complement; see section 4, in particular Proposition 4.2. In continuity points $(\bar{t}, \bar{x})$ the solution coincides with the solution of the classical characteristic equation, i.e., with the solution $(\zeta, v)(\cdot; z, \omega, u_1)$ of the characteristic equation

$$\begin{aligned} \dot{\zeta}(t) &= f'(v(t)), & \zeta(0; z, \omega, u_1) &= z, \\ \dot{v}(t) &= g(t, \zeta(t), v(t), u_1(t, \zeta(t))), & v(0; z, \omega, u_1) &= \omega, \end{aligned}$$

one has with some $\bar{z} > 0$

$$\bar{x} = \zeta(\bar{t}; \bar{z}, u_0(\bar{z}), u_1), \quad y(\bar{t}, \bar{x}; u) = v(\bar{t}; \bar{z}, u_0(\bar{z}), u_1) \tag{3.10}$$

if $(\bar{t}, \bar{x})$ is not on a rarefaction wave; otherwise, with some $\bar{\omega} \in [u_0(\bar{z}-), u_0(\bar{z}+)]$,

$$\bar{x} = \zeta(\bar{t}; \bar{z}, \bar{\omega}, u_1), \quad y(\bar{t}, \bar{x}; u) = v(\bar{t}; \bar{z}, \bar{\omega}, u_1). \tag{3.11}$$

In shock points the same holds for the left and right states $y(\bar{t}, \bar{x}\mp; u)$, respectively. If $(\bar{t}, \bar{x})$ is a continuity point that is no shock generation point and not located on a rarefaction wave or on a forward characteristic emanating from a point $(0, x_i)$, $1 \le x_i \le N$, where u_0 is continuous, then the representation (3.10) is valid on a neighborhood and the first equation in (3.10) can be solved for $\bar{z}$, yielding with the second by the implicit function theorem the differentiability of

$$(w, \sigma) \in W_\rho \longmapsto y(\cdot; u^s(w, \sigma)) \in L^r(J(\bar{x})) \tag{3.12}$$

with a neighborhood $J(\bar{x})$ of $\bar{x}$ for $r = \infty$ and $\rho > 0$ small enough; see Lemmas 5.1 and 5.5 of section 5.1 (Case C^c). If $(\bar{t}, \bar{x})$ is a continuity point in the interior of a rarefaction wave, the same holds by solving in the first equation of (3.11) for $\bar{\omega}$ and inserting it in the second; see Lemmas 5.6 and 5.10 of section 5.2 (Case R^c). If $(\bar{t}, \bar{x})$ is on the boundary of a rarefaction wave (Case RB^c), then (3.2) is Fréchet differentiable for all $r \in [1, \infty)$; see Lemma 5.12 in section 5.4. If $(\bar{t}, \bar{x})$ is a point on a forward characteristic emanating from a point $(0, x_i)$, $1 \le x_i \le N$, where u_0 is continuous (Case CB^c), then (3.2) is Fréchet differentiable in $(0, 0)$ for all $r \in [1, \infty)$; see Lemma 5.11 in section 5.3. Moreover, by (3.10) or (3.11) and the implicit function theorem, it is not difficult to show that the derivative of (3.9) is given by $d_{(w,\sigma)}y \cdot (\delta w, s) = \delta Y(\bar{t}, \cdot)|_{J(\bar{x})}$, where δY is the broad solution of the linearized equation (3.3)–(3.4); see Lemmas 5.1 and 5.6 and Remarks 5.4 and 5.9 in section 5.

Now consider an observation time $\bar{t}$ such that $y(\bar{t}, \cdot; u)$ has only nondegenerate shocks and no shock generation points on $I = [a, b]$ (see Theorem 3.8) with continuity points a, b. Then one can show that $y(\bar{t}, \cdot; u)$ has on I finitely many shocks at positions $a < \bar{x}_1 < \cdots < \bar{x}_K < b$. The previous considerations show already that (3.9) maps Fréchet differentiable to $L^1(J)$. It remains to analyze the shock positions. Using the theory of generalized characteristics we show in section 6 that the shocks are stable

in the above sense; see Lemma 6.2 and Corollary 6.5. To show the differentiability of the shock position stated in Lemma 6.4 and Corollary 6.5, we apply in section 8 an adjoint argument: Let $\bar{x}_k = \bar{x}_k(u)$ be a shock position and set $\tilde{u} \stackrel{\text{def}}{=} u^s(w,\sigma)$. We use the abbreviations $\tilde{y} = y(\cdot;\tilde{u})$, $y = y(\cdot;u)$, $\Delta y \stackrel{\text{def}}{=} \tilde{y} - y$, and $\delta u \stackrel{\text{def}}{=} \tilde{u} - u$. With an ε-neighborhood $J_\varepsilon =]\hat{x}_-, \hat{x}_+[\stackrel{\text{def}}{=}]\bar{x}_k - \varepsilon, \bar{x}_k + \varepsilon[$ the stability properties of the shock yield

$$\frac{1}{[y(\bar{t},\bar{x}_k)]} \int_{\hat{x}_-}^{\hat{x}_+} \Delta y(\bar{t},x)\,dx = \bar{x}_k(\tilde{u}) - \bar{x}_k + O((\varepsilon + \|(w,\sigma)\|_W)\|(w,\sigma)\|_W). \tag{3.13}$$

Therefore, in order to obtain the shock sensitivity we compute the derivative of the functional on the left-hand side by adjoint-based techniques and then take the limit $\varepsilon \to 0$. Our adjoint approach uses first an averaged adjoint equation to avoid a linearization of the conservation law at shocks and then uses stability properties of the averaged equation to derive the actual adjoint equation: The difference of (1.1) for $\tilde{y}$ and y yields

$$\partial_t \Delta y + \partial_x(\tilde{a}\Delta y) = \tilde{b}\Delta y + g(t,x,y,\tilde{u}_1) - g(t,x,y,u_1) \tag{3.14}$$

with the averaged coefficients

$$\tilde{a}(t,x) \stackrel{\text{def}}{=} \int_0^1 f'(y(t,x) + \tau\Delta y(t,x))\,d\tau, \quad \tilde{b}(t,x) \stackrel{\text{def}}{=} \int_0^1 g_y(\cdot, y(t,x) + \tau\Delta y(t,x), \tilde{u}_1)\,d\tau.$$

Denote by D_ε the domain confined by the backward characteristics through $(\bar{t}, \hat{x}_\mp)$. Multiplying (3.14) by a sufficiently regular test function $\tilde{p}$ with $\tilde{p}(\bar{t},\cdot) \equiv 1/[y(\bar{t},\bar{x}_k)]$ and integrating by parts on D_ε yields

$$\begin{aligned}(\tilde{p}(\bar{t},\cdot), \Delta y(\bar{t},\cdot))_{2,J_\varepsilon} &= (\tilde{p}(0,\cdot), \delta u_0)_{2,D_\varepsilon^c \cap \{t=0\}} + (\partial_t\tilde{p} + \tilde{a}\partial_x\tilde{p} + \tilde{b}\tilde{p}, \Delta y)_{2,D_\varepsilon} \\ &\quad + (\tilde{p}, g_{u_1}(\cdot,y,u_1)\delta u_1)_{2,D_\varepsilon} + o(\|(w,\sigma)\|_W).\end{aligned}$$

The left-hand side is identical to the left-hand side of (3.13). The remainder term contains boundary terms along the characteristics that can be estimated by the stability properties of the solution and a distributed term from the Taylor expansion of g w.r.t. u_1. Now we choose $\tilde{p}$ as the solution of the averaged adjoint equation

$$\begin{aligned}&\partial_t\tilde{p} + \tilde{a}\partial_x\tilde{p} = -\tilde{b}\tilde{p}, \quad \tilde{p}(\bar{t},\cdot) = p^{\bar{t}}, \\ &p^{\bar{t}} \equiv 1/[y(\bar{t},\bar{x}_k;u)].\end{aligned} \tag{3.15}$$

Then we obtain the duality relation

$$\begin{aligned}(p^{\bar{t}}, \Delta y(\bar{t},\cdot))_{2,J_\varepsilon} &= (\tilde{p}(0,\cdot), \delta u_0)_{2,D_\varepsilon^c \cap \{t=0\}} + (\tilde{p}, g_{u_1}(\cdot,y,u_1)\delta u_1)_{2,D_\varepsilon} \\ &\quad + o(\|(w,\sigma)\|_W).\end{aligned} \tag{3.16}$$

We will show in section 7 that a special class of *reversible* solutions of this transport equation with discontinuous coefficient is stable w.r.t. $\tilde{a}, \tilde{b}$, sufficiently regular for the above calculations, and that $\tilde{p}$ converges in an appropriate space to the reversible solution $p^k = p$ of the limit adjoint equation (3.5) for end data $1/[y(\bar{t},\bar{x}_k;u)]$. Using these stability properties, we can take the limit $(\delta w, s) \to 0$ and then $\varepsilon \to 0$ in (3.16) and deduce quite immediately (3.8) (at least formally, but we will justify the limit transition). Hereby, we use the following convenient fact: by prescribing just the point

data (3.6) we can achieve that the adjoint state p^k has exactly the desired support $D = D_0$, and thus the integration domain D is automatically provided by p^k. Note moreover, that the limit transition in $(\tilde{p}(0,\cdot), \delta u_0)_{2,D_\varepsilon^{cl}\cap\{t=0\}}$ leads to the last two terms in (3.8), since $\delta u_0 = \delta w_0 + S_{u_0}^{(x_i)}(\delta w_0, s)$ is a shift-variation.

REMARK 3.6. *We note that the duality relation* (3.16) *with reversible solutions* $\tilde{p}$ *of* (3.15) *holds also if* J_ε *is replaced by any other interval* $J =]\hat{x}_-, \hat{x}_+[$ *with continuity points* $\hat{x}_\mp$ *and if* D_ε *is replaced by the domain* D *confined by the backward characteristics through* $(\bar{t}, \hat{x}_\mp)$*, i.e.,*

$$(p^{\bar{t}}, \Delta y(\bar{t},\cdot))_{2,J} = (\tilde{p}(0,\cdot), \delta u_0)_{2,D^{cl}\cap\{t=0\}} + (\tilde{p}, g_{u_1}(\cdot, y, u_1)\delta u_1)_{2,D} + o(\|(w,\sigma)\|_W). \tag{3.17}$$

This holds first for Lipschitz continuous end data $p^{\bar{t}}$*, but* (3.16) *extends to all end data* $p^{\bar{t}}$ *that are the limit of a boundedly everywhere convergent sequence of Lipschitz-functions; see Theorem* 7.11*. We will use this to get a gradient representation for tracking-type functionals* (1.2)*.*

If D *lies between neighbored shocks, then it contains only continuity points, and the limit transition in* (3.17) *yields the classical duality relation*

$$(p^{\bar{t}}, \delta Y(\bar{t},\cdot))_{2,J} = (p(0,\cdot), \delta u_0)_{2,D^{cl}\cap\{t=0\}} + (p, g_{u_1}(\cdot, y, u_1)\delta u_1)_{2,D}, \tag{3.18}$$

where δY *is the solution of the linearized equations* (3.3)–(3.4) *and* p *is the reversible solution of the adjoint equation* (3.5)*. See also Remark* 3.13*.*

REMARK 3.7. *By requiring the duality relation* (3.18) *for all* $\bar{t} \in (0,T]$ *and end data* $p^{\bar{t}} \in C_c^{0,1}(\mathbb{R})$*, one can define a measure solution (*duality solution*) of the linearized state equation* (3.3) *on all of* Ω_T*. The linearization of the initial shift-variation* δu_0 *leads then to the initial measure* $\delta w_0 + \sum_{i=1}^N [u_0(x_i)]_+ s_i \delta(\cdot - x_i)$*, with* $\delta(\cdot)$ *denoting the Dirac measure at* 0*. By taking the limit in* (3.16) *one can show that this is the correct "weak" linearization of* (1.1) *in the space of measures. This concept was introduced in* [1, 2] *for the case without a source term. For the present case see the author's habilitation thesis* [26]*. As already mentioned, the topology of measures is too weak to obtain directly differentiability results for tracking-functionals* (1.2)*. Therefore we show shift-differentiability.*

Our next result concerns the nondegeneracy assumptions of Theorems 3.2 and 3.4. They require that at the observation time $\bar{t}$ on $I = [a,b]$ there are no shock generation points, no shock interaction points, and only finitely many shocks that are all nondegenerate. In general, this situation holds very likely at a given time $\bar{t}$, since the number of shock interaction points and shock generation points is at most countable and degeneracy of a shock is—in contrast to nondegeneracy—not stable under perturbations. In fact, under slightly stronger regularity assumptions on u_0 and u_1 we are able to show that the situation assumed in Theorems 3.2 and 3.4 holds actually for a.a. $\bar{t} \in (0,T]$.

THEOREM 3.8 (nondegeneracy of shocks holds for a.a. $\bar{t} \in (0,T]$)*. Let* (A2)–(A3) *hold and assume in addition that* f *is* C^3 *and* $g \in L^\infty(0,T; C^2_{loc}(\mathbb{R}\times\mathbb{R}\times\mathbb{R}^m))$*. If* $u_0 \in PC^2(\mathbb{R}; x_1, \dots, x_N)$*,* $u_1 \in L^\infty(0,T; C^2_{loc}(\mathbb{R})^m)$*, then the assumptions of Theorem* 3.2 *hold for a.a.* $\bar{t} \in]0,T]$*.*

The proof again uses generalized characteristics and is carried out in section 9.

3.2. Differentiability of tracking-type functionals. Using Lemma 2.3 and Theorem 3.2 we get the Fréchet differentiability of a large class of tracking-type func-

tionals (1.2), i.e.,

$$J(y(u)) = \int_a^b \phi(y(\bar{t}, \cdot; u), y_d)\, dx. \tag{1.2}$$

We have the following result.

THEOREM 3.9 (differentiability of tracking-type functionals). *Let the assumptions of Theorem* 3.2 *hold and let* $J(y(\bar{t}, \cdot; u))$ *be defined as in* (1.2) *with* $\phi \in C^{1,1}_{loc}(\mathbb{R}^2)$ *and* $y_d \in L^\infty(I)$ *being approximately continuous in* $\bar{x}_1, \dots, \bar{x}_K$. *Then the functional*

$$(w_0, w_1, \sigma) \in W \longmapsto J(y(\bar{t}, \cdot; u_0 + S^{(x_i)}_{u_0}(w_0, \sigma), u_1 + w_1)) \tag{3.19}$$

has the following differentiability properties:

(i) *The functional* (3.19) *is Fréchet differentiable at* $(0,0,0)$. *The application of the derivative to a direction* $(\delta w_0, \delta w_1, s) \in W$ *is given by* (2.2), *i.e.,*

$$\begin{aligned} d_{(w,\sigma)} J(y) \cdot (\delta w_0, \delta w_1, s) &= (\phi_y(y(\bar{t}, \cdot), y_d), \delta y^{\bar{t}})_{2,I} \\ &+ \sum_{k=1}^{K} \int_0^1 \phi_y(y(\bar{t}, \bar{x}_k+) + \tau[y(\bar{t}, \bar{x}_k)], y_d(\bar{x}_k)) d\tau\, [y(\bar{t}, \bar{x}_k)]_+ \bar{s}_k, \end{aligned} \tag{3.20}$$

where $y = y(\cdot; u)$ *is the entropy solution of* (1.1) *and* $\delta y^{\bar{t}}$, $\bar{s}_k$ *are given by* (3.7) *and* (3.8) *in Theorem* 3.4.

(ii) *If in addition* y_d *is continuous in a neighborhood of* $\bar{x}_1, \dots, \bar{x}_K$ *and if* $x_1, \dots, x_N$ *are discontinuities of* u_0, *then the functional* (3.19) *is continuously Fréchet differentiable on* $\{\|(w,\sigma)\|_W < \rho\}$ *for* $\rho > 0$ *sufficiently small.*

Proof. The theorem follows directly from Lemma 2.3 and Theorem 3.2. ∎

By inserting the shock sensitivities (3.8) in the second part of (3.20) and superimposing suitable multiples of the solutions $p^k = p$ of (3.5)–(3.6) we can rewrite (3.20) more conveniently, as follows.

COROLLARY 3.10. *Under the assumptions of Theorem* 3.9 *the formula* (3.20) *can equivalently be rewritten as*

$$\begin{aligned} d_{(w,\sigma)} J(y) \cdot (\delta w_0, \delta w_1, s) &= (\phi_y(y(\bar{t}, \cdot), y_d), \delta y^{\bar{t}})_{2,I} \\ &+ (p g_{u_1}(\cdot, y, u_1), \delta w_1)_{2,\Omega_{\bar{t}}} + (p(0, \cdot), \delta w_0)_2 + \sum_{i=1}^{N} p(0, x_i)[u_0(x_i)]_+ s_i, \end{aligned} \tag{3.21}$$

where $y = y(\cdot; u)$ *is the entropy solution of* (1.1), $\delta y^{\bar{t}}$ *is given by* (3.7) *in Theorem* 3.4, *and* p *is the reversible solution (cf. Definition* 7.5*) of the adjoint equation*

$$\partial_t p + f'(y)\partial_x p = -g_y(t, x, y, u_1)p, \quad p(\bar{t}, \cdot) = p^{\bar{t}} \tag{3.22}$$

with end data

$$p^{\bar{t}}(x) = \begin{cases} \int_0^1 \phi_y(y(\bar{t}, x+) + \tau[y(\bar{t}, x)], y_d(x))\, d\tau & \text{if } x \in \{\bar{x}_1, \dots, \bar{x}_K\}, \\ 0 & \text{else.} \end{cases} \tag{3.23}$$

Proof. The proof follows by inserting (3.8) in (3.20) and from the observation that the function $p = \sum_k \int_0^1 \phi_y(y(\bar{t}, \bar{x}_k+) + \tau[y(\bar{t}, \bar{x}_k)], y_d(\bar{x}_k))\, d\tau\, [y(\bar{t}, x_k)]_+ p^k$ with p^k from (3.5)–(3.6) is the reversible solution of (3.22) for data (3.23). ∎

3.3. Adjoint calculus for tracking-type functionals. Although the shift-differential is very useful for analytical purposes, the derivative of tracking-type functionals (1.2), (3.19) can more conveniently be computed via an adjoint-based formula, which yields also a gradient representation for (3.19) w.r.t. the scalar product of $L^2(\mathbb{R}) \times L^2(\Omega_{\bar{t}})^m \times \mathbb{R}^N$.

In fact, by using a duality relation (3.18) (see Remark 3.6) between the solution δY of the variational equation (3.3), and the reversible solution p of the adjoint equation (3.22) for end data $p^{\bar{t}} = \phi_y(y(\bar{t},\cdot), y_d)\, \mathbf{1}_{\{x \notin \{\bar{x}_1,\dots,\bar{x}_K\}\}}$, we will be able to obtain also an adjoint-based formula for the first term of (3.21). This leads to the following result.

THEOREM 3.11 (gradient representation for tracking-type functionals). *Let the assumptions of Theorem* 3.9 *hold and let y_d be PC^1. Then*

$$\begin{aligned} d_{(w,\sigma)}J(y)\cdot(\delta w_0, \delta w_1, s) &= (p g_{u_1}(\cdot, y, u_1), \delta w_1)_{2,\Omega_{\bar{t}}} \\ &\quad + (p(0,\cdot), \delta w_0)_2 + \sum_{i=1}^{N} p(0,x_i)[u_0(x_i)]_+ s_i, \end{aligned} \tag{3.24}$$

where p is the reversible solution (cf. Definition 7.5*) of the adjoint equation*

$$\partial_t p + f'(y)\partial_x p = -g_y(t,x,y,u_1)p, \quad p(\bar{t},\cdot) = p^{\bar{t}}, \tag{3.25}$$

$$p^{\bar{t}}(x) = \bar{\phi}_y(x) \stackrel{\text{def}}{=} \begin{cases} \int_0^1 \phi_y(y(\bar{t},x+) + \tau[y(\bar{t},x)], y_d(x))\, d\tau, & x \in I, \\ 0 & \text{else.} \end{cases} \tag{3.26}$$

Thus, the gradient representation of J w.r.t. the scalar product of the Hilbert space $L^2(\mathbb{R}) \times L^2(\Omega_{\bar{t}})^m \times \mathbb{R}^N$ is given by

$$\nabla_{(w,\sigma)}J(y) = \begin{pmatrix} p(0,\cdot) \\ p\, g_{u_1}(\cdot, y, u_1) \\ (p(0,x_i)[u_0(x_i)]_+)_{1\le i\le N} \end{pmatrix}. \tag{3.27}$$

With the domains S_k, D_k from Theorem 3.2 *holds $p|_{D_k} \in C^{0,1}(D_k \cap \{t > \tau\})$ for all $\tau > 0$. Moreover $p|_{S_k}$ is piecewise $C^{0,1}$ on $S_k \cap \{t > \tau\}$ for all $\tau > 0$ with discontinuities along the backward characteristics emanating from discontinuities of y_d. This remains true for $\tau = 0$ on all D_k, S_k that contain no rarefaction wave.*

The proof of this theorem is given in section 9.

REMARK 3.12.

- *The restriction to source terms g that are affine linear w.r.t. y can be dropped without major changes in our analysis if the stability results of Theorem* 7.10 *below for the adjoint equation* (1.6) *can be extended to a discontinuous zeroth order coefficient $g_y(t,x,y,u_1)$. See Remark* 8.1.
- *By Theorem* 3.8 *nondegeneracy for all shocks at time $\bar{t}$ holds for a.a. $\bar{t}$ under slightly stronger regularity assumptions on (u_0, u_1). In this sense, the case of nondegenerate shocks at the observation time $\bar{t}$ is a generic situation.*

REMARK 3.13. *Once the shift-differentiability of* (3.9) *is shown, one could also use the duality relation* (3.17) *with $\tilde{p}$ being a reversible solution of the averaged adjoint equation* (3.15) *with end data* (3.26) *and then take the limit to deduce the gradient formula* (3.24). *For this approach and a detailed analysis of the adjoint equation see the follow-up paper* [27]. *In the present paper it is most straightforward to use the duality relation* (3.18) *to deduce* (3.24) *from* (3.21).

In the next section we start by collecting several structural results of Dafermos [8] that will form the basis of our analysis.

4. Stability and structure of entropy solutions. Our aim is to derive a shift-differentiability result for entropy solutions $y = y(\cdot; u)$ of

$$\begin{aligned} &\partial_t y + \partial_x f(y) = g(t,x,y,u_1), \quad (t,x) \in \Omega_T \stackrel{\text{def}}{=}]0,T[\times \mathbb{R}, \\ &y(0,x) = u_0(x), \quad x \in \mathbb{R}, \end{aligned} \tag{1.1}$$

w.r.t. the control $u = (u_0, u_1)$, where we consider shift-variations of the initial data u_0 and conventional variations of u_1. As explained in section 3.1, we fix $u_0 \in PC^1(\mathbb{R}; x_1, \dots, x_N)$, $x_1 < x_2 < \cdots < x_N$, $u_1 \in L^\infty(0,T; C^1(\mathbb{R})^m)$, an observation time $\bar{t} \in (0,T]$ and consider the mapping

$$(w_0, w_1, \sigma) \in W \longmapsto y(\bar{t}, \cdot; u_0 + S_{u_0}^{(x_i)}(w_0, \sigma), u_1 + w_1) \in L^1(a,b)$$

for some $a < b$, where $W = PC^1(\mathbb{R}; x_1, \dots, x_N) \times L^\infty(0,T; C^1(\mathbb{R})^m) \times \mathbb{R}^N$.

Our analysis is based on the theory of generalized characteristics introduced in [8] to obtain structural information on the solution combined with a duality argument using the adjoint equation to (1.1). This approach has the advantage that we need not restrict a priori the class of considered entropy solutions. Thus, the results apply to solutions with very complicated structure.

In order to ensure the existence of essentially bounded entropy solutions and to allow the application of the theory of generalized characteristics, we make the following assumptions that we have already introduced in section 3.

ASSUMPTIONS:

(A2) (A1) holds, f is twice continuously differentiable, $g \in L^\infty(0,T; C^1_{loc}(\mathbb{R} \times \mathbb{R} \times \mathbb{R}^m))$, and g is Lipschitz continuous w.r.t. x.

(A3) $f'' \geq m_{f''} > 0$ for some $m_{f''} > 0$.

In the next subsection we summarize existence and stability results for entropy solutions of (1.1). In section 4.2 we collect from [8] the necessary results on the structure of solutions provided by the theory generalized characteristics. These ingredients will be used in sections 5–8 to prepare the proof in section 9 of the shift-differentiability result and the other main results that we have stated in section 3.

4.1. Basic properties of entropy solutions. We recall the following existence, uniqueness, and stability properties of the state equation (1.1); see, e.g., [25] and [21].

THEOREM 4.1. *Let* (A1) *hold. Then for all* $u = (u_0, u_1) \in L^\infty(\mathbb{R}) \times L^\infty(\Omega_T)^m \stackrel{\text{def}}{=} U$ *there exists a unique entropy solution* $y = y(u) \in L^\infty(\Omega_T)$. *After modification on a set of measure zero, one has* $y \in C([0,T]; L^1(-R,R))$ *for all* $R > 0$. *Let* $M_u > 0$ *and* $U_{ad} = \{u \in U : \|u_0\|_\infty \leq M_u, \|u_1\|_\infty \leq M_u\}$. *Then there are* $M_y > 0$ *and* $L_y > 0$ *such that for all* $u, \hat{u} \in U_{ad}$ *the corresponding solutions* $y, \hat{y}$ *satisfy*

(i) $\|y(t,\cdot)\|_\infty \leq M_y$,

(ii) $\|y(t,\cdot) - \hat{y}(t,\cdot)\|_{1,[a,b]} \leq L_y(\|u_0 - \hat{u}_0\|_{1,I_t} + \|u_1 - \hat{u}_1\|_{1,[0,t]\times I_t})$

for all $t \in [0,T]$, $a < b$, *where* $I_t = [a - tM_{f'}, b + tM_{f'}]$ *with* $M_{f'} = \max_{|y| \leq M_y} |f'(y)|$.

Moreover, let (A2), (A3) *hold and set* $\hat{U}_{ad} = \{u \in U_{ad} : \|u_1\|_{L^\infty(0,T;C^1)} \leq M_u\}$. *Then there exists a constant* $M_{cr} > 0$ *such that for all* $u \in \hat{U}_{ad}$ *and all* $t \in]0,T]$ *with* $E = M_{cr} m_{f''}$ *Oleinik's entropy condition*

$$\partial_x y(t,\cdot) \leq \frac{1}{(1 - e^{-Et})/M_{cr} + e^{-Et}/M} \tag{4.1}$$

holds in the sense of distributions whenever $M \in [M_{cr}, \infty]$ *is such that* $\partial_x u_0 \leq M$ *in the sense of distributions. In particular,* $y(t,\cdot) \in BV_{loc}(\mathbb{R})$ *for all* $t \in]0,T]$.

Proof. For the first part, see, e.g., [25]. The Oleinik entropy condition in this form can be deduced by a straightforward extension of the proof in [23] to the inhomogeneous case; see also [21]. ∎

Since for all $u \in U_{ad}$ the corresponding solutions $y = y(u)$ of (1.1) satisfy $\|y\|_\infty \leq M_y$, we may modify g for $|y| > M_y$ in such a way that g satisfies instead of the weaker growth condition in (A1) a global Lipschitz condition w.r.t. y. To study $y(u)$ for $u \in U_{ad}$ we may therefore assume that instead of (A2) the following holds.

ASSUMPTION:

(A2′) (A2) is satisfied and g is globally Lipschitz w.r.t. y.

By this modification of g we can achieve also that the backward solutions of the characteristic equations associated with (1.1) remain bounded for all end data and not only for end data obtained from bounded forward solutions. Since we will always deal with bounded entropy solutions, we will assume without restriction that (A2′) holds.

4.2. Generalized characteristics and the structure of solutions. We assume throughout this section that (A2), (A3) hold and consider controls $u \in \hat{U}_{ad}$ with $\hat{U}_{ad}$ from Theorem 4.1. Hence, we may assume without restriction that (A2′) is also satisfied.

Let $u = (u_0, u_1) \in \hat{U}_{ad}$ be given. Then by Theorem 4.1 (1.1) has a unique entropy solution $y = y(u) \in L^\infty(\Omega_T) \cap C([0,T]; L^1_{loc}(\mathbb{R}))$ with $y(t,\cdot) \in BV_{loc}(\mathbb{R})$, $t \in]0,T]$ and $\|y\|_\infty \leq M_y$. Hence, y admits left and right limits w.r.t. x for all $t \in]0,T]$. Moreover, by (4.1) each discontinuity is admissible, i.e.,

$$y(t,x-) \geq y(t,x+) \quad \forall\, t \in]0,T] \text{ and } \forall\, x \in \mathbb{R}.$$

These properties allow the application of the theory of generalized characteristics [8]. For notational convenience we consider the representative for y with $y(t,x) = y(t,x-)$. A Lipschitz continuous curve $x = \xi(t)$ defined on $t \in [a,b] \subset [0,T]$ is a (generalized) characteristic if the differential inclusion holds:

$$\dot{\xi}(t) \in [f'(y(t,\xi(t)+)), f'(y(t,\xi(t)-))] \quad \text{a.e. on } [a,b].$$

The local existence of a characteristic through any $(\bar{t},\bar{x}) \in \Omega_T$ follows from [9]; see also [8]. Assumption (A1) ensures that $\|y\|_\infty \leq M_y$. Thus, characteristics cannot escape and exist on the whole interval $[0,T]$. Hence, we can always set $[a,b] = [0,T]$. Since y is a weak solution of (1.1), it can be deduced [8] that the following actually holds:

$$\dot{\xi}(t) = \begin{cases} f'(y(t,\xi(t))) & \text{if } y(t,\xi(t)-) = y(t,\xi(t)+), \\ \dfrac{[f(y(t,\xi(t)))]}{[y(t,\xi(t))]} & \text{if } y(t,\xi(t)-) \neq y(t,\xi(t)+) \end{cases} \qquad \text{a.e. on [0,T].}$$

A characteristic is called genuine on $[a,b]$ if $y(t,\xi(t)+) = y(t,\xi(t)-)$ for a.a. $t \in [a,b]$.

The study of generalized characteristics in [8] together with the a priori bound $\|y\|_\infty \leq M_y$ yields the following structure of y.

PROPOSITION 4.2 (structure of entropy solutions [8]). *Let $u = (u_0, u_1) \in \hat{U}_{ad}$ with $u_0 \in BV_{loc}(\mathbb{R})$ and denote by $y = y(u)$ the representative of the entropy solution of (1.1) with $y(t,x) = y(t,x-)$. Then the following holds:*

For each fixed $(\bar{t},\bar{x}) \in \Omega_T$ the one-sided limits $y(\bar{t},\bar{x}\pm)$ exist and satisfy the entropy condition $y(\bar{t},\bar{x}-) \geq y(\bar{t},\bar{x}+)$. Moreover, the minimal and maximal backward characteristics $\xi_\mp(t)$ through $(\bar{t},\bar{x})$ are genuine, i.e., $y(t,\xi_\mp(t)-) = y(t,\xi_\mp(t)+)$,

$t \in]0, \bar{t}[$. *Moreover, for any genuine characteristic* $\xi(t)$ *on* $[0, \bar{t}]$ *one has (with our convention for the choice of* y*)*

(4.2)
$$\xi(t) = \zeta(t),\ t \in [0, \bar{t}], \quad y(t, \xi(t)) = v(t),\ t \in]0, \bar{t}[, \quad u_0(\xi(0)-) \le v(0) \le u_0(\xi(0)+),$$
$$y(\bar{t}, \xi(\bar{t})-) \ge v(\bar{t}) \ge y(\bar{t}, \xi(\bar{t})+),$$

where (ζ, v) *solves the classical characteristic equations*

$$\begin{aligned} \dot{\zeta}(t) &= f'(v(t)), \\ \dot{v}(t) &= g(t, \zeta(t), v(t), u_1(t, \zeta(t))). \end{aligned} \tag{4.3}$$

In particular, two different genuine characteristics may intersect only at their end points. Finally, if ξ *is the minimal characteristic* ξ_- *or the maximal characteristic* ξ_+ *through* $(\bar{t}, \bar{x})$*, then* (4.2) *holds with the solution* (ζ, v) *of* (4.3) *for the initial values*

$$(\zeta(\bar{t}), v(\bar{t})) = \begin{cases} (\bar{x}, y(\bar{t}, \bar{x}-)) & \text{if } \xi = \xi_-, \\ (\bar{x}, y(\bar{t}, \bar{x}+)) & \text{if } \xi = \xi_+. \end{cases} \tag{4.4}$$

We remark that for any minimal or maximal backward characteristic $\xi_\mp(t)$ the point $z = \xi_\mp(0)$ is a continuity point of u_0 or a nonentropy-admissible discontinuity, i.e., $u_0(z-) < u_0(z+)$; cf. (4.2).

Denote for $u_1 \in L^\infty(0, T; C^1(\mathbb{R})^m)$ and $z, w \in \mathbb{R}$ by $(\zeta(\cdot; z, w, u_1), v(\cdot; z, w, u_1))$ the solution of (4.3) with

$$\zeta(0; z, w, u_1) = z, \quad v(0; z, w, u_1) = w. \tag{4.5}$$

Let $(\bar{t}, \bar{x}) \in \Omega_T$ be a point of continuity of y w.r.t. x. Then y is by [8] continuous at $(\bar{t}, \bar{x})$ and the backward characteristic ξ through $(\bar{t}, \bar{x})$ is unique and genuine. Moreover, $\bar{z} = \xi(0)$ is a continuity point of u_0 or $u_0(\bar{z}-) < u_0(\bar{z}+)$. In the first case we have

$$\bar{x} = \zeta(\bar{t}; \bar{z}, u_0(\bar{z}), u_1), \tag{4.6}$$
$$y(\bar{t}, \bar{x}) = v(\bar{t}; \bar{z}, u_0(\bar{z}), u_1). \tag{4.7}$$

In the second case $(\bar{t}, \bar{x})$ lies on a rarefaction wave, i.e.,

$$\bar{x} = \zeta(\bar{t}; \bar{z}, \bar{w}, u_1), \tag{4.8}$$
$$y(\bar{t}, \bar{x}) = v(\bar{t}; \bar{z}, \bar{w}, u_1) \tag{4.9}$$

with some $\bar{w} \in [u_0(\bar{z}-), u_0(\bar{z}+)]$.

To study the smoothness of $(t, x, \hat{u}) \longmapsto y(t, x; \hat{u})$ in a suitable neighborhood of $(\bar{t}, \bar{x}, u)$ we will show that (4.6) can locally be solved for $\bar{z}$ (or (4.8) for $\bar{w}$) as long as $(\bar{t}, \bar{x})$ is not a shock generation point yielding with (4.7) (or (4.9)) an expression for y.

We begin by stating smoothness properties of the functions on the right-hand side in (4.8) and (4.9). The following result on ordinary differential equations will be useful.

PROPOSITION 4.3. *Let* $h(t, X, U) \in L^\infty(0, T; C^1_{loc}(\mathbb{R}^n \times \mathbb{R}^m)^n)$ *and Lipschitz w.r.t.* X*. Set* $\bar{H} \stackrel{\text{def}}{=} \{(Z, U) \in \mathbb{R}^n \times L^\infty(0, T; C^1(\mathbb{R}^n)^m) \ : \ \|U\|_{L^\infty(0,T;C^1(\mathbb{R}^n))} < M\}$ *for some* $M > 0$*. Then for all* $(Z, U) \in \bar{H}$ *there exists a unique solution* $X = X(\cdot; Z, U) \in C^{0,1}([0, T])^n$ *of*

$$\dot{X}(t) = h(t, X(t), U(t, X(t))), \quad X(0) = Z,$$

and the mapping $(Z,U) \in (\bar{H}, \|\cdot\|_{\mathbb{R}^n \times L^2(0,T;C^i(\mathbb{R}^n)^m)}) \longmapsto X(\cdot;Z,U) \in C([0,T])^n$ *is Lipschitz continuous for* $i = 0$ *and continuously Fréchet differentiable for* $i = 1$. *Moreover,* $(t,Z) \in [0,T] \times \mathbb{R}^n \mapsto X(t;Z,U)$ *is Lipschitz continuous for* $U \in L^\infty(0,T;C^1(\mathbb{R}^n)^m)$. *Finally, for any closed set* $S \subset \mathbb{R}^r$ *the mapping*

$$(Z,U) \in C(S) \times L^\infty(0,T;C^1(\mathbb{R}^n)^m) \longmapsto X(\cdot;Z(\cdot),U) \in C([0,T] \times S)^n$$

is continuously Fréchet differentiable.

The proof is standard and can, for example, be obtained by a refinement of the analysis in the appendix of [22] using the fact that the Nemyckii operator

$$(Z,U) \in L^\infty(0,T) \times (L^\infty(0,T;C^1(\mathbb{R})), \|\cdot\|_{L^2(0,T;C^1(\mathbb{R}))}) \longmapsto U(\cdot,Z(\cdot)) \in L^2(0,T)$$

is continuously Fréchet differentiable. Since the remainder term in the first order expansion of $(Z,U) \longmapsto X(\cdot;Z,U)$ can be estimated uniformly for all Z in a compact set, the last assertion follows immediately. We omit the technical details.

Now we obtain the following properties for solutions of (4.3).

LEMMA 4.4. *Let* (A2) *hold and denote for* $(z,w,u_1) \in \mathbb{R} \times \mathbb{R} \times L^\infty(0,T;C^1(\mathbb{R})^m)$ *by* $(\zeta,v)(\cdot;z,w,u_1)$ *the solution of* (4.3) *for initial data* (4.5). *Let* $M_w, M_u > 0$ *be given and*

$$H_i \stackrel{\text{def}}{=} \mathbb{R}^2 \times L^2(0,T;C^i(\mathbb{R})^m), \quad i = 0,1,$$
$$H \stackrel{\text{def}}{=} \{(z,w,u_1) \in H_1 \ : \ |w| < M_w, \ \|u_1\|_{L^\infty(0,T;C^1(\mathbb{R}))} < M_u\}.$$

Then the mapping

$$(z,w,u_1) \in (H, \|\cdot\|_{H_i}) \longmapsto (\zeta,v)(\cdot;z,w,u_1) \in C([0,T])^2 \tag{4.10}$$

is Lipschitz continuous for $i = 0$ *and continuously Fréchet differentiable for* $i = 1$, *and on* H *the right-hand side is uniformly Lipschitz w.r.t. t. Moreover, with* $\delta\nu = (\delta z, \delta w, \delta u_1)$ *one has*

$$d_{(z,w,u_1)}(\zeta,v)(\cdot;z,w,u_1) \cdot \delta\nu = (\delta\zeta,\delta v)(\cdot;z,w,u_1;\delta\nu), \tag{4.11}$$

where $(\delta\zeta,\delta v) = (\delta\zeta,\delta v)(\cdot;z,w,u_1;\delta z,\delta w,\delta u_1)$ *solves the linearized equation*

$$\begin{aligned} \dot{\delta\zeta} &= f''(v)\,\delta v, \\ \dot{\delta v} &= g_x(:)\,\delta\zeta + g_y(:)\,\delta v + g_{u_1}(:)\,(\partial_x u_1(t,\zeta)\delta\zeta + \delta u_1(t,\zeta)), \\ \delta\zeta(0) &= \delta z, \quad \delta v(0) = \delta w, \end{aligned} \tag{4.12}$$

with $(:) = (t,\zeta,v,u_1(t,\zeta))$. *The Fréchet derivative* (4.11) *can be continuously extended to* $\mathcal{L}(H_0, C([0,T]))$ *uniformly bounded on bounded subsets of* H. *Finally, for any closed* $S \subset \Omega_T^{cl}$ *and any bounded interval* J *the mapping*

$$\begin{aligned} (z,u_0,u_1) \in C(S;J) \times C^1(J) \times L^\infty(0,T;C^1(\mathbb{R})^m) \\ \longmapsto (\zeta,v)(\cdot_t;z(\cdot),u_0(z(\cdot)),u_1) \in C(S)^2 \end{aligned} \tag{4.13}$$

is continuously Fréchet differentiable, where $\cdot_t$ *denotes the projection* $(t,x) \longmapsto t$. *If* (A2′) *holds, then the same statements are true for backward solutions of* (4.3).

Proof. We can apply Proposition 4.3 if an a priori bound for v in (4.3) is known, since this ensures with (A2) that the right-hand side in (4.3) admits a Lipschitz

constant w.r.t. (ζ, v, u_1) for all (ζ, v, u_1) of interest. To derive such a bound we use (A1) and get constants $C_1, C_2 > 0$ with $\frac{d}{dt}|v(t)| \le C_1 + C_2|v(t)|$, $t \in]0,T[$. Since $|v(0)| = |w| \le M_w$, the Gronwall lemma yields

$$|v(t)| \le (M_w + C_1 T)e^{C_2 T}, \quad t \in [0,T].$$

Now the proof can be obtained by using Proposition 4.3.

The fact that (4.11) can be continuously extended to $\mathcal{L}(H_0, C([0,T]))$ is obvious from the properties of (4.12). The differentiability of (4.13) follows from Proposition 4.3, since the Nemyckii operator

$$(z, u_0) \in C(S;J) \times C^1(J) \longmapsto u_0(z(\cdot)) \in C(S)$$

is continuously Fréchet differentiable.

If (A2′) is satisfied, then we have the stronger growth estimate $|g(t,x,y,u_1)| \le C_1 + C_2|y|$ and the above arguments can be applied to backward solutions. □

To ensure backward stability of solutions to (4.3) for all end data—not only the relevant ones obtained from forward solutions—it will be convenient to assume (A2′) instead of (A2), which may be done without restriction by our considerations at the end of section 4.1.

4.3. Classification of continuity points. For the following analysis of the structure of entropy solutions we assume that $u_0 \in PC^1(\mathbb{R}; x_1, \dots, x_N)$, $x_1 < \cdots < x_N$, and $u_1 \in L^\infty(0,T; C^1(\mathbb{R})^m)$.

For further reference we distinguish several cases for continuity points $(\bar t, \bar x)$. We denote the genuine backward characteristic through $(\bar t, \bar x)$ by ξ and set $\bar z = \xi(0)$. We now consider the following cases.

Case C. Let $(\bar t, \bar x)$ be a continuity point of $y = y(u)$ such that $\bar z \ne x_i$, $i = 1, \dots, N$. Since $u_0 \in PC^1(\mathbb{R}; x_1, \dots, x_N)$, there is an interval J containing $\bar z$ such that $u_0|_J \in C^1(J)$. Now

(4.14)
$$(z, u_0, u_1) \in J \times C^1(J) \times L^\infty(0,T; C^1(\mathbb{R})^m) \longmapsto (\zeta, v)(\cdot; z, u_0(z), u_1) \in C([0,T])^2$$

is continuously Fréchet differentiable by Lemma 4.4. Hence, $\frac{d}{dz}\zeta(t; z, u_0(z), u_1)$ exists and is continuous on $(t,z) \in [0,T] \times J$. Since genuine characteristics may intersect only at their end points and contain only continuity points, it is obvious that [8]

$$\frac{d}{dz}\zeta(t; z, u_0(z), u_1)|_{z=\bar z} \ge 0, \quad 0 \le t \le \bar t. \tag{4.15}$$

Moreover, if $(\bar t, \bar x)$ is not an element of the shock set, i.e., if the unique forward characteristic is genuine until some $\bar t + \tau$, $\tau > 0$, then there is $\beta > 0$ with

$$\frac{d}{dz}\zeta(t; z, u_0(z), u_1)|_{z=\bar z} \ge \beta > 0, \quad 0 \le t \le \bar t. \tag{4.16}$$

In fact, we have

$$\frac{d}{dz}\zeta(t; z, u_0(z), u_1)|_{z=\bar z} = \delta\zeta(t; \bar z, u_0(\bar z), u_1; 1, u_0'(\bar z), 0) \tag{4.17}$$

with $\delta\zeta$ given by (4.12). Moreover, all points on the genuine backward characteristic $\zeta(t) = \zeta(t; \bar z, u_0(\bar z), u_1)$ are continuity points. Assume that $\delta\zeta(\tilde t) = 0$ for the right-hand side of (4.17) holds at some $\tilde t \in [0, \bar t]$. Then $\delta\dot\zeta(\tilde t) \ne 0$, since otherwise, by the

first line in (4.12), $\delta v(\tilde{t}) = 0$. This is impossible because the unique backward solution of (4.12) would vanish in contradiction to the initial values. From (4.15) we thus have $\delta\dot{\zeta}(\tilde{t}) < 0$, and therefore $\frac{d}{dz}\zeta(t; z, u_0(z), u_1)|_{z=\bar{z}} < 0$ for small $t > \tilde{t}$. Hence, the unique forward characteristic through $(\tilde{t}, \zeta(\tilde{t}))$ cannot be genuine, since the unique candidate $\zeta(t) = \zeta(t; \bar{z}, u_0(\bar{z}), u_1)$ is not admissible by (4.15). Since the left-hand side of (4.16) is continuous on $[0, \bar{t}]$, (4.16) must hold for some $\beta > 0$.

Case CB. $(\bar{t}, \bar{x})$ is a continuity point, $\bar{z} = x_i$, and u_0 is continuous at $\bar{z}$. In this case by the same arguments the one-sided derivatives must satisfy (4.15) and in addition (4.16) if $(\bar{t}, \bar{x})$ is not in the shock set.

Cases R, RB. If $(\bar{t}, \bar{x})$ is a continuity point and $\bar{z} = x_i$ with $u_0(\bar{z}-) < u_0(\bar{z}+)$, then three (essentially two) cases can occur.

Case R. All backward characteristics through $(\bar{t}, x)$ with x in a small neighborhood of $\bar{x}$ meet $t = 0$ in $\bar{z}$. Then similar arguments as above show that with $\bar{w}$ from (4.8)

$$\frac{d}{dw}\zeta(t; \bar{z}, w, u_1)|_{w=\bar{w}} \geq 0, \quad 0 \leq t \leq \bar{t}, \tag{4.18}$$

holds, and if $(\bar{t}, x)$ is not a point of the shock set, there is $\beta > 0$ with

$$\frac{d}{dw}\zeta(t; \bar{z}, w, u_1)|_{w=\bar{w}} \geq \beta t > 0, \quad 0 < t \leq \bar{t}. \tag{4.19}$$

Case RB. $(\bar{t}, \bar{x})$ lies on the left or right boundary of a rarefaction wave. In this case the one-sided derivatives satisfy (4.15) and (4.18) (and moreover (4.16) and (4.19) if $(\bar{t}, x)$ is not a shock generation point), respectively.

Finally, it will be convenient to indicate that one of the above cases holds *and* the point is not in the shock set (i.e., not a shock generation point), which means that the unique forward characteristic remains genuine at least until some $\bar{t} + \tau$, $\tau > 0$.

Case C^c, CB^c, R^c, or RB^c. If $(\bar{t}, \bar{x})$ is of type C and (4.16) holds, then we call $(\bar{t}, \bar{x})$ of type C^c. If $(\bar{t}, \bar{x})$ is of type CB and (4.16) holds for the one-sided derivatives, then we call $(\bar{t}, \bar{x})$ of type CB^c. Similarly, if $(\bar{t}, \bar{x})$ is of type R and (4.19) holds, then we call $(\bar{t}, \bar{x})$ of type R^c. If $(\bar{t}, \bar{x})$ is of type RB and (4.16), respectively, (4.19), holds for the one-sided derivatives (see Case RB above), then we call $(\bar{t}, \bar{x})$ of type RB^c.

4.4. Classification of shock points. Now let $(\bar{t}, \bar{x})$ be a shock point located on the shock curve $\eta(t)$. We know from [8] and section 4.2 that the minimal and maximal backward characteristics $\xi_\mp(t)$ through $(\bar{t}, \bar{x})$ are genuine with initial condition (4.4). Thus, if we set $\bar{z}^\mp = \xi_\mp(0)$ we can classify the left and right states of the shock exactly as continuity points before, depending on whether or not $\bar{z}^\mp = x_i$ for some i. Moreover, the corresponding equations (4.6), (4.7) or (4.8), (4.9) hold with $\bar{z}^\mp$ and $y(\bar{t}, \bar{x}\mp)$ instead of $\bar{z}$, $y(\bar{t}, \bar{x})$. In particular, the following cases will be important.

Case $\mathrm{C}^c\mathrm{C}^c$. The extreme backward characteristics $\xi_\mp$ through $(\bar{t}, \bar{x})$ have the same properties as the backward characteristic through a continuity point of type C^c, i.e., we have $\bar{z}^\mp = \xi_\mp(0) \neq x_i$, $i = 1, \ldots, N$, and (4.16) holds for $\xi_\mp$ in $\bar{z} = \bar{z}^\mp$.

Case $\mathrm{R}^c\mathrm{R}^c$. The extreme backward characteristics $\xi_\mp$ through $(\bar{t}, \bar{x})$ have the same properties as the backward characteristic through a continuity point of type R^c; i.e., with $\bar{z}^\mp = \xi_\mp(0)$, $u_0(\bar{z}-) < u_0(\bar{z}+)$ holds for $\bar{z} = \bar{z}^\mp$, (4.8)–(4.9) hold with $(\bar{x}, \bar{z}) = (\bar{x}\mp, \bar{z}^\mp)$, $\bar{w} = \bar{w}^\mp \in (u_0(\bar{z}-), u_0(\bar{z}+))$, and (4.19) holds for $\xi_\mp$ with $(\bar{z}, \bar{w}) = (\bar{z}^\mp, \bar{w}^\mp)$.

Cases $\mathrm{C}^c\mathrm{R}^c$, $\mathrm{R}^c\mathrm{C}^c$. In Case $\mathrm{C}^c\mathrm{R}^c$ the minimal backward characteristic ξ_- has the properties as in Case C^c, the maximal characteristic ξ_+ has the properties as in Case R^c. The converse is true in Case $\mathrm{R}^c\mathrm{C}^c$.

5. Differentiability at continuity points. We start by analyzing the differentiability properties w.r.t. the control at continuity points that are not shock generation points and that are moreover not located on the boundary of a rarefaction wave (Case C^c or R^c).

5.1. Differentiability in continuity points of class C^c. We study first Case C^c; i.e., $(\bar t, \bar x)$ satisfies Case C and is not a shock generation point. By continuity there are $z_- < \bar z < z_+$ such that

(5.1)
$$u_0 \in C^1(J), \quad \frac{d}{dz}\zeta(t; z, u_0(z), u_1) \geq \beta > 0 \quad \forall\, t \in [0, \bar t],\ \forall\, z \in\,]z_- - \rho, z_+ + \rho[\stackrel{\text{def}}{=} J$$

for some $\beta, \rho > 0$. This allows us to solve (4.6) locally for $\bar z$ yielding with (4.7) a local regularity result for y.

5.1.1. Solution of the characteristic equations. We have the following result on the solvability of (4.6) and the resulting properties of y according to (4.7).

LEMMA 5.1. *Let* (A2)–(A3) *hold, let* $u = (u_0, u_1) \in PC^1(\mathbb{R}; x_1, \dots, x_N) \times L^\infty(0, T; C^1(\mathbb{R})^m)$, *and let* (5.1) *hold for some* $\beta, \rho > 0$. *Then there is* $\tau > 0$ *and a neighborhood*

$$V \subset C^1(J) \times L^\infty(0, T; C^1(\mathbb{R})^m)$$

of $u = (u_0, u_1)$ *such that*

$$\frac{d}{dz}\zeta(t; z, \hat u_0(z), \hat u_1) \geq \frac{\beta}{2} > 0 \quad \forall\ (t, z) \in [0, \bar t + \tau] \times J, \quad \forall\ \hat u \in V. \tag{5.2}$$

Moreover, for all $\hat u \in V$ *and all* (t, x) *in the stripe*

$$S = S(\tau) \stackrel{\text{def}}{=} \{(t, x)\ :\ t \in [0, \bar t + \tau],\ x \in [\xi_-(t), \xi_+(t)]\},$$

where $\xi_\mp(t) = \zeta(t; z_\mp, u_0(z_\mp), u_1)$, *the equation*

$$x = \zeta(t; z, \hat u_0(z), \hat u_1) \tag{5.3}$$

has in J *a unique solution* $z = Z(t, x, \hat u)$. *Set*

$$Y(t, x, \hat u) \stackrel{\text{def}}{=} v(t; Z(t, x, \hat u), \hat u_0(Z(t, x, \hat u)), \hat u_1). \tag{5.4}$$

Then $Z(\cdot, \hat u), Y(\cdot, \hat u) \in C^{0,1}(S)$. *The mappings*

$$(x, \hat u) \in\ ]\xi_-(t), \xi_+(t)[\times V \longmapsto (Z, Y)(t, x, \hat u), \quad t \in [0, \bar t + \tau[\,, \tag{5.5}$$

$$\hat u \in V \longmapsto (Z, Y)(\cdot, \hat u) \in C(S)^2 \tag{5.6}$$

are continuously Fréchet differentiable. The derivatives of (5.5) *are*

$$d_{(x,u)}Z(t, x, \hat u) \cdot (\delta x, \delta u) = \frac{\delta x - \delta\zeta(t; z, \hat u_0(z), \hat u_1; 0, \delta u_0(z), \delta u_1)}{\delta\zeta(t; z, \hat u_0(z), \hat u_1; 1, \hat u_0'(z), 0)}, \tag{5.7}$$

(5.8)
$$\begin{aligned} d_{(x,u)}Y(t, x, \hat u) \cdot (\delta x, \delta u) = {} & \delta v(t; z, \hat u_0(z), \hat u_1; 1, \hat u_0'(z), 0)\ d_{(x,u)}Z(t, x, \hat u) \cdot (\delta x, \delta u) \\ & + \delta v(t; z, \hat u_0(z), \hat u_1; 0, \delta u_0(z), \delta u_1), \end{aligned}$$

where $z = Z(t, x, \hat{u})$ *and* $(\delta\zeta, \delta v)$ *are given by* (4.12). *The derivative of* (5.6) *is*

$$d_u(Z, Y)(\cdot, \hat{u}) \cdot \delta u = d_{(x,u)}(Z, Y)(\cdot, \hat{u}) \cdot (0, \delta u). \tag{5.9}$$

NOTATION 5.2. *Given* $\bar{t}, z_-, z_+$ *satisfying* (5.1), *it will be convenient to indicate by*

$$(Y, Z, V, S(\tau), J) = \mathrm{YC}(u, \bar{t}, [z_-, z_+])$$

that the open interval $J \supset [z_-, z_+]$, *the stripe* $S = S(\tau)$, *the neighborhood* V, *and the functions* Y, Z *are obtained by applying Lemma* 5.1.

REMARK 5.3. *By construction,* $Y(\cdot, \hat{u}) \in C^{0,1}(S)$ *is an* S *classical solution of* (1.1) *for the control* $\hat{u} \in V$.

REMARK 5.4. *It is not difficult to show that* $\delta Y = d_u Y(\cdot, u) \cdot \delta u \in C(S)$ *is the unique broad solution (i.e., solution along characteristics) of the linearized equation*

$$\begin{aligned} &\partial_t \delta Y + \partial_x (f'(Y)\delta Y) = g_y(t, x, Y, u_1)\delta Y + g_{u_1}(t, x, Y, u_1)\delta u_1, \quad (t, x) \in S, \\ &\delta Y(0, x) = \delta u_0(x), \quad x \in [z_-, z_+], \end{aligned} \tag{5.10}$$

where $Y = Y(\cdot, u)$. *By Remark* 5.3 *and the differentiability of* (5.6) *we see from*

$$\partial_t Y(\cdot; \hat{u}) + \partial_x f(Y(\cdot; \hat{u})) = g(t, x, Y(\cdot; \hat{u}), \hat{u}_1)$$

with $\hat{u} = u + \sigma\delta u$ *by applying test functions* $p \in C^1(S)$, *integrating by parts, and taking the derivative in* $\sigma = 0$ *that* δY *is also a weak solution of* (5.10). *Even more, for any domain* $D \subset S$ *with Lipschitz boundary and any* $p \in C^{0,1}(D)$, *one has*

$$\begin{aligned} &(p\,(n_1 + n_2 f'(Y)), \delta Y)_{2,\partial D} \\ &= (\partial_t p + f'(Y)\partial_x p + g_y(t, x, Y, u_1)p, \delta Y)_{2,D} + (p\, g_{u_1}(t, x, Y, u_1), \delta u_1)_{2,D}, \end{aligned} \tag{5.11}$$

where $(n_1, n_2)^T$ *is the unit outer normal of* D. *In section* 9.3 *we will choose* p *as a solution of the adjoint equation* (3.25)–(3.26) *to obtain from* (3.21) *the gradient representation* (3.24) *for tracking-type functionals* (1.2), (3.19).

Proof of Lemma 5.1. Let $(\hat{u}_0, \hat{u}_1) \in C^1(J) \times L^\infty(0, T; C^1(\mathbb{R})^m)$. We have already observed that Lemma 4.4 implies the continuous Fréchet differentiability of (4.14). Thus, we deduce from (5.1) by continuity that (5.2) holds with a sufficiently small neighborhood $V \subset C^1(J) \times L^\infty(0, T; C^1(\mathbb{R})^m)$ of u and $\tau > 0$ small enough. Hence, for all $t \in [0, \bar{t} + \tau]$ and $\hat{u} \in V$ the mapping

$$z \in J \longmapsto \zeta(t; z, \hat{u}_0(z), \hat{u}_1)$$

is strictly monotone increasing and $]\zeta(t; z_-, \hat{u}_0(z_-), \hat{u}_1) - \beta\rho/2, \zeta(t; z_+, \hat{u}_0(z_+), \hat{u}_1) + \beta\rho/2[$ is contained in its image. Hence, for sufficiently small V and τ we get by continuity that $[\xi_-(t), \xi_+(t)]$ is contained in the image for all $t \in [0, \bar{t} + \tau]$ and all $\hat{u} \in V$. As a consequence, for all $(t, x) \in S$ and all $\hat{u} \in V$ there exists exactly one solution $z = Z(t, x, \hat{u}) \in J$ of (5.3). Since $Z(t, x, \hat{u}) \in J$ for all considered $x, \hat{u}$, we conclude from (5.2), (5.3) and the continuous Fréchet differentiability of (4.14) by the implicit function theorem that the first component of (5.5) is continuously Fréchet differentiable. By (5.4) and the continuous differentiability of (4.14) the second component in (5.5) is also continuously Fréchet differentiable. The formula (5.7) is an immediate consequence of the implicit function theorem, and (5.8) follows from (5.4). The

Lipschitz continuity of $Z(\cdot,\hat{u})$ follows directly from (5.2) and the Lipschitz continuity of (5.3) w.r.t. t,x. Now the Lipschitz continuity of $Y(\cdot,\hat{u})$ is clear by (5.4) and the Lipschitz continuity of $v(t;z,w,\hat{u}_1)$ w.r.t. t,z,w.

To show the differentiability of (5.6) we observe that for all $\hat{u}\in V$ the function $Z(\cdot,\hat{u})$ is in $C(S;J)$ and satisfies $F(Z(\cdot,\hat{u}),\hat{u})=0$ with the operator

$$\begin{aligned} F:(z,\hat{u}_0,\hat{u}_1)\in C(S;J)\times C^1(J)\times L^\infty(0,T;C^1(\mathbb{R})^m)\\ \longmapsto (\zeta(\cdot_t;z(\cdot),\hat{u}_0(z(\cdot)),\hat{u}_1)-\cdot_x)\in C(S), \end{aligned}$$

where $\cdot_t,\cdot_x$ denote the projections on the t- and x-components, respectively. F is continuously Fréchet differentiable by Lemma 4.4, and we have obviously

$$d_zF(z,\hat{u})\cdot\delta z=\left((t,x)\longmapsto \frac{d}{d\tilde{z}}\zeta(t;\tilde{z},\hat{u}_0(\tilde{z}),\hat{u}_1)|_{\tilde{z}=z(t,x)}\delta z(t,x)\right).$$

By (5.2) we get that $(d_zF(z,\hat{u}))^{-1}$ exists and is bounded for all $z\in C(S;J)$ and all $\hat{u}\in V$. Hence, the first component of (5.6) is continuously Fréchet differentiable by the implicit function theorem, and now by (5.4) and Lemma 4.4 the second is also. The formula (5.9) is obvious. The properties of these mappings follow directly from Lemma 4.4. ∎

5.1.2. Differentiability result in continuity points of type C^c. We are now able to characterize the properties of y in continuity points of class C^c.

LEMMA 5.5 (differentiability properties in continuity points of class C^c). *Let* (A2)–(A3) *hold and let* $u=(u_0,u_1)\in PC^1(\mathbb{R};x_1,\dots,x_N)\times L^\infty(0,T;C^1(\mathbb{R})^m)$. *Let* $(\bar{t},\bar{x})\in\Omega_T$ *be a point of continuity of* $y=y(\cdot;u)$ *of class* C^c, *i.e., outside the shock set such that* $\bar{z}\neq x_i$, $1\le i\le N$, *holds for* $\bar{z}$ *given by* (4.6). *Then the following hold:*

(i) *There is a maximal nonempty open interval* I *such that* $\{\bar{t}\}\times I$ *does not contain points of the shock set and that all backward characteristics through* $(\bar{t},x)$, $x\in I$, *meet* $t=0$ *not in* x_i, $i=1,\dots,N$. $y(\bar{t},\cdot;u)$ *is continuously differentiable on* I.

(ii) *Let* $\hat{I}=]x_-,x_+[$ *be an arbitrary interval with closure in* I, *denote by* $\xi_\mp$ *the genuine backward characteristics through* $(\bar{t},x_\mp)$, *and set* $z_\mp=\xi_\mp(0)$. *Then there are* $\beta>0$, $\rho>0$ *such that* (5.1) *is satisfied. Using Notation* 5.2 *let* $(Y,Z,V,S(\tau),J)=\mathrm{YC}(u,\bar{t},[z_-,z_+])$ *be obtained according to Lemma* 5.1. *Given* $M_\infty>0$ *there are* $R>0$, $\nu>0$ *such that after a possible reduction of* τ *and* V *the following holds:*

$$\begin{aligned} y(t,x;\hat{u})=Y(t,x,\hat{u}_0|_J,\hat{u}_1)\quad\forall\,(t,x)\in S,\quad\forall\,\hat{u}\in\hat{V},\quad \textit{where}\\ \hat{V}\overset{\mathrm{def}}{=}\{(\hat{u}_0,\hat{u}_1)\in L^\infty(\mathbb{R})\times L^\infty(0,T;C^1(\mathbb{R})^m)\ :\ (\hat{u}_0|_J,\hat{u}_1)\in V,\\ \|\hat{u}_0-u_0\|_{\infty,\mathbb{R}\setminus J}<M_\infty,\|\hat{u}_0-u_0\|_{1,[-R,R]\setminus J}<\nu\}. \end{aligned}$$

Hence, the differentiability results of Lemma 5.1 *for* Y *carry over to* $y|_S$.

Proof. (i) $(\bar{t},\bar{x})$ is of class C^c, since $\bar{z}\neq x_i$, $i=1,\dots,N$, and $(\bar{t},\bar{x})$ is a continuity point outside the shock set. Hence, we know that (4.16) holds for some $\beta>0$, and we find $z_-<\bar{z}<z_+$, $\rho>0$ such that (5.1) is satisfied. Therefore, Lemma 5.1 is applicable, yielding $(Y,Z,V,S(\tau),J)=\mathrm{YC}(u,\bar{t},[z_-,z_+])$; see Notation 5.2. Then clearly $u_0\in C^1(J)$, $i=1,\dots,N$. By the above-mentioned results of [8] we have (cf. (4.6), (4.7))

$$y(\bar{t},\bar{x};u)=Y(\bar{t},\bar{x},u).$$

Now $x \longmapsto y(\bar{t}, x; u)$ is continuous outside a countable set and the identities (4.6), (4.7) with $\tilde{x}, \tilde{z}$ instead of $\bar{x}, \bar{z}$ hold for all continuity points $(\bar{t}, \tilde{x})$. Using the backward stability of (4.3) according to Lemma 4.4 we see that $\tilde{z} \to \bar{z}$ for continuity points $\tilde{x}$ with $\tilde{x} \to \bar{x}$. Hence, for all continuity points $\tilde{x} \in]\xi_-(\bar{t}), \xi_+(\bar{t})[$ with $\tilde{z} \in J$ we must have $\tilde{z} = Z(\bar{t}, \tilde{x}, u)$ and thus

$$y(\bar{t}, x; u) = Y(\bar{t}, x, u)$$

for all $x \in \bar{I} \stackrel{\text{def}}{=}]\bar{x}_-, \bar{x}_+[$, where $\bar{I}$ is sufficiently small with $\bar{x} \in \bar{I}$, $\bar{I} \subset]\xi_-(\bar{t}), \xi_+(\bar{t})[$ (first for the dense set of continuity points $x = \tilde{x} \in \bar{I}$ and thus for all $x \in \bar{I}$ by our convention $y(t, x) = y(t, x-)$ for the choice of y). This shows that $y(\bar{t}, \cdot; u)$ is continuously differentiable on $\bar{I} \neq \emptyset$. Moreover $\{\bar{t}\} \times \bar{I}$ does not contain points of the shock set, since (5.1) holds, and all backward characteristics starting in $\{\bar{t}\} \times \bar{I}$ meet $t = 0$ not in x_i, $i = 1, \ldots, N$. Hence, there exists a maximal open nonempty interval I with the asserted properties.

(ii) Let $\hat{I} \stackrel{\text{def}}{=}]x_-, x_+[\neq \emptyset$ be arbitrary with closure in I. Now for any point $\bar{x} \in [x_-, x_+]$ we can argue as above and find an interval-neighborhood $\bar{I}$ with the above properties. Taking a finite covering, we get the following: Denote by $\xi_\mp$ the genuine backward characteristics through $(\bar{t}, x_\mp)$; set $z_\mp = \xi_\mp(0)$. By the finite covering, we obtain $\rho > 0$ and $\beta > 0$ such that (5.1) holds. Hence, Lemma 5.1 yields $(Y, Z, V, S(\tau), J) = \mathrm{YC}(u, \bar{t}, [z_-, z_+])$. Then we have from the proof of (i) that

$$y(\bar{t}, x; u) = Y(\bar{t}, x, u)$$

for all $x \in \hat{I}$. We show also that $y(\cdot; u) = Y(\cdot, u)$ on $S = S(\tau)$ after a possible reduction of $\tau > 0$. In fact, $(\bar{t}, x_\mp)$ are continuity points of $y(\cdot; u)$, since $y(\bar{t}, \cdot; u)$ is continuous at $x_\mp$; cf. [8]. Hence, by (5.1) and [8, Lem. 5.2], $(\bar{t}, x_\mp)$ are not shock generation points, i.e., the genuine characteristics $\xi_\mp$ remain genuine until $\bar{t} + \tau$ after a possible reduction of $\tau > 0$. Now take any $(t, x) \in S$. The extreme backward characteristics through (t, x) are genuine and may not intersect $\xi_\mp$. Hence, they stay in S and must thus coincide with $\zeta(\cdot; Z(t, x, u), u_0(Z(t, x, u)), u_1)$, which yields

$$y(t, x; u) = Y(t, x, u) \quad \forall\, (t, x) \in S.$$

Since $\hat{I} \stackrel{\text{def}}{=}]x_-, x_+[$ was an arbitrary interval with closure in I, the same arguments apply to $\tilde{I} =]x_- - 3\eta, x_+ + 3\eta[$ for $\eta > 0$ small enough, yielding, after a possible reduction of τ, that

$$y(t, x; u) = Y(t, x, u) \quad \forall\, (t, x) \in \tilde{S}, \quad Z(t, x, u) \in J \quad \forall\, (t, x) \in \tilde{S}$$

holds, where $\tilde{S} = \tilde{S}(\tau)$ is confined by the genuine backward characteristics $\tilde{\xi}_\mp$ through $(\bar{t}, x_\mp \mp 3\eta)$. We can clearly choose $\tau > 0$ small enough such that with $\tilde{t} = \bar{t} + \tau$ the inequalities $\tilde{\xi}_-(\tilde{t}) < \xi_-(\tilde{t}) - 2\eta$ and $\tilde{\xi}_+(\tilde{t}) > \xi_+(\tilde{t}) + 2\eta$ hold. Now let $M_\infty > 0$ be given. We will show that there are $R > 0$, $\nu > 0$ such that after a possible reduction of V and $\tau > 0$

$$y(t, x; \hat{u}) = Y(t, x; \hat{u}) \quad \forall\, (t, x) \in S, \quad \forall\, \hat{u} \in \hat{V} \tag{5.12}$$

holds with $\hat{V}$ defined in (ii). To this purpose we note that independent of ν, R there is $M > 0$ such that $\|\hat{u}_0\|_\infty < M$ and $\|\hat{u}_1\|_\infty < M$ for all $\hat{u} \in \hat{V}$. Thus, by Theorem 4.1(i) there is $M_y > 0$ with $\|y(.; \hat{u})\|_\infty \leq M_y$ for all $\hat{u} \in \hat{V}$, and we can thus for convenience

assume (A2′) instead of (A2), ensuring with Lemma 4.4 the local Lipschitz stability of backward characteristics. Thus, if we denote by $(\zeta^{\tilde{t}}, v^{\tilde{t}})(t; x, w, u_1)$ the backward solution of (4.3) with $(\zeta^{\tilde{t}}, v^{\tilde{t}})(\tilde{t}; x, w, u_1) = (x, w)$, we can reduce V and find $\varepsilon > 0$ such that

(5.13)
$$\zeta^{\tilde{t}}(0; x, w, \hat{u}_1) \in J, \quad \zeta^{\tilde{t}}(t; x, w, \hat{u}_1) \begin{cases} < \xi_-(t) \text{ if } x \in [\tilde{\xi}_-(\tilde{t}), \xi_-(\tilde{t}) - \eta], \\ > \xi_+(t) \text{ if } x \in [\xi_+(\tilde{t}) + \eta, \tilde{\xi}_+(\tilde{t})], \end{cases} \quad t \in [0, \tilde{t}],$$

whenever $\hat{u} \in \hat{V}$, $|w - y(\tilde{t}, x; u)| \le \varepsilon$. Now we have for $I_R = [-R, R]$ with sufficiently large $R > 0$ by Theorem 4.1(ii) the local L^1-stability estimate

$$\|y(\tilde{t}, \cdot; \hat{u}) - y(\tilde{t}, \cdot; u)\|_{1, [\tilde{\xi}_-(\tilde{t}), \tilde{\xi}_+(\tilde{t})]} \le C \left(\|\hat{u}_0 - u_0\|_{1, I_R} + \|\hat{u}_1 - u_1\|_{1, [0,T] \times I_R} \right) \tag{5.14}$$

for all $\hat{u} \in \hat{V}$ with C only depending on f, g, u, V, M_∞. From (5.14) and the definition of $\hat{V}$ we deduce that for $\nu > 0$ small enough

$$\operatorname*{ess\,inf}_{x \in I(\tilde{\xi}_\mp(\tilde{t}), \xi_\mp(\tilde{t}) \mp \eta)} |y(\tilde{t}, \cdot; \hat{u}) - y(\tilde{t}, \cdot; u)| < \frac{C\nu}{\eta} \le \varepsilon \tag{5.15}$$

whenever $\hat{u} \in \hat{V}$. Since $y(\tilde{t}, \cdot; \hat{u}) \in BV_{loc}(\mathbb{R})$, we obtain by combining (5.13) and (5.15) that for any $\hat{u} \in \hat{V}$ we can find continuity points $\hat{x}_\mp \in I(\tilde{\xi}_\mp(\tilde{t}), \xi_\mp(\tilde{t}) \mp \eta)$ of $y(\tilde{t}, \cdot; \hat{u})$ such that the genuine backward characteristics $\hat{\xi}_\mp(t) = \zeta^{\tilde{t}}(t; \hat{x}_\mp, y(\tilde{t}, \hat{x}_\mp; \hat{u}), \hat{u}_1)$ satisfy (5.13) with $x = \hat{x}_\mp$, $w = y(\tilde{t}, \hat{x}_\mp; \hat{u})$, respectively. Since $\hat{\xi}_\mp(0) \in J$, we must have $\hat{\xi}_\mp(0) = Z(\tilde{t}, \hat{x}_\mp, \hat{u})$ and therefore $y(\tilde{t}, \hat{x}_\mp; \hat{u}) = Y(\tilde{t}, \hat{x}_\mp, \hat{u})$. Now any genuine backward characteristic ξ of $y(\cdot; \hat{u})$ through $(t, x) \in S$ may not leave the area confined by the genuine characteristics $\hat{\xi}_\mp$. Therefore, $\xi(0) \in J$ and ξ must by Lemma 5.1 coincide with the unique genuine forward characteristic starting in $Z(t, x, \hat{u})$ and, consequently, $y(t, x; \hat{u}) = Y(t, x, \hat{u})$. Thus, (5.12) is proven. ∎

5.2. Differentiability in the interior of rarefaction waves (Case R^c). In a second step, we look at continuity points of class R^c, i.e., located outside the shock set in the interior of a rarefaction wave. Then (4.8), (4.9) hold with $\bar{z} = x_i$, $u_0(\bar{z}-) < u_0(\bar{z}+)$ for some $i \in \{1, \dots, N\}$ and some $\bar{w} \in]u_0(\bar{z}-), u_0(\bar{z}+)[$. Moreover, we obtain from (4.19) by continuity $w_- < \bar{w} < w_+$ and $\beta, \rho > 0$ with

$$\frac{d}{dw} \zeta(t; \bar{z}, w, u_1) \ge \beta t > 0 \quad \forall\, t \in]0, \bar{t}], \quad \forall\, w \in]w_- - \rho, w_+ + \rho[\stackrel{\text{def}}{=} J_w. \tag{5.16}$$

Then we can locally solve (4.8) yielding with (4.9) a representation for rarefaction waves.

5.2.1. Solution of the characteristic equations. The following lemma is a counterpart of Lemma 5.1 for the solution of (4.8) w.r.t. $\bar{w}$ and the resulting properties of y in (4.9).

LEMMA 5.6. *Let assumptions* (A2)–(A3) *hold, let* $(u_0, u_1) \in PC^1(\mathbb{R}; x_1, \dots, x_N) \times L^\infty(0, T; C^1(\mathbb{R})^m)$, *and let* (5.16) *hold for some* $\beta, \rho > 0$. *Then there is* $\tau > 0$ *and a neighborhood* $V_1 \subset L^\infty(0, T; C^1(\mathbb{R})^m)$ *of* u_1 *such that*

$$\frac{d}{dw} \zeta(t; \bar{z}, w, \hat{u}_1) \ge \frac{\beta}{2} t > 0 \quad \forall\, (t, w) \in]0, \bar{t} + \tau] \times J_w, \quad \forall\, \hat{u}_1 \in V_1. \tag{5.17}$$

Moreover, for all $\hat{u}_1 \in V_1$ and all (t,x) in the stripe

$$S \stackrel{\text{def}}{=} \{(t,x) \,:\, t \in]0, \bar{t}+\tau], x \in [\xi_-(t), \xi_+(t)]\}\,, \quad \xi_\mp(t) = \zeta(t; \bar{z}, w_\mp, u_1),$$

the equation

$$x = \zeta(t; \bar{z}, w, \hat{u}_1) \tag{5.18}$$

has in J_w a unique solution $w = W(t,x,\hat{u}_1)$. Set

$$Y_r(t,x,\hat{u}_1) \stackrel{\text{def}}{=} v(t; \bar{z}, W(t,x,\hat{u}_1), \hat{u}_1). \tag{5.19}$$

Then $W(\cdot,\hat{u}_1), Y_r(\cdot,\hat{u}_1) \in C^{0,1}(S \cap \{t \geq \tilde{t}\})$ for any $\tilde{t} \in]0,\bar{t}[$ and the mappings

$$(x,\hat{u}_1) \in\ ]\xi_-(t), \xi_+(t)[\times V_1 \longmapsto (W,Y_r)(t,x,\hat{u}_1), \quad t \in]\tilde{t}, \bar{t}+\tau[\,, \tag{5.20}$$

$$\hat{u}_1 \in V_1 \longmapsto (W,Y_r)(\cdot,\hat{u}_1) \in C(S \cap \{t \geq \tilde{t}\})^2 \tag{5.21}$$

are continuously Fréchet differentiable. The derivatives of (5.20) *are*

$$d_{(x,u_1)} W(t,x,\hat{u}_1) \cdot (\delta x, \delta u_1) = \frac{\delta x - \delta\zeta(t; \bar{z}, w, \hat{u}_1; 0, 0, \delta u_1)}{\delta\zeta(t; \bar{z}, w, \hat{u}_1; 0, 1, 0)}, \tag{5.22}$$

$$\begin{aligned} d_{(x,u_1)} Y_r(t,x,\hat{u}_1) \cdot (\delta x, \delta u_1) = {} & \delta v(t; \bar{z}, w, \hat{u}_1; 0, 1, 0)\ d_{(x,u_1)} W(t,x,\hat{u}_1) \cdot (\delta x, \delta u_1) \\ & + \delta v(t; \bar{z}, w, \hat{u}_1; 0, 0, \delta u_1), \end{aligned} \tag{5.23}$$

where $w = W(t,x,\hat{u}_1)$ and $(\delta\zeta, \delta v)$ are given by (4.12). *The derivative of* (5.21) *is*

$$d_{u_1}(W,Y_r)(\cdot,\hat{u}_1) \cdot \delta u_1 = d_{(x,u_1)}(W,Y_r)(\cdot,\hat{u}_1) \cdot (0, \delta u_1). \tag{5.24}$$

Finally, there is $C > 0$ such that

$$|d_{u_1}(W,Y_r)(\cdot,\hat{u}_1) \cdot \delta u_1| \leq C\, t\, \|\delta u_1\|_{L^\infty(0,T;C(\mathbb{R}))} \quad \text{in } S,$$

which allows a continuous extension to $(0,\bar{z})$ by the value 0.

Proof. Except for the last assertion the proof is very similar to the one of Lemma 5.1 and therefore omitted. For the last statement we observe that with a generic constant C an application of Gronwall's lemma to (4.12) yields $|\delta v(t; \bar{z}, w, \hat{u}_1; 0, 0, \delta u_1)| \leq C\, t\, \|\delta u_1\|_{L^\infty(0,T;C(\mathbb{R}))}$ and then $|\delta\zeta(t; \bar{z}, w, \hat{u}_1; 0, 0, \delta u_1)| \leq C\, t^2\, \|\delta u_1\|_{L^\infty(0,T;C(\mathbb{R}))}$ by the first equation in (4.12). Using this together with (5.17) in (5.22), (5.23) gives the asserted bound. ∎

NOTATION 5.7. *Given $\bar{t}, \bar{z}, w_-, w_+$ satisfying* (5.16) *for some $\beta, \rho > 0$, we indicate by*

$$(Y_r, W, V_1, S(\tau), J_w) = \mathrm{YR}(u, \bar{t}, \bar{z}, [w_-, w_+])$$

that the open interval $J_w \supset [w_-, w_+]$, the stripe $S = S(\tau)$, the neighborhood V_1, and the functions Y_r, Z are obtained by applying Lemma 5.6.

REMARK 5.8. *By construction, $Y_r(\cdot,\hat{u}_1) \in C^{0,1}(S \cap \{t > \tilde{t}\})$ is for any $\tilde{t} \in]0,\bar{t}[$ on $S \cap \{t > \tilde{t}\}$ a classical solution of* (1.1) *for the control $\hat{u}_1 \in V_1$.*

REMARK 5.9. *It can easily be verified that $\delta Y = d_{u_1} Y_r(\cdot, u_1) \cdot \delta u_1 \in C(S)$ is a broad solution of the linearized equation*

$$\begin{aligned} &\partial_t \delta Y + \partial_x (f'(Y_r)\delta Y) = g_y(t,x,Y_r,\hat{u}_1)\delta Y + g_{u_1}(t,x,Y_r,\hat{u}_1)\delta u_1, \quad (t,x) \in S, \\ &\lim_{(t,x)\in S, t\to 0} \delta Y(t,x) = 0. \end{aligned}$$

Moreover, using the same arguments as in Remark 5.4, δY *is also a weak solution, and for every domain* $D \subset S$ *with Lipschitz boundary and any* $p \in C(D)$ *with* $p \in C^{0,1}(D \cap \{t \geq \tilde{t}\})$ *for all* $\tilde{t} \in]0, \bar{t}[$, *the identity* (5.11) *holds. This follows by integrating by parts over* $D \cap \{t > \tilde{t}\}$ *and letting* $\tilde{t} \to 0$.

5.2.2. Differentiability result in continuity points of type $\mathbf{R^c}$. Using Lemma 5.6 we obtain the following regularity result at continuity points that are located on the interior of a rarefaction wave.

LEMMA 5.10 (differentiability properties in continuity points of class R^c). *Let* (A2)–(A3) *hold and let* $u = (u_0, u_1) \in PC^1(\mathbb{R}; x_1, \dots, x_N) \times L^\infty(0, T; C^1(\mathbb{R})^m)$. *Let* $(\bar{t}, \bar{x}) \in \Omega_T$ *be a point of continuity of* $y = y(\cdot; u)$ *outside the shock set such that with* $\bar{z}$, $\bar{w}$ *in* (4.8), (4.9), $u_0(\bar{z}-) < \bar{w} < u_0(\bar{z}+)$ *holds. Then the following statements are true:*

(i) *There is a maximal nonempty open interval* I *such that* $\{\bar{t}\} \times I$ *does not contain points of the shock set and such that all backward characteristics through* $(\bar{t}, x)$, $x \in I$, *meet* $t = 0$ *in* $\bar{z}$. $y(\bar{t}, \cdot; u)$ *is continuously differentiable on* I.

(ii) *Let* J *be an arbitrary neighborhood of* $\bar{z}$. *Let* $\hat{I} =]x_-, x_+[$ *be an arbitrary interval with closure in* I, *and denote by* $\xi_\mp$ *the genuine backward characteristics through* $(\bar{t}, x_\mp)$ *and by* $w_\mp \in]u_0(\bar{z}-), u_0(\bar{z}+)[$ *the corresponding values of* $\bar{w}$ *in* (4.8), (4.9). *Then there are* $\beta > 0$, $\rho > 0$ *such that* (5.16) *is satisfied. Using Notation* 5.7 *let* $(Y_r, W, V_1, S(\tau), J_w) = \mathrm{YR}(u, \bar{t}, \bar{z}, [w_-, w_+])$ *be obtained according to Lemma* 5.6. *Given* $M_\infty > 0$ *and* $\tilde{t} \in]0, \bar{t}[$ *there are* $R > 0$, $\nu > 0$ *such that after a possible reduction of* τ *and* V_1 *the following holds:*

$$y(t, x; \hat{u}) = Y_r(t, x, \hat{u}_1) \quad \forall\, (t, x) \in S \cap \left\{t \geq \tilde{t}\right\}, \quad \forall\, \hat{u} \in \hat{V}, \quad \textit{where}$$
$$\hat{V} \stackrel{\text{def}}{=} \{(\hat{u}_0, \hat{u}_1) \in L^\infty(\mathbb{R}) \times L^\infty(0, T; C^1(\mathbb{R})^m) \ : \ \hat{u}_1 \in V_1, \|\hat{u}_0 - u_0\|_{\infty, \mathbb{R} \setminus J} < M_\infty,$$
$$\|\hat{u}_0 - u_0\|_{\infty, J} < \nu, \quad \|\hat{u}_0 - u_0\|_{1, [-R, R] \setminus J} < \nu\}.$$

Hence, the differentiability properties of Y_r *according to Lemma* 5.6 *hold also for* $y|_S$.

Proof. The proof is very similar to the proof of Lemma 5.5, and we only sketch the differences. One has to use (4.19) instead of (4.16) to obtain the applicability of Lemma 5.6. Then by backward stability all characteristics ξ starting in continuity points $(\bar{t}, \tilde{x})$ close to $(\bar{t}, \bar{x})$ must hit $t = 0$ in $\bar{z}$, since $y(t, \xi(t))$ converges for $t \to 0$, $\tilde{x} \to \bar{x}$ to $\bar{w} \in]u_0(\bar{z}-), u_0(\bar{z}+)[$. Now by Lemma 5.6 the same arguments as in the proof of Lemma 5.5 can be used. Finally, one shows with the L^1-stability estimate that for sufficiently small $\nu > 0$, $\tau > 0$ and V_1 also the backward characteristics of $y(\cdot; \hat{u})$, $\hat{u} \in \hat{V}$, through $(\hat{t}, \hat{x}) \in S \cap \left\{t \geq \tilde{t}\right\}$ must hit $t = 0$ in $\bar{z}$. ∎

Before we proceed to the more involved analysis of shock points in section 6, we consider continuity points of class CB^c and continuity points of class RB^c, i.e., on the boundary of a rarefaction wave.

5.3. Differentiability in continuity points of class $\mathbf{CB^c}$. Next we consider the situation of continuity points $(\bar{t}, \bar{x})$ of class CB^c, i.e., outside of the shock set such that the backward characteristic ξ through $(\bar{t}, \bar{x})$ meets $t = 0$ in a point $\bar{z} = x_i$, $i \in \{1, \dots, N\}$, where u_0 is continuous. Then u_0 is not necessarily differentiable at $\bar{z} = x_i$, and a shift-variation of u_0 can create a discontinuity at x_i, since it allows locally a variation in $PC^1(J; x_i)$ with an interval J containing x_i. Therefore, the

entropy solution $y(\cdot;\hat{u})$ for varied controls $\hat{u}$ can develop a shock or rarefaction wave arbitrarily close to $(\bar{t},\bar{x})$.

We have observed in section 4.3 that the one-sided derivatives in $\bar{z}$ satisfy (4.16). Thus, if we define the C^1-prolongations

$$\begin{aligned} u_0^- &\stackrel{\text{def}}{=} u_0|_{\{z<\bar{z}\}} + (u_0(\bar{z}-) + u_0'(\bar{z}-)(\cdot-\bar{z}))|_{\{z\geq\bar{z}\}}, \\ u_0^+ &\stackrel{\text{def}}{=} u_0|_{\{z>\bar{z}\}} + (u_0(\bar{z}+) + u_0'(\bar{z}+)(\cdot-\bar{z}))|_{\{z\leq\bar{z}\}}, \end{aligned} \tag{5.25}$$

then there are $z_- < \bar{z} < z_+$, $\rho,\beta > 0$ such that (5.1) holds for $u_0^\mp$ instead of u_0. Thus, Lemma 5.1 is applicable for $u_0^\mp$ instead of u_0, yielding with Notation 5.2

$$(Y_\mp, Z_\mp, V_\mp, S_\mp(\tau), J) = \mathrm{YC}(u_0^\mp, u_1, [z_-, z_+]).$$

Now it is obvious from Lemma 5.1 and (5.25) that for $\delta > 0$ small enough and

$$V = \{\hat{u} = (\hat{u}_0, \hat{u}_1) \in \mathcal{V} : \ \|\hat{u} - u\|_{\mathcal{V}} < \delta\}, \quad \mathcal{V} \stackrel{\text{def}}{=} PC^1(J;\bar{z}) \times L^\infty(0,T;C^1(\mathbb{R})^m) \tag{5.26}$$

that the mappings

$$\hat{u} \in V \longmapsto (Z_\mp, Y_\mp)(\cdot, \hat{u}_0^\mp, \hat{u}_1) \in C(S_\mp) \tag{5.27}$$

are continuously differentiable. The next lemma shows that piecing together $Y_\mp$ along ξ yields a first order approximation of $y(\cdot,\hat{u})$ in u for $\hat{u} \in V$.

Lemma 5.11 (differentiability properties in continuity points of class CB^c). *Let* (A2)–(A3) *hold and let* $(\bar{t},\bar{x})$ *be a continuity point of* $y(\cdot;u)$ *of type* CB^c, *i.e., outside of the shock set such that the backward characteristic* ξ *through* $(\bar{t},\bar{x})$ *meets* $t = 0$ *in a continuity point* $\bar{z} = x_i$, $i \in \{1,\dots,N\}$, *of* u_0. *As explained above, define the prolongations* $u_0^\mp$ *in* (5.25) *and let* $z_- < \bar{z} < z_+$ *be close enough to* $\bar{z}$ *such that Lemma* 5.1 *can be applied for* $u_0^\mp$, *yielding with Notation* 5.2 $(Y_\mp, Z_\mp, V_\mp, S_\mp(\tau), J) = \mathrm{YC}(u_0^\mp, u_1, [z_-, z_+])$. *Finally, let* $V, \mathcal{V}$ *be defined as in* (5.26) *with* $\delta > 0$ *small enough. Then the mapping*

$$\begin{aligned} &\hat{u} \in V \longmapsto \tilde{y}(\cdot;\hat{u}) \in C(S_- \cap \{x \leq \xi(t)\}) \cap C(S_+ \cap \{x > \xi(t)\}), \\ &\tilde{y}(t,x;\hat{u}) \stackrel{\text{def}}{=} \begin{cases} Y_-(t,x,\hat{u}_0^-|_J,\hat{u}_1), & (t,x) \in S_- \cap \{x \leq \xi(t)\}, \\ Y_+(t,x,\hat{u}_0^+|_J,\hat{u}_1), & (t,x) \in S_+ \cap \{x > \xi(t)\} \end{cases} \end{aligned} \tag{5.28}$$

is continuously Fréchet differentiable.

For any $M_\infty > 0$ *there are* $R > 0$, $\nu > 0$ *and a neighborhood* $\hat{I} =]x_-, x_+[$ *of* $\bar{x}$ *such that with*

$$\begin{aligned} \hat{V} \stackrel{\text{def}}{=} \{(\hat{u}_0,\hat{u}_1) \in L^\infty(\mathbb{R}) \times L^\infty(0,T;C^1(\mathbb{R})^m) \ : \ &(\hat{u}_0|_J, \hat{u}_1) \in V, \\ &\|\hat{u}_0 - u_0\|_{\infty,\mathbb{R}\setminus J} < M_\infty, \|\hat{u}_0 - u_0\|_{1,[-R,R]\setminus J} < \nu\}, \end{aligned}$$

where V *is given in* (5.26), *for all* $r \in [1,\infty[$

$$\lim_{\substack{\hat{u}\in\hat{V} \\ \|\hat{u}-u\|_{\mathcal{V}}\to 0}} \frac{\|y(\bar{t},\cdot;\hat{u}) - \tilde{y}(\bar{t},\cdot;\hat{u})\|_{r,\hat{I}}}{\|\hat{u}-u\|_{\mathcal{V}}} = 0$$

holds. In particular, $\hat{u} \in (\hat{V}, \|\cdot\|_{\mathcal{V}}) \longmapsto y(\bar{t},\cdot;\hat{u}) \in L^r(\hat{I})$ *is Fréchet differentiable at* u.

Proof. The construction of J, V, $Z_\mp$, and $Y_\mp$ was already justified above. Set $J_\mp = J \cap \{\mp(z-\bar z)>0\}$. Moreover, since (5.27) are continuously Fréchet differentiable, the same is true for (5.28). Since $(\bar t, x)$ is a continuity point outside the shock set, we know that (5.1) holds on an interval J containing $\bar z$ for u_0 (and also for $u_0^\mp$ instead of u_0), and we obtain, as at the beginning of the proof of Lemma 5.5, $x_- < \bar x < x_+$ such that

$$y(\bar t, x; u) = \tilde y(\bar t, x; u) \quad \forall\, x \in \hat I \stackrel{\text{def}}{=}]x_-, x_+[.$$

Given $M_\infty > 0$ and $\varepsilon > 0$ we thus find by Lemma 5.5(ii) $R > 0$ and $\nu > 0$ such that after a possible reduction of δ (and hence V in (5.26)) one has

$$y(\bar t, x; \hat u) = \tilde y(\bar t, x; \hat u) \quad \forall\, x \in \hat I \setminus [\bar x - \varepsilon/2, \bar x + \varepsilon/2], \quad \forall\ \hat u \in \hat V. \tag{5.29}$$

Hereby we have used the fact that $\hat u_0|_{J_\mp} = \hat u_0^\mp|_{J_\mp}$. One easily sees that $R > 0$ and $\nu > 0$ can be chosen independent of $\varepsilon > 0$. In particular, for $\varepsilon > 0$ small enough the genuine backward characteristics $\hat\xi_\mp$ of $y(\cdot;\hat u)$ through $(\bar t, \bar x \mp \varepsilon)$ satisfy $\hat\xi_\mp(0) = Z_\mp(\bar t, \bar x \mp \varepsilon, \hat u) \in J$, and thus $\xi(0) \in J$ holds for all genuine backward characteristics ξ through $(\bar t, x)$, $|x - \bar x| \le \varepsilon$. Let x be an arbitrary continuity point with $|x - \bar x| \le \varepsilon$ and set $z = \xi(0)$. Then we obviously have

$$\begin{aligned} |y(\bar t, x; \hat u) - y(\bar t, x; u)| &\le |v(\bar t; z, \hat u_0(z), \hat u_1) - v(\bar t; z, u_0(z), u_1)| \\ &\quad + L_x |\zeta(\bar t; z, \hat u_0(z), \hat u_1) - \zeta(\bar t; z, u_0(z), u_1)| \\ &\le (1 + L_x) L (|\hat u_0(z) - u_0(z)| + \|\hat u_1 - u_1\|_{L^\infty(0,T;C(\mathbb{R}))}), \end{aligned}$$

where L_x is a Lipschitz constant of $y(\bar t, \cdot; u)$ on $\hat I$, and L is a Lipschitz constant of v; cf. Lemma 4.4. Moreover, $y(\bar t, x; u)$ coincides with $Y_-(\bar t, x, u_0^-, u_1)$ if $x < \bar x$ and with $Y_+(\bar t, x, u_0^+, u_1)$ if $x > \bar x$. Thus, using the local Lipschitz continuity of (5.27), we obtain with a constant $L_Y > 0$

$$|\tilde y(\bar t, x; \hat u) - y(\bar t, x; u)| \le \max\big(|Y_\mp(\bar t, x, \hat u_0^\mp, \hat u_1) - Y_\mp(\bar t, x, u_0^\mp, u_1)|\big) \le L_Y \|\hat u - u\|_{\mathcal{V}}.$$

This together gives, for $r \in [1, \infty[$,

$$\|y(\bar t, \cdot; \hat u) - \tilde y(\bar t, \cdot; \hat u)\|_{r,\hat I} \le C\, \varepsilon^{1/r}\, \|\hat u - u\|_{\mathcal{V}}$$

with $C > 0$ independent of $\hat u \in \hat V$ and ε. In view of (5.29) the proof is complete by letting ε tend to zero. ∎

5.4. Differentiability on the boundary of rarefaction waves (Case RB^c). Finally, we consider continuity points $(\bar t, \bar x)$ of class RB^c, i.e., outside the shock set and on the boundary of a rarefaction wave. If $(\bar t, \bar x)$ is for concreteness on the left boundary of a rarefaction wave, then the backward characteristic meets $t = 0$ in a discontinuity $\bar z = x_i$ of u_0 with $u_0(x_i-) < u_0(x_i+)$, and (4.8), (4.9) hold with $\bar w = u_0(x_i-)$. Finally (4.16) holds for the left-sided, (4.19) for the right-sided derivative, respectively. Hence, using Notation 5.7 we get

$$(Y_r, W, V_1, S_+(\tau), J_w) = \mathrm{YR}(u, \bar t, \bar z, [w_-, w_+])$$

by Lemma 5.6 for suitable $w_- < \bar w < w_+$ and with Notation 5.2

$$(Y_-, Z_-, V_-, S_-(\tau), J) = \mathrm{YC}(u_0^-, u_1, [z_-, z_+])$$

by Lemma 5.1 for suitable $z_- < \bar{z} < z_+$, where u_0^- is the C^1-prolongation (5.25). For V according to (5.26) with $\delta > 0$ small enough we can achieve that $\hat{u} \in V$ implies $\hat{u}_1 \in V_1$ and $(\hat{u}_0^-, \hat{u}_1) \in V_-$. Now it is obvious that $Y_-(\cdot, \hat{u})$ and $Y_r(\cdot, \hat{u}_1)$ can be glued together continuously along the forward characteristic $\hat{\xi}(t) \stackrel{\text{def}}{=} \zeta(t; \bar{z}, \hat{u}_0(\bar{z}-), \hat{u}_1)$ for all $\hat{u} \in V$, yielding a classical solution on $(S_- \cap \{x \le \hat{\xi}(t)\}) \cup (S_+ \cap \{x > \hat{\xi}(t)\})$. The next lemma shows that this function coincides locally with $y(\cdot; \hat{u})$.

LEMMA 5.12 (differentiability properties in continuity points of class RB^c). *Let* (A2)–(A3) *hold and let* $(\bar{t}, \bar{x})$ *be a continuity point of* $y(\cdot; u)$ *outside the shock set on the boundary of a rarefaction wave (Case* RB^c*), i.e., the backward characteristic meets* $t = 0$ *in* $\bar{z} = x_i$ *with* $u_0(\bar{z}-) < u_0(\bar{z}+)$*, and* (4.8), (4.9) *hold with* $\bar{w} = u_0(\bar{z}-)$ *or* $\bar{w} = u_0(\bar{z}+)$.

Given $M_\infty > 0$ *there are* $\delta > 0$, $R > 0$, *and* $\nu > 0$ *such that, with* V *according to* (5.26) *and* $\tilde{V}$ *as in Lemma* 5.11, *the following holds:*

$\hat{\xi} \stackrel{\text{def}}{=} \zeta(\cdot; \bar{z}, \hat{u}(\bar{z}-), \hat{u}_1)$ *is on* $[0, \bar{t}]$ *for all* $\hat{u} \in \hat{V}$ *a genuine forward characteristic of* $y(\cdot; \hat{u})$ *through* $(0, \bar{z})$. *Furthermore:*

(i) *If* $\bar{w} = u_0(\bar{z}-)$, *then let* $(Y_-, Z_-, V_-, S_-(\tau), J) = \mathrm{YC}(u_0^-, u_1, [z_-, z_+])$ *as well as* $(Y_r, W, V_1, S_+(\tau), J_w) = \mathrm{YR}(u, \bar{t}, \bar{z}, [w_-, w_+])$ *be defined as explained above. Then there is a neighborhood* $\hat{I} =]x_-, x_+[$ *of* $\bar{x}$ *such that for all* $\hat{u} \in \hat{V}$

$$y(\bar{t}, x; \hat{u}) \stackrel{\text{def}}{=} \begin{cases} Y_-(\bar{t}, x, \hat{u}_0^-, \hat{u}_1), & x \in]x_-, \hat{\xi}(\bar{t})[\,, \\ Y_r(\bar{t}, x, \hat{u}_1), & x \in]\hat{\xi}(\bar{t}), x_+[\end{cases} \tag{5.30}$$

holds. If $\hat{V}$ *is equipped with the norm*

$$\|(\hat{u}_0, \hat{u}_1)\|_{\mathcal{V}} \stackrel{\text{def}}{=} \|\hat{u}_0\|_{C^1(J_-)} + \|\hat{u}_1\|_{L^\infty(0,T;C^1(\mathbb{R}))},$$

where $J_- \stackrel{\text{def}}{=} J \cap \{z < \bar{z}\}$, *then*

$$\hat{u} \in (\hat{V}, \|\cdot\|_{\mathcal{V}}) \longmapsto y(\bar{t}, \cdot; \hat{u}) \in L^r(\hat{I}) \tag{5.31}$$

is continuously Fréchet differentiable for all $r \in [1, \infty[$ *with derivative*

$$\begin{aligned} d_u y(\bar{t}, \cdot; \hat{u}) \cdot (\delta u_0, \delta u_1) = {} & \mathbf{1}_{]x_-, \hat{\xi}(\bar{t})[}(d_u Y_-(\bar{t}, \cdot, \hat{u}_0^-, \hat{u}_1) \cdot (\delta u_0|_{J_-}, \delta u_1)) \\ & + \mathbf{1}_{]\hat{\xi}(\bar{t}), x_+[}(d_{u_1} Y_r(\bar{t}, \cdot, \hat{u}_1) \cdot \delta u_1). \end{aligned}$$

(ii) *If* $\bar{w} = u_0(\bar{z}+)$, *a completely analogous result holds with left state* Y_r *and right state* Y_+ *obtained from* $(Y_+, Z_+, V_+, S_+(\tau), J) = \mathrm{YC}(u_0^+, u_1, [z_-, z_+])$.

Proof. We only sketch the proof, since it is similar to the proof of Lemma 5.11. First, we obtain that (5.30) holds for u and an appropriate neighborhood $\hat{I}$ of $\bar{x}$. For given $M_\infty > 0$ and $\varepsilon > 0$, we conclude, by applying Lemmas 5.5 and 5.10, that (5.30) holds also on $\hat{I} \cap \{|x - \bar{x}| > \varepsilon/2\}$ for all $\hat{u} \in \hat{V}$ with $\hat{V}$ as in Lemma 5.11 and V as in (5.26). Hence, $y(\bar{t}, x; \hat{u})$ for $x \in [\bar{x} - \varepsilon, \bar{x} + \varepsilon]$ depends only on the values of $\hat{u}_0$ on J_-. But gluing together $Y_-(\cdot, \hat{u})$ and $Y_r(\cdot, \hat{u}_1)$ along $\hat{\xi}$ yields a classical solution that is compatible with the initial data $\hat{u}_0|_{J_-}$ and coincides with $y(\cdot; \hat{u})$ along all genuine backward characteristics through $(\bar{t}, x)$, $x \in \hat{I} \setminus [\bar{x} - \varepsilon, \bar{x} + \varepsilon]$. Hence, (5.30) holds by the uniqueness of entropy solutions.

From (5.30) the continuous differentiability of (5.31) is obvious, since by Lemmas 5.1 and 5.6 the mappings $\hat{u} \in (\hat{V}, \|\cdot\|_{\mathcal{V}}) \longmapsto (Y_-(\bar{t}, \cdot; \hat{u}_0^-, \hat{u}_1), Y_r(\bar{t}, \cdot; \hat{u}_1)) \in C(\hat{I})^2$ are continuously Fréchet differentiable, the functions on the right remain bounded in $C^1(\hat{I})$ on $\hat{V}$, and $\hat{\xi}(\bar{t})$ depends Lipschitz continuously on $\hat{u}$. The formula for the derivative is obvious. ∎

6. Stability of shocks and differentiability of shock position. We turn now to the study of points $(\bar{t},\bar{x})$ in the shock set. Hereby, the following nondegeneracy of shocks will be important.

DEFINITION 6.1 (nondegeneracy of shocks). *Let $u_0 \in PC^1(\mathbb{R};x_1,\dots,x_n)$. A point $(\bar{t},\bar{x}) \in \Omega_T^{cl}$ of the shock set is called* nondegenerate *if $(\bar{t},\bar{x})$ is not a shock interaction point and moreover is of type C^cC^c, R^cR^c, C^cR^c, or R^cC^c according to section* 4.4.

We will show that for a nondegenerate shock point $(\bar{t},\bar{x})$ the following holds for quite general variations $(\hat{u}_0,\hat{u}_1)$ of (u_0,u_1):

- The shock is stable and separates states that coincide locally with representations Y or Y_r obtained by Lemma 5.1 or 5.6, respectively.

Moreover, if $g(t,x,y,u)$ is affine linear w.r.t. y, then the following hold:

- The shock position depends Fréchet differentiably on $(\hat{u}_0,\hat{u}_1)$.
- The sensitivity of the shock position w.r.t. $(\hat{u}_0,\hat{u}_1)$ can be explicitly computed by the adjoint formula (3.8).

6.1. Stability and structure of shocks. The nonlinear stability of shock waves for a single convex conservation law (i.e., $f''>0$, $g\equiv 0$) under variations of the initial data in the Schwartz space $\mathcal{S}$ was shown by Golubitsky and Schaeffer [14]. However, there seem to be no results in the literature on the nonlinear stability of shocks under perturbations of initial data and source term that are directly applicable in our framework. In the following lemma we give a stability result of shock waves based on generalized characteristics that fits our purposes.

For concreteness we consider a shock point $(\bar{t},\bar{x})$ of class $\mathrm{C}^c\mathrm{C}^c$ (see section 4.4).

LEMMA 6.2 (stability of nondegenerate shocks of type $\mathrm{C}^c\mathrm{C}^c$). *Let* (A2)–(A3) *hold, let $u=(u_0,u_1)\in PC^1(\mathbb{R};x_1,\dots,x_N)\times L^\infty(0,T;C^1(\mathbb{R})^m)$, and let $(\bar{t},\bar{x})\in\Omega_T$ be a point of discontinuity of $y(.;u)$ on a shock curve $\eta(t)$. Denote by $\xi_\mp(t)$ the minimal and maximal backward characteristic through $(\bar{t},\bar{x})$ and set $\bar{z}^\mp=\xi_\mp(0)$. Assume that $(\bar{t},\bar{x})$ is a nondegenerate shock point of class C^cC^c according to Definition* 6.1*, i.e., $(\bar{t},\bar{x})$ is not a shock interaction point, $\bar{z}^\mp\neq x_i$, $1\le i\le N$, and* (4.16) *holds with $\bar{z}=\bar{z}^\mp$ for some $\beta>0$. Then there are $z_-^\mp<\bar{z}^\mp<z_+^\mp$, $\rho>0$, and $\beta>0$ such that* (5.1) *holds.*

Using Notation 5.2 *let $(Y_\mp,Z_\mp,V_\mp,S_\mp(\tau),J_\mp)=\mathrm{YC}(u,\bar{t},[z_-^\mp,z_+^\mp])$ be obtained by Lemma* 5.1 *with $z_\mp^-$, $z_\mp^+$ close enough to $\bar{z}^\mp$, respectively. After a possible reduction of $V_\mp$, $\tau>0$ there is a neighborhood $I=]x_-,x_+[$ of $\bar{x}$ such that the following hold:*

(i) *$y(\cdot;u)$ is locally given by*

$$y(t,x;u)=\begin{cases} Y_-(t,x,u), & (t,x)\in S_-\cap\{x<\eta(t)\},\\ Y_+(t,x,u), & (t,x)\in S_+\cap\{x>\eta(t)\}.\end{cases}$$

(ii) *Let $M_\infty>0$ be given. Then there are $R>0$, $\nu>0$ such that for*

$$\hat{V}=\{(\hat{u}_0,\hat{u}_1)\in L^\infty(\mathbb{R})\times L^\infty(0,T;C^1(\mathbb{R})^m):(\hat{u}_0,\hat{u}_1)\in V_-\cap V_+,\ \|\hat{u}_0\|_\infty<M_\infty,\ \|\hat{u}_0-u_0\|_{1,[-R,R]}<\nu\}$$

equipped with the norm

$$\|(\hat{u}_0,\hat{u}_1)\|_{\mathcal{V}}\overset{\text{def}}{=}\|\hat{u}_0\|_{C^1(J_-\cup J_+)}+\|\hat{u}_0\|_{1,[\bar{z}^-,\bar{z}^+]}+\|\hat{u}_1\|_{L^\infty(0,T;C^1(\mathbb{R}))},$$

there is a Lipschitz continuous function

$$x_s:\hat{u}\in(\hat{V},\|\cdot\|_{\mathcal{V}})\longmapsto x_s(\hat{u}) \tag{6.1}$$

with $x_s(u) = \bar{x}$ such that for all $\hat{u} = (\hat{u}_0, \hat{u}_1) \in \hat{V}$ the following holds:

$$y(\bar{t}, x; \hat{u}) = \begin{cases} Y_-(\bar{t}, x, \hat{u}_0|_{J_-}, \hat{u}_1), & x \in]x_-, x_s(\hat{u})[\,, \\ Y_+(\bar{t}, x, \hat{u}_0|_{J_+}, \hat{u}_1), & x \in]x_s(\hat{u}), x_+[\,. \end{cases} \tag{6.2}$$

Hereby, $\hat{u} \in (\hat{V}, \|\cdot\|_{\mathcal{V}}) \longmapsto Y_\mp(\bar{t}, \cdot, \hat{u}_0|_{J_\mp}, \hat{u}_1) \in C(I)$ are continuously Fréchet differentiable.

REMARK 6.3. *We note that all results so far require only* (A2)–(A3) *but not* (A4). *See also Remark* 8.1.

Proof. $(\bar{t}, \bar{x})$ is a point of discontinuity on a shock curve η. Thus, we have by Oleinik's entropy condition (4.1) that $y(\bar{t}, \eta(\bar{t})-; u) > y(\bar{t}, \eta(\bar{t})+; u)$. Since (4.16) holds for $\bar{z} = \bar{z}^\mp$ we obtain by continuity $\beta > 0$, $J_\mp$ such that (5.1) holds. Now Lemma 5.1 yields $(Y_\mp, Z_\mp, V_\mp, S_\mp(\tau), J_\mp) = \mathrm{YC}(u, \bar{t}, [z_-^\mp, z_+^\mp])$ as asserted. By Lemma 5.1 and the definition of $\hat{V}$ in (ii)

$$(\hat{u}_0, \hat{u}_1) \in (\hat{V}, \|\cdot\|_{\mathcal{V}}) \longmapsto Y_\mp(\bar{t}, \cdot, \hat{u}_0|_{J_\mp}, \hat{u}_1) \in C(I)$$

are obviously continuously Fréchet differentiable.

(i) Since $(\bar{t}, \bar{x})$ is not a shock interaction point, we know from [8] that $\bar{t}$ is a continuity point of $t \longmapsto y(t, \eta(t)\pm; u)$. Then by [8, Thm. 4.5] $(\bar{t}, \bar{x})$ is a continuity point of $y(\cdot; u)$ relative to the sets $\mathcal{L}_- = \{(t, x) : x < \eta(t)\}$, $\mathcal{L}_+ = \{(t, x) : x > \eta(t)\}$, the limit being $y(\bar{t}, \eta(\bar{t})\mp; u)$. The extreme backward characteristics $\xi_\mp(t)$ through $(\bar{t}, \bar{x})$ satisfy (4.3) and $\xi_\mp(0) = \bar{z}_\mp \in J_\mp$. Using the backward stability of solutions of (4.3) we find $\delta > 0$ such that the genuine backward characteristic through any continuity point in $Q_\mp \stackrel{\text{def}}{=} \mathcal{L}_\mp \cap \{\|(t, x) - (\bar{t}, \bar{x})\|_\infty \le \delta\}$ meets $t = 0$ in $J_\mp$, respectively. Hence, we must have $y(t, x; u) = Y_\mp(t, x, u)$ for any continuity point in $Q_\mp$, and since the continuity points are dense, for all points in $Q_\mp$ by our convention $y(t, x) = y(t, x-)$. The same is clearly true on the domain covered by these backward characteristics. Possibly after reducing $\tau > 0$ and choosing $z_\mp^-$, $z_\mp^+$ closer to $\bar{z}_\mp$, respectively, we achieve that $S_\mp \cap \mathcal{L}_\mp$ is covered by the backward characteristics through $Q_\mp$. Hence, (i) is proven.

(ii) For notational convenience we write y instead of $y(\cdot; u)$. By the definition of $S_\mp$ it is clear that for sufficiently small $\tau > 0$, $\sigma > 0$ and $x_\mp = \bar{x} \mp \sigma$ one has

$$Q \stackrel{\text{def}}{=}]\bar{t} - \tau, \bar{t} + \tau[\times]x_-, x_+[\subset S_- \cap S_+.$$

Set $I \stackrel{\text{def}}{=}]x_-, x_+[$. As in the proof of Lemma 5.5 there is $M_y > 0$ with $\|y(\cdot; \hat{u})\|_\infty \le M_y$ for all $\hat{u} \in \hat{V}$, and we may choose M_y such that in addition $|Y_\mp(\cdot, \hat{u})| \le M_y$ on $S_\mp$.

By continuity, we may reduce τ such that for some $\Delta > 0$

$$y(t, \eta(t)-) - y(t, \eta(t)+) \ge 2\Delta > 0 \quad \forall\, t \in]\bar{t} - \tau, \bar{t} + \tau[.$$

Since $f'' \ge m_{f''} > 0$, we have for $|t - \bar{t}| < \tau$

$$\dot{\eta}(t) = \frac{f(y(t, \eta(t)-)) - f(y(t, \eta(t)+))}{y(t, \eta(t)-) - y(t, \eta(t)+)} \begin{cases} \le f'(y(t, \eta(t)-)) - m_{f''}\Delta, \\ \ge f'(y(t, \eta(t)+)) + m_{f''}\Delta. \end{cases}$$

In particular, $\dot{\eta}(t)$ is continuous in $\bar{t}$, and we may reduce $V_\mp$, ν, τ, σ such that with $\varepsilon > 0$ the following hold:

$$f'(Y_-(t, x, \hat{u})) \ge \max_{|t-\bar{t}|<\tau} \dot{\eta}(t) + \varepsilon, \tag{6.3}$$

$$Y_-(t, x, \hat{u}) - Y_+(t, x, \hat{u}) \ge \Delta \tag{6.4}$$

for all $(t, x, \hat{u}) \in Q \times \hat{V}$. Finally, reduce $\tau > 0$ such that $\eta(t) \in I$ and

$$|\dot{\eta}(t) - \max_{|t-\bar{t}|<\tau} \dot{\eta}(t)| < \frac{m_{f''}\varepsilon}{8M_{f''}} =: \hat{\varepsilon} \tag{6.5}$$

for all t, $|t - \bar{t}| < \tau$, where $M_{f''} \stackrel{\text{def}}{=} \sup_{|y|\le M_y} f''(y)$.

The following property of genuine backward characteristics will be crucial: Let $\zeta(t)$ be a backward characteristic such that

$$|\zeta(\hat{t}) - \eta(\hat{t})| < \delta, \quad \dot{\zeta}(\hat{t}) < \dot{\eta}(\hat{t}) - 2\hat{\varepsilon} \quad \text{for some } \hat{t} \in [\bar{t} - \tau/2, \bar{t}]. \tag{6.6}$$

From (4.3) we get a constant $M > 0$ with

$$|\ddot{\zeta}(t)| \le M,$$

and we can reduce $\tau > 0$ such that $M\tau < \hat{\varepsilon}/2$. Then

$$\zeta(t) - \eta(t) > \delta \quad \text{for } t = \bar{t} - \tau \text{ if } \delta < \hat{\varepsilon}\tau/8.$$

In fact, by using (6.5) we get the following with $s = \min_{|t-\bar{t}|<\tau} \dot{\eta}(t)$ and $t = \bar{t} - \tau$:

$$\begin{aligned}
\zeta(t) - \eta(t) &\ge \zeta(\hat{t}) - \eta(\hat{t}) + (s - \dot{\zeta}(\hat{t}))(\hat{t} - t) - \tfrac{M}{2}(\hat{t} - t)^2 \\
&\ge -\delta + (s - \dot{\eta}(\hat{t}) + 2\hat{\varepsilon})(\hat{t} - t) - \tfrac{M}{2}(\hat{t} - t)^2 \\
&\ge -\delta + (-\hat{\varepsilon} + 2\hat{\varepsilon} - M\tau)\frac{\tau}{2} \ge \frac{\hat{\varepsilon}\tau}{4} - \delta > \delta.
\end{aligned}$$

Fix these τ and δ and set $Q_\delta \stackrel{\text{def}}{=} Q \cap \{(t, x) : |x - \eta(t)| < \delta\}$. Using a finite covering of $(Q^{cl} \cap \mathcal{L}_\mp) \setminus Q_\delta$ with the stripes S obtained by Lemma 5.5 for varying t, we may reduce $V_\mp$, ν such that for all $\hat{u} \in \hat{V}$

$$y(t, x; \hat{u}) = Y_\mp(t, x, \hat{u}_0|_{J_\mp}, \hat{u}_1) \quad \forall\, (t, x) \in (Q \cap \mathcal{L}_\mp) \setminus Q_\delta, \tag{6.7}$$

respectively. For the rest of the proof it is convenient to set $Y_\mp(\cdot, \hat{u}) = Y_\mp(\cdot, \hat{u}_0|_{J_\mp}, \hat{u}_1)$ for $\hat{u} \in \hat{V}$ on $S_\mp$.

Assume that (6.2) does not hold for all $\hat{u} \in \hat{V}$. Then there is $\hat{u} \in \hat{V}$ and without restriction a continuity point $\hat{x}$, $|\hat{x} - \bar{x}| < \delta$ with $y(\bar{t}, \hat{x}; \hat{u}) \ne Y_\mp(\bar{t}, \hat{x}, \hat{u})$ and thus $\hat{\xi}(0) < \inf J_+$ for the genuine backward characteristic $\hat{\xi}$ through $(\bar{t}, \hat{x})$. For convenience we set $\hat{y} \stackrel{\text{def}}{=} y(\cdot; \hat{u})$. Let $\tilde{x}$ be the infimum of these $\hat{x}$. $(\bar{t}, \tilde{x})$ cannot be a continuity point, since otherwise by continuity $\tilde{\xi}(0) \in J_-$ for the backward characteristic $\tilde{\xi}$ through $(\bar{t}, \tilde{x})$, and thus by Lemma 5.5 $\hat{y}(\bar{t}, x) = Y_-(\bar{t}, x, \hat{u})$ for x close enough to $\tilde{x}$, which is a contradiction. Thus, $(\bar{t}, \tilde{x})$ lies on a shock. We set $\tilde{\eta}(\bar{t}) = \tilde{x}$. By construction the maximal backward characteristic $\tilde{\xi}_+$ through $(\bar{t}, \tilde{x})$ satisfies $\sup J_- < \tilde{\xi}_+(0) < \inf J_+$. Since genuine characteristics may intersect only at their end points, we have by (6.7) necessarily that $(t, \tilde{\xi}_+(t)) \in Q_\delta$ for $\bar{t} - \tau < t \le \bar{t}$. Moreover, it is obvious from $\sup J_- < \tilde{\xi}_+(0) < \inf J_+$ that $\hat{y}(t, \tilde{\xi}_+(t)) \ne Y_\mp(t, \tilde{\xi}_+(t), \hat{u})$ for $\bar{t} - \tau < t \le \bar{t}$. Thus, we find as above for all these t a discontinuity point $\tilde{\eta}(t)$ with left state $Y_-(t, \tilde{\eta}(t), \hat{u})$. Now the unique forward characteristic through $(t - \tau, \tilde{\eta}(t - \tau))$ must be a shock, and it must coincide with $\tilde{\eta}$, since every $(t, \tilde{\eta}(t))$ lies on a shock with left state $Y_-(t, \tilde{\eta}(t), \hat{u})$ and therefore cannot be a shock generation point. Hence, $\tilde{\eta}(t)$ is a shock with left state $Y_-(t, \tilde{\eta}(t), \hat{u})$, and we must have $|\tilde{\eta}(t) - \eta(t)| < \delta$ for $\bar{t} - \tau < t \le \bar{t}$, because $\tilde{\eta}$ may not enter the domain of continuity. Hence, there must exist $\hat{t} \in [\bar{t} - \tau/2, \bar{t}]$ with

$$\dot{\tilde{\eta}}(\hat{t}) \le \max_{|t-\bar{t}|<\tau} \dot{\eta}(t) + \hat{\varepsilon} \tag{6.8}$$

because $\hat{\varepsilon}\tau/2 > 2\delta$ by the choice of δ. Set $y_\mp = \hat{y}(\hat{t}, \tilde{\eta}(t)\mp)$. We show that necessarily $f'(y_+) \le s - 2\hat{\varepsilon}$ with $s = \min_{|t-\bar{t}|<\tau} \dot{\eta}(t)$ as defined above. If not, then we have by (6.3)

$$h \stackrel{\text{def}}{=} f'(y_-) - f'(y_+) \ge s + \varepsilon - (s - 2\hat{\varepsilon}) \ge \varepsilon + 2\hat{\varepsilon} > \varepsilon.$$

Moreover, $y_- - y_+ \ge \frac{h}{M_{f''}}$, and thus with $f'' \ge m_{f''}$ and (6.5)

$$\dot{\tilde{\eta}}(\hat{t}) = \frac{f(y_-) - f(y_+)}{y_- - y_+} \ge f'(y_+) + \frac{m_{f''} h}{2M_{f''}} > s - 2\hat{\varepsilon} + \frac{m_{f''} h}{2M_{f''}} > \max_{|t-\bar{t}|<\tau} \dot{\eta}(t) - 3\hat{\varepsilon} + 4\hat{\varepsilon}.$$

This contradicts (6.8). We conclude that the maximal backward characteristic ζ through $(\hat{t}, \tilde{\eta}(\hat{t}))$ satisfies the scenario (6.6), since $\dot{\zeta}(\hat{t}) = f'(y_+) \le s - 2\hat{\varepsilon}$. Therefore, it would hold $\zeta(\hat{t} - \tau) - \eta(\hat{t} - \tau) > \delta$. This is a contradiction, since then ζ would intersect $\tilde{\xi}_+$. Hence, the assumption was wrong and (6.2) is shown.

It remains to show that $\hat{u} \in (\hat{V}, \|\cdot\|_{\mathcal{V}})) \mapsto x_s(\hat{u})$ is Lipschitz. But this follows directly from (6.2), (6.4) and the L^1-stability according to Theorem 4.1(ii). The continuous differentiability of $\hat{u} \longmapsto Y_\mp(\bar{t}, \cdot, \hat{u}) \in C(I)$ was already shown at the beginning of the proof. ∎

Very similar results can be obtained if left and/or right states are rarefaction waves, i.e., in the Cases $\mathrm{C}^c\mathrm{R}^c$, $\mathrm{R}^c\mathrm{C}^c$, $\mathrm{R}^c\mathrm{R}^c$. See Corollary 6.5 below.

6.2. Differentiability of the shock position and an adjoint formula. We come now to the key point of our analysis. In this subsection we state a differentiability result for the shock position x_s of Lemma 6.2 and derive an explicit formula for the derivative by using an appropriate adjoint state. The advantage of the adjoint approach lies in the fact that we do not have to impose regularity assumptions on y for $t < \bar{t}$.

LEMMA 6.4 (differentiability of shock position and adjoint formula, Case $\mathrm{C}^c\mathrm{C}^c$). *With the assumptions and notations of Lemma* 6.2 *let* $J_0 =]\bar{z}_-, \bar{z}_+[\setminus (J_- \cup J_+)$ *and let* $J_s \subset J_0$ *be an open set that contains only down-jumps of* u_0. *Finally, let* $J \subset J_s$ *be an open set with* $J^{cl} \subset J_s$ *and set* $J_r = J_0 \setminus J$. *Define with* $\hat{V}$ *from Lemma* 6.2

$$\tilde{V} \stackrel{\text{def}}{=} \{(\hat{u}_0, \hat{u}_1) \in \hat{V} \ : \ \partial_x(\hat{u}_0 - u_0)|_{J_s} \le M_L\}$$

for some fixed $M_L > 0$ *equipped with the norm*

$$\|(\hat{u}_0, \hat{u}_1)\|_{\tilde{\mathcal{V}}} \stackrel{\text{def}}{=} \|\hat{u}_0\|_{C^1(J_- \cup J_+)} + \|\hat{u}_0\|_{\infty, J_r} + \|\hat{u}_0\|_{1,J} + \|\hat{u}_1\|_{L^\infty(0,T;C^1(\mathbb{R}))}.$$

Then the following hold:

(i) *Assume that* (A4) *holds, i.e.,* g *is affine linear w.r.t.* y. *Then the shock position*

$$x_s : \hat{u} \in (\tilde{V}, \|\cdot\|_{\tilde{\mathcal{V}}}) \longmapsto x_s(\hat{u})$$

is continuously Fréchet differentiable and

$$d_u x_s(\hat{u}) \cdot (\delta u_0, \delta u_1) = (p(0, \cdot), \delta u_0)_2 + (p\, g_{u_1}(\cdot, \hat{y}, \hat{u}_1), \delta u_1)_{2,\Omega_{\bar{t}}}, \tag{6.9}$$

where p *is the reversible solution of the adjoint equation*

$$\partial_t p + f'(\hat{y})\partial_x p = -g_y(t, x, \hat{y}, \hat{u}_1)p, \quad p(\bar{t}, \cdot) = p^{\bar{t}}, \tag{6.10}$$

$$p(\bar{t}, \cdot) = \frac{1}{[y(\bar{t}, x_s(\hat{u}); \hat{u})]}\, \mathbf{1}_{\{x = x_s(\hat{u})\}}(\cdot). \tag{6.11}$$

Denoting by $\hat{\xi}_{\mp}$ *the minimal/maximal backward characteristic of* $\hat{y} = y(\cdot;\hat{u})$ *through* $(\bar{t}, x_s(\hat{u}))$ *and by* D *the domain confined by them, one has* $p = \hat{p}\,\mathbf{1}_{D^{cl}}(\cdot)$, *where* $\hat{p} \in C^{0,1}_{loc}((0,\bar{t})\times\mathbb{R})$ *is the reversible solution of* (6.10) *for the constant end data* $\hat{p}(\bar{t},\cdot) = 1/[y(\bar{t}, x_s(\hat{u}); \hat{u})]$.

(ii) *If* g *does not depend on* y, *then the same holds for* $x_s : \hat{u} \in (\hat{V}, \|\cdot\|_{\mathcal{V}}) \mapsto x_s(\hat{u})$ *in* (6.1), *and the reversible solution of equation* (6.10) *is simply* $p = 1/[y(\bar{t}, x_s(\hat{u}); \hat{u})]\,\mathbf{1}_{D^{cl}}(\cdot)$.

The proof of this lemma will be postponed to section 8, since part (i) requires an existence and stability result for backward transport equations of the form (6.10) that will be provided in the following section. The essential feature of the adjoint equation is that the coefficient $f'(\hat{y})$ is discontinuous but satisfies by $f'' > 0$ and Oleinik's entropy condition (4.1) a one-sided Lipschitz condition of the form

$$\partial_x f'(\hat{y}(t,\cdot)) \le \frac{C}{t + 1/M}, \quad t \in]0, T[\,,$$

where $M \in [M_{cr}, \infty]$ is such that $\partial_x \hat{u}_0 \le M$.

If the left and/or right state of the shock is a rarefaction wave, then an analogue of Lemmas 6.2 and 6.4 holds.

COROLLARY 6.5 (differentiability of shock position, Cases $\mathrm{R}^c\mathrm{C}^c$, $\mathrm{C}^c\mathrm{R}^c$, $\mathrm{R}^c\mathrm{R}^c$). *Let* $(\bar{t}, \bar{x})$ *be of class* $R^c C^c$, *i.e., with the notations of Lemmas* 6.2 *and* 6.4 $\bar{z}^- = x_i$, $i \in \{1, \dots, N\}$, *holds,* $u_0(x_i-) < \bar{w} < u_0(x_i+)$ *with* $\bar{w}$ *according to* (4.8), *and* (4.19) *holds for the left-sided derivative with* $\bar{z} = \bar{z}^-$, *while the right state is as in Lemmas* 6.2 *and* 6.4.

Using Notation 5.7 *let* $(Y_-, W, V_1, S_-(\tau), J_w) = \mathrm{YR}(u, \bar{t}, \bar{z}^-, [w_-, w_+])$ *according to Lemma* 5.6 *with appropriate* $w_- < \bar{w} < w_+$. *Then the results of Lemmas* 6.2 *and* 6.4 *remain true if we set* $J_- = \emptyset$ *and define* V_- *for an arbitrary open interval* J *containing* $\bar{z}^-$ *and* $\nu > 0$ *sufficiently small as*

$$V_- \stackrel{\text{def}}{=} \{\hat{u}_0 \in L^\infty(J) \,:\, \|\hat{u}_0 - u_0\|_\infty < \nu\} \times V_1.$$

An analogue result holds if the right state is a rarefaction wave, i.e., $\bar{z}^+ = x_i$, $u_0(x_i-) < \bar{w} < u_0(x_i+)$.

7. The adjoint equation. We consider the backward problem for a transport equation

$$\partial_t p + a\partial_x p = -bp, \quad p(T,\cdot) = p^T \tag{7.1}$$

with $b \in L^\infty(0,T; C^{0,1}(\mathbb{R}))$ and $a \in L^\infty(\Omega_T)$ satisfying the one-sided Lipschitz condition

$$a_x(t,\cdot) \le \alpha(t), \quad \alpha \in L^1(0,T), \tag{7.2}$$

or at least the weakened one-sided Lipschitz condition

$$a_x(t,\cdot) \le \alpha(t), \quad \alpha \in L^1(\sigma,T) \quad \forall\, \sigma \in (0,T). \tag{7.3}$$

The latter case is appropriate by Oleinik's entropy condition (4.1) if we want to consider the adjoint equation for solutions with rarefaction waves (generated by up-jumps of u_0). It is well known [7] that under this one-sided Lipschitz condition on a for any $p^T \in \mathrm{Lip}(\mathbb{R})$ there exists a Lipschitz continuous solution to (7.1) which is not

necessarily unique. For the case $b \equiv 0$ the stability of p w.r.t. a for the special class of *reversible solutions* was extensively studied by Bouchut and James in the recent paper [1]. We will use the results of [1] to study the stability w.r.t. a and b. Due to space limitations we will restrict ourselves to summarizing the definition of *reversible solutions* for (7.1) and the necessary existence, stability, and regularity properties.[1] A detailed analysis is given in the follow-up paper [27]. Our results extend the stability results in [1] to the case $b \neq 0$, which is not straightforward, since the definition of reversible solutions in [1] cannot be directly used.

Denote by $\mathcal{L}_h$ the space of Lipschitz continuous solutions to

$$\partial_t p + a\partial_x p = 0, \quad (t,x) \in \Omega_T. \tag{7.4}$$

We recall the following definition of [1].

DEFINITION 7.1. *$p \in \mathcal{L}_h$ is called a* reversible solution *of* (7.4) *if there exist $p_1, p_2 \in \mathcal{L}_h$ such that $\partial_x p_1 \geq 0$, $\partial_x p_2 \geq 0$, and $p = p_1 - p_2$.*

Then the following holds [1].

THEOREM 7.2. *Let $a \in L^\infty(\Omega_T)$ satisfy the one-sided Lipschitz condition* (7.2)*, i.e., $a_x(t,\cdot) \leq \alpha(t)$ with $\alpha \in L^1(0,T)$. Then for any $p^T \in \mathrm{Lip}(\mathbb{R})$ there exists a unique reversible solution $p \in C^{0,1}_{loc}(\Omega_T^{cl})$ of* (7.4) *with $p(T,\cdot) = p^T$. Moreover,*

$$\|p(t,\cdot)\|_{\infty,I} \leq \|p^T\|_{\infty,J}, \quad \|\partial_x p(t,\cdot)\|_{\infty,I} \leq e^{\int_t^T \alpha} \|\partial_x p^T\|_{\infty,J}$$

with $I =]x_1, x_2[$, $J =]x_1 - \|a\|_\infty (T-t), x_2 + \|a\|_\infty (T-t)[$.

This concept of reversible solution is not directly extendible to the case $b \neq 0$. However, a natural extension can be obtained by using the generalized backward flow associated with a defined in [1].

DEFINITION 7.3. *Let $D_b = \{(s,t) \in \mathbb{R}^2 : 0 \leq t \leq s \leq T\}$. Then the generalized backward flow $X : D_b \times \mathbb{R} \longrightarrow \mathbb{R}$ is defined by the requirement that for any $s \in]0,T]$ $X(s;\cdot,\cdot)$ is the unique reversible solution to*

$$\partial_t X + a\partial_x X = 0, \quad (t,x) \in]0,s[\times \mathbb{R}, \quad X(s;s,x) = x, \quad x \in \mathbb{R}.$$

Moreover, we set $X(0;0,x) = x$.

One can show [1] that $\partial_x X \geq 0$ and $x \longmapsto X(s;t,x)$ is surjective for all $(s,t) \in D_b$ and that for all $0 \leq t \leq \sigma \leq s \leq T$ and $x \in \mathbb{R}$ the composition formula

$$X(s;\sigma, X(\sigma;t,x)) = X(s;t,x) \tag{7.5}$$

is satisfied. Moreover, for all $(t,x) \in \Omega_T$ the following holds [1]:

$$\partial_s X(s;t,x) \in [a(s, X(s;t,x)+), a(s, X(s;t,x)-)] \quad \text{for a.a. } s \in]t,T[. \tag{7.6}$$

Note that by the one-sided Lipschitz condition $\partial_x a(t,\cdot) \leq \alpha(t)$, $\alpha \in L^1(0,T)$, the following holds: $a(t,\cdot) \in BV_{loc}(\mathbb{R})$ for a.a. t. Thus the left- and right-sided limits in (7.6) exist.

REMARK 7.4. *Since $X(t;t,x) = x$, $X(\cdot;t,x)$ is by* (7.6) *the unique forward characteristic through (t,x) in the sense of Filippov. Thus, if y is the entropy solution of* (1.1) *and $a = f'(y)$, then $X(\cdot;t,x)$ is the generalized forward characteristic through (t,x).*

[1] For convenience, the proofs are added in the appendix.

Finally, it is shown in [1] that for $b \equiv 0$ the reversible solution of (7.1) is given by $p(t,x) = p^T(X(T;t,x))$ and is thus the broad solution along the generalized characteristics. This leads us to the following definition of reversible solutions for (7.1).

DEFINITION 7.5. *Denote by $B(\mathbb{R})$ the Banach space of bounded functions equipped with the* sup*-norm and let*

$$B_{\mathrm{Lip}}(\mathbb{R}) \stackrel{\mathrm{def}}{=} \left\{ w \in B(\mathbb{R}) \colon \begin{array}{l} w \text{ is the pointwise everywhere limit of a sequence} \\ (w_n) \text{ in } C^{0,1}(\mathbb{R}),\ (w_n) \text{ bounded in } C(\mathbb{R}) \cap H^{1,1}_{loc}(\mathbb{R}) \end{array} \right\}.$$

Let $a \in L^\infty(\Omega_T)$, $\partial_x a(t,\cdot) \le \alpha(t)$, $\alpha \in L^1(0,T)$, and $b \in L^\infty(0,T;C^{0,1}(\mathbb{R}))$. Then a reversible solution of (7.1) *is defined as follows. For any $z \in \mathbb{R}$ define $p(t, X(t;0,z))$ as a solution of*

$$\begin{aligned} &p(T, X(T;0,z)) = p^T(X(T;0,z)), \\ &\tfrac{d}{dt} p(t, X(t;0,z)) = -b(t, X(t;0,z))\, p(t, X(t;0,z)), \quad t \in [0,T]. \end{aligned} \tag{7.7}$$

If it holds only that $\alpha \in L^1(\sigma,T)$ for all $\sigma > 0$, then we define p first on the domains $(\sigma,T)\times\mathbb{R}$ and then on Ω_T by exhaustion.

REMARK 7.6. *Since by* (7.5) *with $x = X(s;0,z)$, $X(t;s,x) = X(t;0,z)$ holds for $s \le t \le T$, then* (7.7) *implies that for all $0 \le s < T$ and $x \in \mathbb{R}$*

$$\begin{aligned} &p(T, X(T;s,x)) = p^T(X(T;s,x)), \\ &\tfrac{d}{dt} p(t, X(t;s,x)) = -b(t, X(t;s,x)) p(t, X(t;s,x)), \quad t \in [s,T]. \end{aligned} \tag{7.8}$$

With (7.6) *this shows that the value of $p(s,x)$ depends only on the values of a and b in the triangle $\{(t,z) \in \Omega_T \,:\, t \in [s,T], z \in [x - \|a\|_\infty (t-s), x + \|a\|_\infty (t-s)]\}$.*

With this definition of reversible solutions we have the following existence and uniqueness result under the strong one-sided Lipschitz condition (7.2).

THEOREM 7.7. *Let $a \in L^\infty(\Omega_T)$, $\partial_x a(t,\cdot) \le \alpha(t)$ with $\alpha \in L^1(0,T)$, and $b \in L^\infty(0,T;C^{0,1}(\mathbb{R}))$. Then for all $p^T \in C^{0,1}(\mathbb{R})$ there exists a unique reversible solution p of* (7.1)*. Moreover, $p \in C^{0,1}(\Omega_T)$ and p solves* (7.1) *a.e. on Ω_T.*

Furthermore, for all $t \in [0,T]$, $z_1 < z_2$, and $0 \le t_1 < t_2 \le T$, the following hold with $I = [z_1, z_2]$, $J = [z_1 - \|a\|_\infty (T-t), z_2 + \|a\|_\infty (T-t)]$:

$$\|p(t,\cdot)\|_{\infty,I} \le \|p^T\|_{\infty,J} e^{\|b\|_\infty (T-t)}, \tag{7.9}$$

$$\|p(t_2,\cdot) - p(t_1,\cdot)\|_{1,I} \le (t_2 - t_1)\, C, \tag{7.10}$$

where C depends only on I, $\|p^T\|_{H^{1,1}(J)}$, $\|a\|_{\infty,[0,T]\times J}$, $\|b\|_{L^\infty(0,T;C^{0,1}(J))}$. Finally, one has $p \in H^{1,1}(]0,T[\times I)$, where $\|p\|_{H^{1,1}(]0,T[\times I)}$ does not depend on α.

The proof can be obtained by using the properties of the backward flow X. A detailed analysis can be found in the follow-up paper [27].[2]

A key point for our analysis is the following stability result.

THEOREM 7.8. *Let (a_n) be a bounded sequence in $L^\infty(\Omega_T)$ with $a_n \to a$ in $L^\infty(\Omega_T)$-weak*, $\partial_x a(t,\cdot) \le \alpha(t)$, $\alpha \in L^1(0,T)$, and $\partial_x a_n(t,\cdot) \le \alpha_n(t)$, (α_n) bounded in $L^1(0,T)$. Let (b_n) be a bounded sequence in $L^\infty(0,T;C^{0,1}(\mathbb{R}))$ with $b_n \to b$ in $L^\infty(0,T;C_{loc}(\mathbb{R}))$, $b \in L^\infty(0,T;C^{0,1}(\mathbb{R}))$. Finally, let p_n^T be bounded in $C^{0,1}(\mathbb{R})$ with $p_n^T \to p^T$ in $C_{loc}(\mathbb{R})$. Then the reversible solutions p_n of*

$$\partial_t p_n + a_n \partial_x p_n = -b_n p_n, \quad p_n(T,\cdot) = p_n^T$$

[2]For convenience, a proof can be found in the appendix.

converge for any $R > 0$ in $C([0,T] \times [-R,R])$ to the reversible solution p of (7.1).

The proof can be obtained by using the following stability result of [1]: if X and X_n denote the backward flows according to Definition 7.3 for a and a_n, respectively, then by [1] $X_n \to X$ in $C(D_b \times [-R,R])$ for any $R > 0$. Again, the details can be found in the follow-up paper [27] and a proof can be found in the appendix.

REMARK 7.9. *Theorem* 7.8 *shows in particular that reversible solutions are stable w.r.t. smoothing of the coefficients and data by convolution with an averaging kernel.*

By Oleinik's entropy condition (4.1) the case $\alpha(t) = C/t$ is of special interest if the initial data contain an up-jump. Then merely $\alpha \in L^1(\sigma,T)$ for $\sigma > 0$ (cf. the weakened one-sided Lipschitz condition (7.3)) and Theorem 7.7 is only applicable on $]\sigma,T[\times\mathbb{R}$, $\sigma > 0$, instead of Ω_T. For $\sigma \to 0+$ this yields the definition of p on the open set Ω_T. Moreover, we see from (7.9)–(7.10) that $p \in H^{1,1}(]0,T[\times]-R,R[) \cap L^\infty(\Omega_T) \cap C^{0,1}(]0,T];L^1_{loc}(\mathbb{R}))$ for all $R > 0$ and thus admits an L^1-trace $p(0,\cdot) \in L^\infty(\Omega_T)$ with $p(t,\cdot) \to p(0,\cdot)$ as $t \to 0+$ in $L^r_{loc}(\mathbb{R})$ for all $r \in [1,\infty[$. Finally, $p \in L^\infty(0,T;H^{1,1}_{loc}(\mathbb{R}))$, and thus $p(0,\cdot) \in BV_{loc}(\mathbb{R})$. So we have the following extension of Theorem 7.8.

THEOREM 7.10. *Let the assumptions of Theorem* 7.8 *hold with the relaxation that α, α_n are only uniformly bounded in $L^1(\sigma,T)$ for each fixed $\sigma > 0$. Then there exist unique reversible solutions $p, p_n \in C^{0,1}_{loc}(\Omega_T) \cap C^{0,1}(]0,T];L^1_{loc}(\mathbb{R}))$ satisfying* (7.8) *for all $s \in]0,T[$. Moreover,* (7.9)–(7.10) *hold, p, p_n are uniformly bounded in $L^\infty(\Omega_T) \cap H^{1,1}(\Omega_T) \cap C^{0,1}(]0,T];L^1_{loc}(\mathbb{R}))$ and can thus be extended to $C^{0,1}([0,T];L^1_{loc}(\mathbb{R}))$. The traces $p(0,\cdot), p_n(0,\cdot)$ at $t = 0$ are uniformly bounded in $L^\infty(\mathbb{R}) \cap BV_{loc}(\mathbb{R})$. Finally, $p_n \to p$ in $C_{loc}(\Omega_T) \cap C([0,T];L^r_{loc}(\mathbb{R}))$ for all $r \in [1,\infty[$.*

Proof. The statements are immediately clear by Theorems 7.7 and 7.8, Remark 7.6, and the previous considerations. ∎

It will be useful for our adjoint calculus to extend reversible solutions to the case where $p^T \in B_{\mathrm{Lip}}(\mathbb{R})$ with $B_{\mathrm{Lip}}(\mathbb{R})$ as in Definition 7.5. Note that Definition 7.5 covers this case and makes perfect sense. The following theorem shows that the corresponding broad solutions are pointwise limits of Lipschitz backward solutions if p^T is in $B_{\mathrm{Lip}}(\mathbb{R})$.

THEOREM 7.11. *Let a, b be as in Theorem* 7.10. *Then for $p^T \in B_{\mathrm{Lip}}(\mathbb{R})$ the corresponding reversible solution p according to Definition* 7.5 *is unique, and $p \in B(\Omega_T)$ and satisfies* (7.9). *If $p^T_n \in C^{0,1}(\mathbb{R})$ is a sequence, bounded in $C(\mathbb{R}) \cap H^{1,1}_{loc}(\mathbb{R})$, that converges boundedly everywhere to p^T, then the corresponding reversible solutions p^n according to Theorem* 7.7 *converge boundedly everywhere to p.*

The proof follows easily by the definition of reversible solutions as broad solutions; see [27].

8. Differentiability of shock position: Proof of Lemma 6.4. We are now in the position to prove Lemma 6.4.

Proof of Lemma 6.4. We prove (i) and will easily deduce (ii). Throughout the proof we will use the notations of Lemma 6.2.

Let $\hat{u} \in \tilde{V}$ be fixed and $\tilde{u} \in \tilde{V}$ be arbitrary. In the following we write $\hat{y}$ and $\tilde{y}$ instead of $y(\cdot;\hat{u})$ and $y(\cdot;\tilde{u})$ and set $\Delta y \overset{\text{def}}{=} \tilde{y} - \hat{y}$, $\delta u = \tilde{u} - \hat{u}$.

For all sufficiently small $\varepsilon > 0$ one has

$$\hat{x}_\mp \overset{\text{def}}{=} x_s(\hat{u}) \mp \varepsilon \in]x_-, x_+[\tag{8.1}$$

with $x_\mp$ as in Lemma 6.2. Our proof is based on the observation that

$$\int_{\hat{x}_-}^{\hat{x}_+} \Delta y(\bar{t},x)\,dx = (x_s(\tilde{u}) - x_s(\hat{u}))\,[\hat{y}(\bar{t},\hat{x}_s)] + O((\varepsilon + \|\delta u\|_{\mathcal{V}})\|\delta u\|_{\mathcal{V}}). \tag{8.2}$$

In fact, let $\hat{x}_s = x_s(\hat{u})$ and $x_{s,\mp} \stackrel{\text{def}}{=} \min/\max\{x_s(\hat{u}), x_s(\tilde{u})\}$, respectively. Since obviously $[\hat{y}(\bar{t}, \hat{x}_s)] = Y_-(\bar{t}, \hat{x}_s, \hat{u}) - Y_+(\bar{t}, \hat{x}_s, \hat{u})$, we obtain by Lemma 6.2(ii)

$$\begin{aligned}
&\int_{\hat{x}_-}^{\hat{x}_+} \Delta y(\bar{t}, x)\, dx - (x_s(\tilde{u}) - x_s(\hat{u}))\, [\hat{y}(\bar{t}, \hat{x}_s)] \\
&= \int_{\hat{x}_s - \varepsilon}^{x_{s,-}} (Y_-(\bar{t}, x, \tilde{u}) - Y_-(\bar{t}, x, \hat{u}))\, dx + \int_{x_{s,+}}^{\hat{x}_s + \varepsilon} (Y_+(\bar{t}, x, \tilde{u}) - Y_+(\bar{t}, x, \hat{u}))\, dx \\
&\quad + \int_{x_{s,-}}^{\hat{x}_s} (Y_-(\bar{t}, \hat{x}_s, \hat{u}) - Y_-(\bar{t}, x, \hat{u}) + Y_+(\bar{t}, x, \tilde{u}) - Y_+(\bar{t}, \hat{x}_s, \hat{u}))\, dx \\
&\quad + \int_{\hat{x}_s}^{x_{s,+}} (Y_+(\bar{t}, \hat{x}_s, \hat{u}) - Y_+(\bar{t}, x, \hat{u}) + Y_-(\bar{t}, x, \tilde{u}) - Y_-(\bar{t}, \hat{x}_s, \hat{u}))\, dx \\
&\stackrel{\text{def}}{=} R_1 + R_2 + R_3 + R_4.
\end{aligned}$$

Using the Lipschitz continuity of $Y_\mp$ w.r.t. $\tilde{u}$ (cf. Lemma 5.1), we get

$$|R_1 + R_2| \le \varepsilon C \|\delta u\|_{\mathcal{V}}$$

with C independent of ε. Since x_s in (6.1) is Lipschitz by Lemma 6.2 and there is by Lemma 5.1 also a uniform Lipschitz constant of $Y_\mp$ w.r.t. x for all $\tilde{u} \in \tilde{V}$, we have

$$|R_3 + R_4| \le C \|\delta u\|_{\mathcal{V}}^2$$

with C not depending on ε.

Main line of the proof. The main line of the proof for the analysis of the shock sensitivity is therefore to compute the derivative of the special functional $\int_{\hat{x}_-}^{\hat{x}_+} \Delta y(\bar{t}, x)\, dx$ appearing on the left-hand side of (8.2) with $\hat{x}_\mp$ according to (8.1) by adjoint-based techniques and then to take the limit $\varepsilon \to 0$. See also the proof sketch in section 3.1.

To this end we compute the left-hand side of (8.2) by using an adjoint equation with "averaged" coefficients and then using the stability w.r.t. the coefficients to derive the actual adjoint equation for the shock sensitivity.

The difference of (1.1) for $\tilde{y}$ and $\hat{y}$ yields

$$\partial_t \Delta y + \partial_x (f(\tilde{y}) - f(\hat{y})) = g(t, x, \tilde{y}, \tilde{u}_1) - g(t, x, \hat{y}, \hat{u}_1). \tag{8.3}$$

We now define the functions

$$a \stackrel{\text{def}}{=} f'(\hat{y}), \quad \tilde{a}(t, x) \stackrel{\text{def}}{=} \int_0^1 f'(\hat{y}(t, x) + \tau \Delta y(t, x))\, d\tau, \tag{8.4}$$

$$\tilde{b} \stackrel{\text{def}}{=} g_y(\cdot, \tilde{u}_1), \quad b \stackrel{\text{def}}{=} g_y(\cdot, \hat{u}_1). \tag{8.5}$$

Then by the definition of $\tilde{a}, \tilde{b}$ and since g is affine linear in y, we can rewrite (8.3) as

$$\partial_t \Delta y + \partial_x (\tilde{a} \Delta y) = \tilde{b} \Delta y + g(t, x, \hat{y}, \tilde{u}_1) - g(t, x, \hat{y}, \hat{u}_1). \tag{8.6}$$

Let $\tilde{p}$ be a test function satisfying

$$\tilde{p} \in C^{0,1}([\sigma, \bar{t}] \times [-R, R]) \cap C([0, \bar{t}]; L^2_{loc}(\mathbb{R})) \quad \forall\, \sigma > 0, R > 0. \tag{8.7}$$

Denote by $\hat\zeta_\mp$ the genuine backward characteristic of $\hat y = y(\cdot;\hat u)$ through $(\bar t, \hat x_\mp)$. Then by (8.1) and Lemma 6.2 one has with the minimal/maximal backward characteristics $\hat\xi_\mp$ through $\hat x_s$

$$\hat\zeta_\mp(0) \in J_\mp \quad \text{and} \quad \hat\zeta_-(t) \le \hat\xi_-(t) \le \hat\xi_+(t) \le \hat\zeta_+(t) \quad \forall\, t \in [0,\bar t].$$

By Lemma 5.5 there are $\delta > 0$ and $\eta > 0$ such that $\tilde y = Y_\mp(\cdot,\tilde u)$ on $\{(t,x) \ : \ t \in [0,\bar t], |x - \hat\zeta_\mp(t)| < \delta\}$ if $\|\delta u\|_{\tilde{\mathcal{V}}} \le \eta$.

Denote by D_ε the domain between $\hat\zeta_\mp$ for $t \in [0,\bar t]$. Oleinik's entropy condition (4.1) implies $\tilde y, \hat y \in BV([\sigma,\bar t] \times [-R,R])$ for all $\sigma > 0$ and $R > 0$; see Theorem 4.1. Hence, the same is true for Δy, $\tilde a \Delta y$. Therefore, if we multiply (8.6) by $\tilde p$ satisfying (8.7) and integrate over the domain $D_\varepsilon \cap \{t \ge \sigma\}$ between $\hat\zeta_\mp$ for $t \in [\sigma,\bar t]$ we may apply integration by parts (first for $\tilde p$ in C^1 (see [11]) but then for $\tilde p$ satisfying (8.7) by a density argument in $C \cap H^{1,1}$). Now $\Delta y(\sigma,\cdot) \to \delta u_0$ and $\tilde p(\sigma,\cdot) \to \tilde p(0,\cdot)$ in $L^2_{loc}(\mathbb{R})$ as $\sigma \to 0+$ by the initial condition and (8.7). Therefore, the integration by parts can be extended until $\sigma = 0$, i.e., over all of D_ε. Since $\frac{d}{dt}\hat\zeta_\mp(t) = f'(\hat y(t,\hat\zeta_\mp(t)))$ this gives

$$\begin{aligned}
&\int_{\hat x_-}^{\hat x_+} \tilde p(\bar t,x)\Delta y(\bar t,x)\,dx \\
&= \int_{\hat\zeta_-(0)}^{\hat\zeta_+(0)} \tilde p(0,x)\delta u_0(x)\,dx + \int_0^{\bar t}\int_{\hat\zeta_-(t)}^{\hat\zeta_+(t)} \Delta y\,(\partial_t\tilde p + \tilde a\partial_x\tilde p + \tilde b\tilde p)\,dxdt \\
&\quad + \int_0^{\bar t}\int_{\hat\zeta_-(t)}^{\hat\zeta_+(t)} \tilde p\ (g(t,x,\hat y,\tilde u_1) - g(t,x,\hat y,\hat u_1))\,dxdt \\
&\quad + \int_0^{\bar t} \tilde p(t,\hat\zeta_-(t))\ (-f'(\hat y)\Delta y + f(\tilde y) - f(\hat y))(t,\hat\zeta_-(t))\,dt \\
&\quad + \int_0^{\bar t} \tilde p(t,\hat\zeta_+(t))\ (f'(\hat y)\Delta y - f(\tilde y) + f(\hat y))(t,\hat\zeta_+(t))\,dt \\
&\overset{\text{def}}{=} I_1 + I_2 + I_3 + I_4 + I_5.
\end{aligned} \tag{8.8}$$

Our aim is to choose $\tilde p$ as a reversible solution of the "averaged" adjoint equation

$$\partial_t\tilde p + \tilde a\partial_x\tilde p = -\tilde b\tilde p, \quad \tilde p(\bar t,\cdot) = p^{\bar t} \equiv 1/[y(\bar t, x_s(\hat u);\hat u)]. \tag{8.9}$$

In the following, we justify that $\tilde p$ exists and satisfies (8.7), and we study $\tilde p$ for $\tilde u \to \hat u$.

From the stability estimates of Theorem 4.1 we know that $\|\tilde y\|_\infty \le M_y$ for all $\tilde u \in \tilde V$ and

$$\|\Delta y\|_{1,loc} \to 0 \quad \text{as } \|\delta u\|_{\tilde{\mathcal{V}}} \to 0.$$

Hence, we have $\tilde a \in L^\infty(\Omega_{\bar t})$ uniformly bounded and

$$\tilde a \to f'(\hat y) \overset{\text{def}}{=} a \quad \text{in } L^1_{loc}(\Omega_{\bar t}) \text{ and in } L^\infty(\Omega_{\bar t})\text{-weak}^* \quad \text{as } \|\delta u\|_{\tilde{\mathcal{V}}} \to 0. \tag{8.10}$$

Moreover, since $f'' > 0$ we get by Oleinik's entropy condition (4.1) that there exists a constant $C > 0$ with

$$\partial_x\tilde a(t,\cdot), \partial_x a(t,\cdot) \le \frac{C}{t}. \tag{8.11}$$

By assumption (A4), the coefficients $\tilde{b}, b$ in (8.5) do not depend on y, and we have $\tilde{b}, b \in L^\infty(0,T;C^{0,1}(\mathbb{R}))$ by (A2).

Thus, (8.9) has by Theorem 7.10 a unique reversible solution $\tilde{p}$ satisfying (8.7) ($\tilde{p} \in C([0,T];L^2_{loc}(\mathbb{R}))$ follows from $\tilde{p} \in C([0,T];L^1_{loc}(\mathbb{R})) \cap L^\infty(\Omega_T)$). Denote moreover by p the reversible solution of (6.10) (i.e., of (8.9) with a, b instead of $\tilde{a}, \tilde{b}$), which exists for the same reasons by Theorem 7.10.

Let $M_u > 0$ such that $\|\tilde{u}_1\|_\infty \le M_u$ for all $\tilde{u} \in \tilde{V}$. By (A2) and (A4) g_y admits a Lipschitz constant $L_{g'}$ for u_1 on $[-M_u, M_u]^m$ and thus

$$\|\tilde{b} - b\|_{L^\infty(0,T;C(\mathbb{R}))} \le L_{g'} \|\delta u_1\|_\infty \le L_{g'} \|\delta u\|_{\tilde{\mathcal{V}}}.$$

Thus, we obtain with (8.10) by Theorem 7.10 that the reversible solutions $\tilde{p}$ of (8.9) and p of (6.10) satisfy for all $\sigma > 0$, $R > 0$

$$\tilde{p}, p \in C^{0,1}([\sigma, \bar{t}] \times [-R, R]) \tag{8.12}$$

and for all $r \in [1, \infty)$

$$\tilde{p} \to p \text{ in } C([\sigma,\bar{t}] \times [-R,R]) \cap C([0,\bar{t}];L^r(-R,R)) \quad \text{as } \|\delta u\|_{\tilde{\mathcal{V}}} \to 0. \tag{8.13}$$

We show that even

$$\|\tilde{p}(0,\cdot) - p(0,\cdot)\|_{C(J)} \to 0 \quad \text{as } \|\delta u\|_{\tilde{\mathcal{V}}} \to 0. \tag{8.14}$$

In fact, since u_0 has only admissible discontinuities (down-jumps) on J_s, there is by the definition of $\tilde{V}$ some $M_0 > 0$ with $\partial_x \tilde{u}_0|_{J_s} \le M_0$ in the sense of distributions for all $\tilde{u} \in \tilde{V}$. Since J has its closure in J_s, there exists $\rho > 0$ such that $J_{2\rho} \subset J_s$ for the 2ρ-neighborhood $J_{2\rho}$ of J. Moreover, $\|\tilde{a}\|_\infty \le M_a$ for $M_a = \sup_{|y| \le M_y} |f'(y)|$ and all $\tilde{u} \in \tilde{V}$. Thus, there is $\sigma_0 > 0$ such that

$$K_p \overset{\text{def}}{=} \{(t,x) \in [0,\bar{t}] \times \mathbb{R} : \operatorname{dist}(x,J) \le M_a t\} \subset ([0,\sigma] \times J_\rho) \cup ([\sigma,\bar{t}] \times \mathbb{R})$$

for all $\sigma \in]0, \sigma_0]$. The propagation speed of $\tilde{y}$ is uniformly bounded by M_a for all $\tilde{u} \in \tilde{V}$. Hence, there is $\sigma \in]0, \sigma_0]$ such that $\tilde{y}|_{[0,\sigma]\times J_\rho}$ does only depend on $(\tilde{u}_0|_{J_s}, \tilde{u}_1)$. Smoothing $\tilde{u}_0$ outside J_s thus yields by (4.1) that $\partial_x \tilde{y}|_{[0,\sigma]\times J_\rho} \le M_1$ for some $M_1 > 0$ and thus $\partial_x \tilde{a}|_{[0,\sigma]\times J_\rho} \le M_2$. Together with (8.11) we have, possibly after increasing M_2,

$$\partial_x \tilde{a}|_{K_p} \le M_2.$$

By Remark 7.6, $\tilde{p}|_{K_p}$ depends only on the values of $\tilde{a}$, $\tilde{b}$ on K_p. Thus, by Theorem 7.8

$$\tilde{p} \to p \text{ in } C(K_p) \quad \text{as } \|\delta u\|_{\tilde{\mathcal{V}}} \to 0,$$

which implies (8.14).

We now consider the terms $I_1, \dots, I_5$ in (8.8). Since $\tilde{p}$ solves (8.9) a.e. by Theorem 7.10, we have $I_2 = 0$. Moreover, (7.9) yields $M_p > 0$ with $\|\tilde{p}\|_\infty \le M_p$ for all $\tilde{u} \in \tilde{V}$, and we obtain by Lemma 5.1

$$|I_{4/5}| \le \int_0^{\bar{t}} M_p\, M_{f''}\, |\Delta y(t, \hat{\zeta}_\mp(t))|^2\, dt \le \bar{t}\, M_p\, M_{f''} \|Y_\mp(\cdot;\tilde{u}) - Y_\mp(\cdot;\hat{u})\|^2_{\infty, S_\mp} \le C\|\delta u\|^2_{\mathcal{V}}$$

with C independent of ε. Moreover, we have with (A2) and the dominated convergence theorem

$$\|g(\cdot,\hat{y},\tilde{u}_1) - g(\cdot,\hat{y},\hat{u}_1) - g_{u_1}(\cdot,\hat{y},\hat{u}_1)\delta u_1\|_{1,D_\varepsilon} = o(\|\delta u_1\|_\infty)$$

independently of ε. Therefore,

$$\left| I_3 - \int_{D_\varepsilon} p\, g_{u_1}(t,x,\hat{y},\hat{u}_1)\delta u_1 \, d(t,x) \right| \le \|p\|_\infty o(\|\delta u_1\|_\infty) + C\|\tilde{p}-p\|_{1,D_\varepsilon}\|\delta u_1\|_\infty$$

with C independent of ε. From the backward stability of genuine backward characteristics we find $L_\zeta > 0$ independent of ε with $|\hat{\zeta}_\mp(t) - \hat{\xi}_\mp(t)| \le L_\zeta \varepsilon$ for all $t \in [0,\bar{t}]$. Thus, we obtain with the last inequality

$$\left| I_3 - \int_0^{\bar{t}} \int_{\hat{\xi}_-(t)}^{\hat{\xi}_+(t)} p\, g_{u_1}(t,x,\hat{y},\hat{u}_1)\delta u_1 \, dxdt \right| \le C(\varepsilon + \|\tilde{p}-p\|_{1,D_\varepsilon})\|\delta u\|_{\mathcal{V}} + o(\|\delta u\|_{\mathcal{V}})$$

for some $C > 0$ independent of ε. Moreover, since $\hat{\xi}_\mp(0), \hat{\zeta}_\mp(0) \in J_\mp$

$$\begin{aligned}\left| I_1 - \int_{\hat{\xi}_-(0)}^{\hat{\xi}_+(0)} p\,\delta u_0 \, dx \right| &\le (L_\zeta\varepsilon\|\tilde{p}(0,\cdot)\|_\infty + \|(\tilde{p}-p)(0,\cdot)\|_{1,J_r\cup J_-\cup J_+})\|\delta u_0\|_{\infty,J_r\cup J_-\cup J_+} \\ &\quad + \|(\tilde{p}-p)(0,\cdot)\|_{\infty,J}\|\delta u_0\|_{1,J} \\ &\le C(\varepsilon + \|(\tilde{p}-p)(0,\cdot)\|_{1,J_r\cup J_-\cup J_+} + \|(\tilde{p}-p)(0,\cdot)\|_{\infty,J})\|\delta u\|_{\tilde{\mathcal{V}}}.\end{aligned}$$

Using the estimates for $I_1,\dots,I_5$, $R_1,\dots,R_4$ together with (8.13), (8.14), the Fréchet differentiability of $x_s : (\tilde{V}, \|\cdot\|_{\tilde{\mathcal{V}}}) \longrightarrow \mathbb{R}$ is obvious and the adjoint formula (6.9) is shown with $p\mathbf{1}_{D^{cl}}(\cdot)$ instead of p, where $D = D_0$ is the domain between $\hat{\xi}_\mp$, $0 \le t \le \bar{t}$. But it is obvious from Definition 7.5 that $p\mathbf{1}_{D^{cl}}(\cdot)$ is nothing else but the reversible solution of (6.10) for the initial data (6.10), i.e., $p^{\bar{t}} = 1/[y(\bar{t},x_s(\hat{u});\hat{u})]\,\mathbf{1}_{\{x=x_s(\hat{u})\}}(\cdot)$, since then p has automatically the support D^{cl}, and coincides on D^{cl} with the reversible solution for the end data $p^{\bar{t}} \equiv 1/[y(\bar{t},x_s(\hat{u});\hat{u})]$. In fact, p is by Definition 7.5 and Remark 7.4 transported along the backward characteristics.

The derivative is also continuous. In fact, denote by $\tilde{p}$ now the reversible solution of (8.9) with $\tilde{p}^{\bar{t}} = 1/[\tilde{y}(\bar{t},x_s(\tilde{u}))]$, $\tilde{a} = f'(\tilde{y})$, $\tilde{b} = g_y(\cdot,\tilde{u}_1)$. Then we already know that $\tilde{p}^{\bar{t}} \to p^{\bar{t}}$ in $C(\mathbb{R})$ as $\|\delta u\|_{\tilde{\mathcal{V}}} \to 0$ and $\tilde{a}$, $\tilde{b}$ have the same properties as above. Thus, we obtain again that (8.13), (8.14) hold. Together with the Lipschitz continuity of $\hat{u} \in (\tilde{V}, \|\cdot\|_{\tilde{\mathcal{V}}}) \longmapsto \hat{\xi}_\mp(\cdot) \in C([0,T])$ and the definition of $\|.\|_{\tilde{\mathcal{V}}}$ this shows immediately that the derivative of x_s is continuous.

(ii) In this case the adjoint equation has no source term and (8.9) as well as (6.10) have both the reversible solutions $\tilde{p} = p \equiv 1/[y(\bar{t},x_s(\hat{u});\hat{u})]$. Taking $(\tilde{V}, \|\cdot\|_{\tilde{\mathcal{V}}})$ instead of $(\hat{V}, \|.\|_{\mathcal{V}})$ was only necessary to ensure (8.12)–(8.14), which are now trivial. The proof is now exactly as before but less technical since $\tilde{p} = p = \text{const}$. □

REMARK 8.1. *If g is not affine linear w.r.t. y, then the mean value coefficient $\tilde{b}$ in* (8.5) *has the form*

$$\tilde{b} = \int_0^1 g_y(\cdot,\hat{y}(t,x+) + \tau\Delta y(t,x),\tilde{u}_1)\,d\tau, \tag{8.15}$$

where we define $\Delta y(t,x)$ everywhere by setting

$$\Delta y(t,x) = \max\left(\tilde{y}(t,x-) - \hat{y}(t,x+), \tilde{y}(t,x+) - \hat{y}(t,x-)\right). \tag{8.16}$$

Then $\tilde{b}$ is discontinuous at shocks of $\hat{y}$ and $\tilde{y}$. Nevertheless, we can again define solutions of the "averaged" adjoint equation (8.9) *along the characteristics as in Definition* 7.5. *But now the definition of the limit coefficient b at the shocks is important, since it influences the value of the adjoint state on the domain covered by backward characteristics emanating from both sides of the shock. Since the backward characteristics corresponding to the mean value coefficient $\tilde{a}$ propagate close to a shock of $\hat{y}$ in the area between the shock fronts of $\hat{y}$, the limit coefficient b in the adjoint equation has to be defined by*

$$b = \int_0^1 g_y(\cdot, \hat{y}(t, x+) + \tau[\hat{y}(t, x)], \tilde{u}_1)\, d\tau.$$

One can now show by the results of sections 4–8 *that $\tilde{b} \to b$ in all continuity points of $\hat{y}$ and in all nondegenerate shock points. The limit coefficient b is discontinuous, but the singular parts of $\partial_x \tilde{b}, \partial_x b$ have densities w.r.t. $\partial_x \tilde{a}, \partial_x a$, respectively. We believe that this allows us to derive at least a bound for $\tilde{p}, p$ in $H^{1,2}$.*

We briefly sketch how a stability result can be obtained in the case of a piecewise Lipschitz continuous solution with a nondegenerate shock $\hat{\eta}$: then we know that the shock is stable and also $\tilde{y}$ has a shock $\tilde{\eta}$ that connects uniformly varying Lipschitz continuous states. Denote the area between the shock fronts $\hat{\eta}$ and $\tilde{\eta}$ by S. Generalized backward characteristics of the averaged equation cannot enter S from outside. Thus they start in $\{\bar{t}\} \times I(\hat{\eta}, \tilde{\eta})$ and leave at some point $(s, \hat{\eta}(s))$ or $(s, \tilde{\eta}(s))$. They define a Lipschitz continuous function $\tilde{p}|_S$ that converges uniformly to $p(t, \hat{\eta}(t))$, the reversible solution of the limit equation along the shock. In particular, $\tilde{p}(t, \hat{\eta}(t)), \tilde{p}(t, \tilde{\eta}(t)) \to p(t, \hat{\eta}(t))$ uniformly. Outside of S, $\tilde{p}$ is obtained by integrating along the characteristics starting from the Lipschitz continuous boundary data $\tilde{p}(t, \hat{\eta}(t))$, $\tilde{p}(t, \tilde{\eta}(t))$, and $p^{\bar{t}}(x)$, $x \notin I(\hat{\eta}, \tilde{\eta})$, respectively. Thus, $\tilde{p}$ is Lipschitz continuous and converges uniformly to p, since p is obtained by integrating along the characteristics starting from the boundary data $p(t, \hat{\eta}(t))$, $p^{\bar{t}}(x)$ that are the uniform limit of the boundary data of $\tilde{p}$.

A complete study of the adjoint equation in the general case is in progress. We emphasize that our analysis extends to general source terms as soon as the stability of the adjoint state is established.

REMARK 8.2. *The applied techniques for the shock sensitivity analysis are quite universal and can formally be extended to systems. The domain D_ε must then be chosen between the fastest characteristic field through $\hat{x}_-$ and the slowest through $\hat{x}_+$ corresponding to the mean value Jacobian $\tilde{a}$. Then similar arguments can be used if appropriate stability properties of shock, state, and adjoint equation hold. The appropriate adjoint state must be stable w.r.t. the coefficients. This is an interesting and challenging topic for future research.*

9. Proof of the main results. In this section we will use the results of the previous sections to prove the main results of this section that we have already stated in section 3.

9.1. Shift-differentiability of entropy solutions. In this subsection we will prove the shift-differentiability result for entropy solutions stated in Theorem 3.2 as well as the formula for the shift derivative given in Theorem 3.4.

Proof of Theorem 3.2 *and Theorem* 3.4. The proof has three parts. We consider first shock points then continuity points.

Step 1. We show first the shift-differentiability in a neighborhood of a shock point. Set $y = y(\cdot; u)$. By assumption, $y(\bar{t}, \cdot)$ contains on I finitely many shocks

at $\bar{x}_1, \dots, \bar{x}_K$ satisfying the assumptions of Lemma 6.2 or Corollary 6.5. Let us for concreteness assume that $\bar{x}_k$ is of type $\mathrm{C}^c\mathrm{C}^c$, i.e., satisfies the scenario of Lemma 6.2. The case where left and/or right states lie on a rarefaction wave can be treated very similarly using Corollary 6.5.

Thus, for $\bar{x} = \bar{x}_k$ Lemmas 6.2 and 6.4 are applicable. We denote by $\xi_\mp$ the extreme backward characteristics through $(\bar{t}, \bar{x})$ and by D_k the confined domain. We verify that with the notation of Lemma 6.2 for $\|(w,\sigma)\|_W < \rho$, $\rho > 0$ sufficiently small, one has

(9.1)
$$u^s \stackrel{\text{def}}{=} (u_0 + w_0^s, u_1 + w_1) \in \hat{V}, \quad \|(w_0^s, w_1)\|_{\mathcal{V}} \le C\,\|(w,\sigma)\|_W \text{ for } w_0^s \stackrel{\text{def}}{=} S_{u_0}^{(x_i)}(w_0,\sigma).$$

In fact, there is $\rho > 0$ with $w_0^s|_{J_\mp} = w_0|_{J_\mp}$ whenever $\|\sigma\|_2 < \rho$, and for all $R > 0$

$$\|S_{u_0}^{(x_i)}(w_0,\sigma)\|_{1,[-R,R]} \le 2R\|w_0\|_\infty + 2\|y(\bar{t},\cdot;u)\|_\infty \|\sigma\|_1 \le C\,\|(w,\sigma)\|_W.$$

Hence, for sufficiently small $\rho > 0$, (9.1) is obvious and Lemma 6.2 holds for $\hat{u} = u^s$ according to (9.1) if $\|(w,\sigma)\|_W < \rho$. As a consequence, $y(\bar{t},\cdot;u^s)$ has in a neighborhood $\hat{I} =]x_-, x_+[$ of $\bar{x}$ the form (6.2), and the shock position $x_s(u^s)$ in (6.1) depends by (9.1) and Lemma 6.2(ii) Lipschitz continuously on $(w,\sigma) \in W$.

We now show that $(w,\sigma) \in W \mapsto x_s(u^s)$ is Fréchet differentiable at 0 by applying the differentiability result of Lemma 6.4(i). Since $S_{u_0}^{(x_i)}(w_0,\sigma)$ can create small additional up-jumps, it will be useful to introduce the slight modification

$$\tilde{S}_{u_0}^{(x_i)}(w_0,\sigma)(x) \stackrel{\text{def}}{=} w_0(x) + \sum_{i=1}^{N} [(u_0+w_0)(x_i), u_0(x_i)]_+ \operatorname{sgn}(\sigma_i)\mathbf{1}_{I(x_i, x_i+\sigma_i)}(x),$$

where

$$[(u_0+w_0)(x_i), u_0(x_i)]_+ \stackrel{\text{def}}{=} \begin{cases} [(u_0+w_0)(x_i)]_+ & \text{if } [u_0(x_i)]_+ > 0, \\ 0 & \text{else.} \end{cases}$$

Now choose the open sets J_s and $J \subset J_s$ in Lemma 6.4(i) such that J and J_s contain all down-jumps of u_0 between J_- and J_+. Then we have for $\|(w,\sigma)\|_W < \rho$, $\rho > 0$ sufficiently small, in addition to (9.1), that with $\tilde{V}, \tilde{\mathcal{V}}$ from Lemma 6.4

$$\left.\begin{array}{c} (u_0 + \tilde{w}_0^s, u_1 + w_1) \in \tilde{V} \\ \|(\tilde{w}_0^s, w_1)\|_{\tilde{\mathcal{V}}} \le C\,\|(w,\sigma)\|_W \end{array}\right\} \quad \text{for } \tilde{w}_0^s \stackrel{\text{def}}{=} \tilde{S}_{u_0}^{(x_i)}(w_0,\sigma). \tag{9.2}$$

In fact, since $\tilde{S}_{u_0}^{(x_i)}(w_0,\sigma)$ shifts only down-jumps without creating additional jumps, there are $M_L > 0$, $\rho > 0$ such that $\|(w,\sigma)\|_W < \rho$ implies $\tilde{w}_0^s|_{J_\mp \cup J_r} = w_0|_{J_\mp \cup J_r}$ and $\partial_x(u_0 + \tilde{w}_0^s)|_{J_s} \le M_L$.

We are now in the position to show that $(w,\sigma) \in W \longmapsto x_s(u^s)$ is Fréchet differentiable at 0. Consider $(\delta w, s) \in W$ with $\|(\delta w, s)\|_W < \rho$ and set

$$\hat{u} = (u_0 + \delta w_0^s, u_1 + \delta w_1), \quad \tilde{u} = (u_0 + \delta\tilde{w}_0^s, u_1 + \delta w_1)$$

with

$$\delta w_0^s = S_{u_0}^{(x_i)}(\delta w_0, s), \quad \delta\tilde{w}_0^s = S_{u_0+\delta w_0}^{(x_i)}(\delta w_0, s). \tag{9.3}$$

Then for $\|(\delta w, s)\|_W < \rho$, and by using (9.2), Lemma 6.4(i) yields

$$|x_s(\tilde{u}) - x_s(u) - d_u x_s(u) \cdot (\delta \tilde{w}_0^s, \delta w_1)| = o(\|\delta w, s\|_W).$$

Moreover, by (9.3) it holds that

$$\|\delta \tilde{w}_0^s - \delta w_0^s\|_1 \leq \|(\delta \tilde{w}_0^s, \delta w_1) - (\delta w_0^s, \delta w_1)\|_{\mathcal{V}} \leq \sum_{i=1}^{N} |[\delta w_0(x_i)]|\, |s_i| \leq C\, \|(\delta w, s)\|_W^2$$

and thus by the Lipschitz continuity of (6.1) also that

$$|x_s(\hat{u}) - x_s(u) - d_u x_s(u)(\delta w_0^s, \delta w_1)| = o(\|(\delta w, s)\|_W),$$

where we utilize that by Lemma 6.4, and by (6.9) with the reversible solution p of (6.10)–(6.11), it holds that

$$|d_u x_s(u) \cdot (\delta \tilde{w}_0^s - \delta w_0^s, 0)| \leq \|p(0,\cdot)\|_\infty \|\delta \tilde{w}_0^s - \delta w_0^s\|_1 = o(\|(\delta w, s)\|_W).$$

Finally, we have again by Lemma 6.4 with $\bar{s}_k$ according to (3.8) that

$$|d_u x_s(u) \cdot (\delta w_0^s, \delta w_1) - \bar{s}_k| \leq \sum_{i=1}^{N} [u_0(x_i)]_+ \left| \int_{x_i}^{x_i + s_i} |p(0,x) - p(0,x_i)|\, dx \right| = o(\|s\|_2),$$

where we have used that $p(0,x)$ is continuous in all down-jumps x_i; cf. (8.14) and recall that p vanishes in a neighborhood of the remaining down-jumps, since $\operatorname{supp} p = D_k^{cl}$. This concludes the proof that

$$(w_0, w_1, \sigma) \in W \longmapsto x_s(u^s) \tag{9.4}$$

is Fréchet differentiable in 0 with derivative (3.8). We can now show that

$$(w_0, w_1, \sigma) \in W \longmapsto y(\bar{t}, \cdot; u^s)|_{\hat{I}} \tag{9.5}$$

is shift-differentiable at 0. We know that (6.2) holds on $\|(w,\sigma)\|_W < \rho$. Hereby, the mappings $\bar{u} \in (\hat{V}, \|\cdot\|_{\mathcal{V}}) \longmapsto Y_\mp(\bar{t}, \cdot, \bar{u}_0|_{J_\mp}, \bar{u}_1) \in C(\hat{I})$ are continuously differentiable. We have already observed that $w_0^s|_{J_\mp} = w_0|_{J_\mp}$ on $\|(w,\sigma)\|_W < \rho$ and deduce directly from (9.1) that

$$(w_0, w_1, \sigma) \in W \longmapsto Y_\mp(\bar{t}, \cdot, u_0^s|_{J_\mp}, u_1^s) \in C(\hat{I}) \tag{9.6}$$

are continuously differentiable on $\|(w,\sigma)\|_W < \rho$. According to Remark 5.4

$$\delta Y_\mp \overset{\text{def}}{=} (d_u Y_\mp(\cdot, u) \cdot (\delta w_0|_{J_\mp}, \delta w_1))|_{S_\mp} \tag{9.7}$$

are on $S_\mp$ broad solutions of the variational equation (3.3) and satisfy on $J_\mp = S_\mp^{cl} \cap \{t = 0\}$ the initial condition (3.4). Now set

$$\delta y^{\bar{t}} = \delta Y_-(\bar{t}, \cdot)|_{]x_-, x_s(u)[} + \delta Y_+(\bar{t}, \cdot)|_{]x_s(u), x_+[} \tag{9.8}$$

as claimed in (3.7). With

$$\tilde{y}(\bar{t}, \cdot) \overset{\text{def}}{=} Y_-(\bar{t}, \cdot, \hat{u})|_{]x_-, \bar{x}[} + Y_+(\bar{t}, \cdot, \hat{u})|_{]\bar{x}, x_+[}, \quad \hat{y}(\bar{t}, \cdot) = y(\bar{t}, \cdot; \hat{u})$$

we get by (9.7), (9.8) that $\|\tilde{y}(\bar{t},\cdot) - y(\bar{t},\cdot) - \delta y\|_{\infty,\hat{I}} = o(\|\delta w, s\|_W)$ and thus

$$\begin{aligned}(9.9)\qquad &\|\hat{y}(\bar{t},\cdot) - y(\bar{t},\cdot) - S^{(\bar{x}_k)}_{y(\bar{t},\cdot)}(\delta y^{\bar{t}}, \bar{s})\|_{1,\hat{I}} \\ &\quad = \|\hat{y}(\bar{t},\cdot) - \tilde{y}(\bar{t},\cdot) - \operatorname{sgn}(\bar{s}_k)[y(\bar{t},\bar{x})]_+ \mathbf{1}_{I(\bar{x},\bar{x}+\bar{s}_k)}\|_{1,\hat{I}} + o(\|\delta w, s\|_W).\end{aligned}$$

Using (6.2) and the definition of $\tilde{y}$, we obtain

$$(9.10)\qquad \hat{y}(\bar{t},\cdot)|_{\hat{I}} - \tilde{y}(\bar{t},\cdot)|_{\hat{I}} = \operatorname{sgn}(x_s(\hat{u}) - \bar{x})(Y_-(\bar{t},\cdot,\hat{u}) - Y_+(\bar{t},\cdot,\hat{u}))\, \mathbf{1}_{I(\bar{x},x_s(\hat{u}))},$$

and since $Y_\mp(\bar{t},\cdot,\cdot)$ are Lipschitz continuous at $(\bar{x},u)$ by Lemma 5.1, we have

$$(9.11)\qquad \|Y_-(\bar{t},\cdot,\hat{u}) - Y_+(\bar{t},\cdot,\hat{u}) - [y(\bar{t},\bar{x})]\|_{\infty,I(\bar{x},x_s(\hat{u}))} = O(\|(\delta w, s)\|_W).$$

Oleinik's entropy condition yields $[y(\bar{t},\bar{x})] = [y(\bar{t},\bar{x})]_+$, and thus we have by (9.10) and (9.11)

$$\begin{aligned}&\|\hat{y}(\bar{t},\cdot) - \tilde{y}(\bar{t},\cdot) - \operatorname{sgn}(\bar{s}_k)[y(\bar{t},\bar{x})]_+ \mathbf{1}_{I(\bar{x},\bar{x}+\bar{s}_k)}\|_{1,\hat{I}} \\ &\quad = [y(\bar{t},\bar{x})]_+ |\bar{x} + \bar{s}_k - x_s(\hat{u})| + O(\|(\delta w, s)\|_W^2) = o(\|(\delta w, s)\|_W).\end{aligned}$$

Inserting this in (9.9) yields the shift-differentiability of (9.5) on $\hat{I}$ with $(\delta y^{\bar{t}}|_{\hat{I}}, \bar{s}_k) = T_{s,\hat{I}}(0)\cdot(\delta w, s)$ according to (3.7), (3.8). As mentioned above shock points with a rarefaction wave as left and/or right states can be treated similarly as in Corollary 6.5.

Step 2. We show that (9.5) is even continuously shift-differentiable on $\{\|(w,\sigma)\|_W < \rho\}$ for sufficiently small $\rho > 0$. In fact, on $\{\|(w,\sigma)\|_W < \rho\}$ the function on the right-hand side of (9.5) is of the form (6.2). Fix an arbitrary $(w,\sigma) \in W$, $\|(w,\sigma)\|_W < \rho$ and consider the corresponding control

$$(9.12)\qquad (u_0^s, u_1^s) = u^s \overset{\text{def}}{=} (u_0 + S^{(x_i)}_{u_0}(w_0,\sigma), u_1 + w_1).$$

Then u_0^s is again piecewise C^1, and u_1^s has the same regularity as u_1. Moreover,

$$u_0 + S^{(x_i)}_{u_0}(w_0 + \delta w_0, \sigma + s) = u_0^s + S^{(x_i+\sigma_i)}_{u_0^s}(\delta w_0, s);$$

i.e., varying $(\delta w, s)$ produces a shift-variation of u_0^s. Thus, the arguments of Step 1 show that (9.5) is also shift-differentiable at (w,σ). As above, the shift-derivative $(\delta y^{\bar{t},s}|_{\hat{I}}, \bar{s}^s) = T_{s,\hat{I}}(w,\sigma)\cdot(\delta w, s)$ on $\hat{I}$ is given by (3.7), (3.8) with u, y, x_i replaced by $u^s, y(.;u^s), x_i^s = x_i + \sigma_i$. Consequently, it suffices to show the continuity of the shift-derivative in 0. The continuity of (9.4) and (9.6) implies the continuity of $(w,\sigma) \longmapsto y(\bar{t}, x_s(u^s)\pm; u^s)$ for u^s according to (9.12). It remains to show that $(w,\sigma) \in W \longmapsto T_{s,\hat{I}}(w,\sigma) \in \mathcal{L}(W, L^r(\hat{I}) \times \mathbb{R})$ is continuous for some $r > 1$. Since (9.4) is continuous and (9.6) are continuously Fréchet differentiable, it is obvious from (9.7), (9.8) with u^s instead of u that the mapping $(w,\sigma) \in W \longmapsto \delta y^{\bar{t},s}|_{\hat{I}} \in L^r(\hat{I})$ depends continuously on $(w,\sigma) \in W$ in a neighborhood of 0 for all $r \in]1,\infty[$.

We denote again the extreme backward characteristics of y through $(\bar{t}, x_s(u))$ by $\xi_\mp$, the confined domain by D_k, and the extreme backward characteristics of $y^s \overset{\text{def}}{=} y(\cdot;u^s)$ through $(\bar{t}, x_s(u^s))$ by $\xi^s_\mp$. With $I_k \overset{\text{def}}{=} D_k^{cl} \cap \{t = 0\}$ we obtain by (3.8)

$$\begin{aligned}|\bar{s}_k^s - \bar{s}_k| &\le \|p^s(0,\cdot) - p(0,\cdot)\|_{1,I_k} \|\delta w_0\|_{\infty,I_k} + (\|\xi_- - \xi_-^s\|_{\infty,[0,\bar{t}]} + \|\xi_+ - \xi_+^s\|_{\infty,[0,\bar{t}]}) \\ &\quad \cdot (\|p^s(0,\cdot)\|_{\infty,J_-\cup J_+} \|\delta w_0\|_{\infty,J_-\cup J_+} + \|p^s\, g_{u_1}(\cdot, y^s, u_1^s)\|_{\infty,\Omega_{\bar{t}}} \|\delta w_1\|_{\infty,\Omega_{\bar{t}}}) \\ &\quad + \|p^s\, g_{u_1}(\cdot, y^s, u_1^s) - p\, g_{u_1}(\cdot, y, u_1)\|_{1,D_k} \|w_1\|_{\infty,D_k} \\ &\quad + \sum_{x_i \in I_k} (|[u_0^s(x_i^s)]_+ - [u_0(x_i)]_+| |p^s(0,x_i^s)| + |[u_0(x_i)]_+| |p^s(0,x_i^s) - p(0,x_i)|)\, |s_i|,\end{aligned}$$

where p is the reversible solution of (3.5) and p^s is the reversible solution of (3.5) for y^s instead of y and $x_s(u^s)$ instead of $\bar{x}_k = x_s(u)$. We deduce that

$$\sup_{\|(\delta w,s)\|_W=1} |\bar{s}_k^s - \bar{s}_k| \to 0 \quad \text{as} \quad \|(w,\sigma)\|_W \to 0$$

from the following observations. By (6.2) and the continuity of (9.4), (9.6) it follows that $\|\xi_\mp^s - \xi_\mp\|_{\infty,[0,\bar{t}]} \to 0$ as $\|(w,\sigma)\|_W \to 0$. Moreover, y^s is bounded in L^∞ with $y^s \to y$ in $L^1_{loc}(\Omega_T)$, and exactly the same arguments as in the proof of Lemma 6.4(i) show with Theorem 7.10 that for all $\tau > 0$, $R > 0$ we have $p^s, p \in C^{0,1}([\tau,\bar{t}]\times[-R,R])$ and for all $r \in [1,\infty[$

$$p^s \to p \ \text{ in } \ C([\tau,\bar{t}]\times[-R,R]) \cap C([0,\bar{t}]; L^r_{loc}(\mathbb{R})) \ \text{ as } \ \|(w,\sigma)\|_W \to 0.$$

In particular, we obtain $\|p^s\, g_{u_1}(\cdot, y^s, u_1^s) - p\, g_{u_1}(\cdot, y, u_1)\|_{1,D_k} \to 0$ by the Lebesgue dominated convergence theorem and a subsequence-subsequence argument, since g_{u_1} is by (A2) a Carathéodory function. To estimate the sum in the last term we observe that $[u_0^s(x_i+\sigma_i)] = [u_0(x_i)]$ for $\|\sigma\|_2$ small enough. We still have to show that $|p^s(0,x_i+\sigma_i) - p(0,x_i)| \to 0$ if $[u_0(x_i)]_+ \neq 0$. Recall that $p^s(0,\cdot)$, $p(0,\cdot)$ are continuous in the admissible discontinuities $x_i+\sigma_i$, x_i of u_0^s and u_0; cf. (8.14). Fix an arbitrarily small $\tau > 0$. We know that $p(\tau,\cdot) \in C^{0,1}_{loc}(\mathbb{R})$ and that $\|p^s(\tau,\cdot) - p(\tau,\cdot)\|_{C([-R,R])} \to 0$. Let η, η^s be the shocks of y, y^s emanating from x_i and $x_i+\sigma_i$, respectively. Clearly, for all $\tau > 0$ sufficiently small $y(\tau,\cdot)$ satisfies the framework of Lemma 6.2 in a neighborhood of $\eta(\tau)$. Hence, we know from the first part of the proof that $\eta^s(\tau) \to \eta(\tau)$ as $\|(w,\sigma)\|_W \to 0$. $p(t,\eta(t))$ and $p^s(t,\eta^s(t))$ satisfy by Remarks 7.4 and 7.6 an ordinary differential equation of the form (7.8). We thus obtain

$$\begin{aligned} |p^s(0,x_i^s) - p(0,x_i)| \le\ & |p^s(\tau,\eta_s(\tau)) - p(\tau,\eta(\tau))| \\ & + \tau\, \|p\, g_{u_1}(\cdot,y,u_1) - p^s\, g_{u_1}(\cdot,y^s,u_1^s)\|_{\infty,D_k}, \end{aligned}$$

and the right-hand side becomes arbitrarily small for τ and $\|(w,\sigma)\|_W$ sufficiently small.

Step 3. It remains to consider the continuity points. Since $y(\bar{t},\cdot;u)$ contains no shock generation points, all continuity points satisfy the scenario of Lemma 5.5, 5.10, 5.11, or 5.12, respectively, corresponding to the Cases C^c, R^c, CB^c, and RB^c. In Case C^c there are by Lemma 5.5(ii) $\hat{I} =]x_-, x_+[\ni \bar{x}$ and $z_- < \bar{z} < z_+$ such that for $(Y,Z,V,S,J) = \mathrm{YC}(u,\bar{t},[z_-,z_+])$ obtained by Lemma 5.1 it holds that

$$y(\bar{t},\cdot;\hat{u})|_{\hat{I}} = Y(\bar{t},\cdot;\hat{u}_0|_J, \hat{u}_1)|_{\hat{I}} \tag{9.13}$$

for all $\hat{u} \in \hat{V}$ with $\hat{V}$ defined in Lemma 5.5(ii). Hereby, $x_i \notin J$, $i = 1,\dots,N$. Let $M_\infty > 0$ be chosen large enough in the definition of $\hat{V}$. Then we find $\rho > 0$ such that on $\{\|(w,\sigma)\|_W < \rho\}$ with u^s in (9.1), $u_0^s|_J = (u_0+w_0)|_J \in C^1(J)$ holds, and moreover $u^s \in \hat{V}$. Hence, (9.13) holds on $\{\|(w,\sigma)\|_W < \rho\}$, and it follows by Lemma 5.1 that

$$(w,\sigma) \in W \longmapsto y(\bar{t},\cdot;u^s)|_{\hat{I}} \in C(\hat{I}) \tag{9.14}$$

is continuously Fréchet differentiable on $\{\|(w,\sigma)\|_W < \rho\}$. From Lemma 5.1 we have with Remark 5.4 that

$$\delta Y \overset{\text{def}}{=} (d_u Y(\cdot,u)(\delta w_0|_J, \delta w_1))|_S$$

is the broad solution of the linearized state equation (3.3) on S for initial data (3.4). Hence, the derivative of (9.14) in 0 is given by (3.7).

In Case CB^c Lemma 5.11 can be applied. Then $\bar{z} = x_i$ and u_0 is continuous (but not necessarily differentiable) in x_i. Let $\hat{I} =]x_-, x_+[\ni \bar{x}$, J with $\bar{z} = \xi(0) \in J$, and let $S_\mp$, $Y_\mp$, and $\hat{V}$ be given according to Lemma 5.11. Then for $\rho > 0$ sufficiently small we have $u^s \in \hat{V}$ on $\{\|(w,\sigma)\|_W < \rho\}$, since $u_0^s|_J = (u_0 + w_0)|_J \in PC^1(J;\bar{z})$. Hence,

$$(w,\sigma) \in W \longmapsto y(\bar{t},\cdot;u^s) \in L^r(\hat{I}) \tag{9.15}$$

is by Lemma 5.11 Fréchet differentiable at 0 for all $r \in [1,\infty[$, and the derivative coincides with the derivative of $(w,\sigma) \in W \longmapsto \tilde{y}(\bar{t},\cdot;u^s) \in L^r(\hat{I})$, where $\tilde{y}$ is defined in (5.28). Using (5.28), the derivative of (9.15) is thus clearly

$$d_u\tilde{y}(\bar{t},\cdot;u)\cdot(\delta w_0|_J, \delta w_1) = \delta Y_-(\bar{t},\cdot)|_{]x_-,\bar{x}[} + \delta Y_+(\bar{t},\cdot)|_{]\bar{x},x_+[} \overset{\text{def}}{=} \delta y^{\bar{t}}, \tag{9.16}$$

where with $S_{\mp,\xi} \overset{\text{def}}{=} S_\mp \cap \{\mp(x - \xi(t)) > 0\}$ and the prolongations $u_0^\mp$ and $w_0^\mp$ as in (5.25),

$$\delta Y_\mp = (d_u Y_\mp(\cdot; u_0^\mp|_J, u_1)\cdot(\delta w_0^\mp|_J, \delta w_1))|_{S_{\mp,\xi}}.$$

Since ξ is a genuine forward characteristic of $y(\cdot,u)$ through $(0,\bar{z})$, it is clear that $\delta Y_\mp$ depend in $S_{\mp,\xi}$ only on the values of $u_0, \delta w_0$ on $J \cap \{\mp(x-\bar{z}) > 0\}$, respectively, and are thus by Remark 5.4 on $S_{\mp,\xi}$ broad solutions of the linearized state equation (3.3) for initial values (3.4). Hence, (9.16) is exactly (3.7) on $\hat{I}$.

If $\bar{x}$ is a continuity point on a rarefaction wave (Case R^c) or on the boundary of a rarefaction wave (Case RB^c), then Lemma 5.10 or 5.12 is applicable. Now the Fréchet differentiability of (9.14) or (9.15) can be proven completely analogous. In Case R^c (9.14) is continuously Fréchet differentiable on $\{\|(w,\sigma)\|_W < \rho\}$, $\rho > 0$ small enough; in Case RB^c at least (9.15) is continuously Fréchet differentiable for all $r \in [1,\infty[$. On the rarefaction wave the broad solution δY of (3.3) in (3.7) is now defined according to Remark 5.9 and thus satisfies the initial condition (3.4).

Now combining the shift-differentiability of (9.5) in a neighborhood of a shock point and the Fréchet differentiability of (9.14) or (9.15), respectively, in a neighborhood of continuity points, the shift-differentiability of (3.2) in 0 is obvious by selecting a finite covering of I. Moreover, we have shown that $(\delta y^{\bar{t}}, \bar{s}) = T_s(0)\cdot(\delta w, s)$ is actually given by (3.7), (3.8). By using the local properties of the derivatives according to Lemmas 5.1 and 5.6 we conclude that $T_s(0) \in \mathcal{L}(W, PC(I;\tilde{x}_1,\dots,\tilde{x}_{\tilde{K}})\times\mathbb{R}^K)$ with $(\tilde{x}_k)$ comprising the shock locations and the points of class CB^c.

Consider finally the case that $u_0(x_i-) \neq u_0(x_i+)$ for all x_i. Then the scenario of Lemma 5.11, i.e., Case CB^c, does not occur. We have seen that (9.5) is continuously shift-differentiable and (9.14) in Cases C^c, R^c, or at least (9.15) for all $r \in [1,\infty[$ in Case RB^c are continuously Fréchet differentiable on $\{\|(w,\sigma)\|_W < \rho\}$, $\rho > 0$ sufficiently small. Hence, we obtain in fact continuous shift-differentiability of (3.2) on $\{\|(w,\sigma)\|_W < \rho\}$ by choosing a finite covering of I by intervals $\hat{I}$. □

Remark 9.1. *At the boundary of a rarefaction wave and along the backward characteristic through points $(\bar{t},\bar{x})$ of class CB^c, the broad solutions of the linearized equation according to Remark* 5.4 *or* 5.9 *cannot be pieced together continuously. Nevertheless, the obtained broad solution is a weak solution across the jump, since obviously the jump condition is satisfied.*

REMARK 9.2. *We have already mentioned in Remark* 3.5 *that weak solutions* δY *of the variational equation* (5.10) *can be defined on all of* $\Omega_{\bar{t}}$. *However, since the broad solutions to both sides of a shock do not in general satisfy the jump condition across shocks,* δY *is a measure with singular part at shocks. This reflects the fact that the mapping* $u \mapsto y(t,\cdot;u)$ *is at best differentiable in* $\mathcal{M}_{loc}(\mathbb{R})$–$w^*$. *For the conservative case (i.e.,* $g_y \equiv 0$, $g_{u_1} \equiv 0$*) the concept of duality solutions introduced in* [1] *yields solutions of this type; see also* [12]. *An analysis of duality solutions for the general variational equation* (5.10) *can be found in the author's habilitation thesis* [26]. *See also Remark* 3.7.

9.2. Differentiability of tracking-type functionals (proof of Theorem 3.9 and Corollary 3.10). Now since Theorems 3.2 and 3.4 are proven, also the differentiability results for tracking-type functionals (1.2), (3.19) in Theorem 3.9, and Corollary 3.10 in section 3.2 are shown. (Relying on Theorems 3.2 and 3.4 we have already proven them in section 3.2 by using Lemma 2.3.)

9.3. Adjoint calculus for tracking-type functionals (proof of Theorem 3.11). We are finally in the position to justify the adjoint-based gradient representation for tracking-type functionals (1.2), (3.19) stated in Theorem 3.11.

Proof of Theorem 3.11. The differentiability of J in (1.2), (3.19) is a direct consequence of Theorem 3.9. It remains to justify (3.24). We will show that (3.24) follows from (3.21).

We use the following notation: p denotes the reversible solution of (3.25)–(3.26) according to Theorem 7.11, which appears in (3.24). $\tilde{p}$ denotes the reversible solution of (3.25) for end data (3.23); then (3.21) holds with $\tilde{p}$ instead of p, i.e.,

$$\begin{aligned} d_{(w,\sigma)}J(y)\cdot(\delta w_0,\delta w_1,s) &= (\phi_y(y(\bar{t},\cdot),y_d),\delta y^{\bar{t}})_{2,I} \\ &\quad + (\tilde{p}g_{u_1}(\cdot,y,u_1),\delta w_1)_{2,\Omega_{\bar{t}}} + (\tilde{p}(0,\cdot),\delta w_0)_2 + \sum_{i=1}^{N}\tilde{p}(0,x_i)[u_0(x_i)]_+ s_i. \end{aligned} \tag{9.17}$$

Let D_k and S_k be defined as in Theorem 3.4. p and $\tilde{p}$ are by Definition 7.5 and Remark 7.4 transported along the backward characteristics. Hence, the values of the end data $p^{\bar{t}}$ and $\tilde{p}^{\bar{t}}$ on an interval $[x_-,x_+]$ determine the values of p and $\tilde{p}$ in the domain confined by the minimal/maximal backward characteristic through $x_\mp$, respectively. Thus, p vanishes outside the domain confined by the genuine backward characteristics through the continuity points $(\bar{t},a)$ and $(\bar{t},b)$.

Moreover, p and $\tilde{p}$ coincide on each domain D_k^{cl} and $\operatorname{supp}\tilde{p} \subset \bigcup_k D_k^{cl}$. Therefore, using the solution p for end data (3.26) we can rewrite (9.17) in the form

$$\begin{aligned} d_{(w,\sigma)}J(y)\cdot(\delta w,s) &= (p^{\bar{t}},\delta y^{\bar{t}})_{2,I} + \sum_{i=1}^{N} p(0,x_i)[u_0(x_i)]_+ s_i \\ &\quad + \sum_{k=1}^{K}\Big((p\,g_{u_1}(\cdot,y,u_1),\delta w_1)_{2,D_k} + (p(0,\cdot),\delta w_0)_{2,I_k}\Big), \end{aligned} \tag{9.18}$$

where $I_k \overset{\text{def}}{=} D_k^{cl} \cap \{t=0\}$, $k=1,\dots,K$. For the second term we have used that all down-jumps x_i are contained in the union of I_k, $k=1,\dots,K$.

We still have to rewrite the first expression on the right-hand side by means of the adjoint state. Fix an arbitrary stripe S_k. Since y is continuous on S_k, all characteristics are genuine. If S_k contains backward characteristics emanating at $t=\bar{t}$

from the finitely many discontinuities of y_d in I or forward characteristics emanating from $(0, x_i)$ with $u_0(x_i-) = u_0(x_i+)$, these characteristics divide S_k in finitely many substripes. Consider one of these substripes S. Then there are $x_- < x_+$ such that S is the domain confined by the maximal/minimal backward characteristic through $(\bar t, x_\mp)$, respectively, and Remark 5.4 applies with $Y = y(\cdot; u)$ outside of rarefaction waves, and Remark 5.9 applies with $Y_r = y(\cdot; u)$ inside of rarefaction waves. By the choice of S we have $y(\bar t, .), y_d \in C^{0,1}(]x_-, x_+[)$, and thus the function $p^{\bar t}$ in (3.26) is in $C^{0,1}(]x_-, x_+[)$. Since the reversible solution p of (3.25) depends on S only on $p^{\bar t}|_{]x_-,x_+[}$, we conclude exactly as in the proof of Lemma 6.4 that $p|_S \in C^{0,1}(S \cap \{t > \tau\}) \cap C([0, \bar t], L^2_{loc}(\mathbb{R}))$ for all $\tau > 0$. Remarks 5.4 and 5.9 thus yield (5.11) for $D = S \cap \{t > \tau\}$. Since the L^2-traces of p and y exist at $t = 0$ and $p|_S$ satisfies (3.25) a.e. by Theorem 7.7, (5.11) gives for $\tau \to 0$

$$(p^{\bar t}, \delta y^{\bar t})_{2,]x_-,x_+[} = (p(0,\cdot), \delta w_0)_{2,S^{cl}\cap\{t=0\}} + (p\, g_{u_1}(t, x, Y, u_1), \delta w_1)_{2,S};$$

see (3.18) and Remark 3.6. The boundary terms along the confining characteristics $\xi_\mp$ drop out, since the outer normal $(n_1, n_2)^T$ is a multiple of $(f'(y(t, \xi_\mp(t))), -1)^T$. Summing over all stripes S and inserting the result in (9.18) yields (3.24).

We already know from Theorem 7.10 that for all $\tau > 0$, $p|_{D_k} \in C^{0,1}(D_k \cap \{t > \tau\})$ holds and that $p|_{S_k}$ is piecewise $C^{0,1}$ on $S_k \cap \{t > \tau\}$ with discontinuities along the backward characteristics emanating from discontinuities of y_d and that the same holds for $\tau = 0$ whenever S_k or D_k contains no rarefaction wave. ∎

9.4. Nondegeneracy of shocks (proof of Theorem 3.8). We finally show that the nondegeneracy assumptions of Theorem 3.2 hold for a.a. $\bar t \in [0, T]$ under slightly stronger smoothness assumptions on (u_0, u_1).

Proof of Theorem 3.8. By [8, Lem. 4.1, Cor. 4.2] the set of shocks is at most countable, and for each shock η the functions $t \mapsto y(t, \eta(t)\mp)$ are continuous outside of an at most countable set. Thus, for all except countably many times $\bar t \in]0, T[$ for all shocks η the functions $t \mapsto y(t, \eta(t)\mp)$ are continuous at $\bar t$ and the extreme backward characteristics through $(\bar t, \eta(\bar t))$ do not propagate at the boundary of a rarefaction wave or reach $t = 0$ in one of the points x_i, $i = 1, \dots, N$, where u_0 is continuous. Denote by $B \subset]0, T[$ the set of all $\bar t$ having this property. Fix some $\bar t \in B$. If all shocks satisfy the framework of Lemma 6.2 or Corollary 6.5, i.e., are nondegenerate according to Definition 6.1, then $y(\bar t, \cdot)$ has obviously only finitely many shocks on each compact interval I and Theorem 3.2 is applicable.

Now assume that there exists a shock η and a set $R \subset B$ with outer measure $\mu^*(R) > 0$ such that η for all $\bar t \in R$ does not satisfy the framework of Lemma 6.2 or Corollary 6.5. Fix an arbitrary density point $\hat t \in R$ of R w.r.t. μ^*.

Assume for concreteness that η has no rarefaction wave as left or right state at time $\hat t$. Since $t \mapsto y(t, \eta(t)\mp)$ are continuous at $\hat t$, the left or right state is also not a rarefaction wave for all times t in a sufficiently close neighborhood S of $\hat t$, and since $\hat t \in R$ is a density point of R, we may reduce R such that $R = R \cap S$. By assumption, (4.19) must be violated for all $\bar t \in R$ with $\bar z$ being one of the intersection points $z_\mp = \xi_\mp(0; \bar t, \eta(\bar t))$ of the minimal or maximal backward characteristic of y through $(\bar t, \eta(\bar t))$ with $\{t = 0\}$. Without restriction assume that for all $\bar t \in R$ (4.19) is violated on the left side of the shock, i.e., $z(\bar t) \overset{\text{def}}{=} \bar z = \xi_-(0; \bar t, \eta(\bar t))$. Then with $F(t, z) \overset{\text{def}}{=} \frac{d}{dz}\zeta(t; z, u_0(z), u_1)$ we must have

$$F(\bar t, z(\bar t)) = 0, \quad F(t, z(\bar t)) > 0, \quad 0 \le t < \bar t \quad \forall\, \bar t \in R. \tag{9.19}$$

Since genuine backward characteristics may not intersect, the mapping $t \in S \mapsto z(t)$ is strictly monotone decreasing and thus almost everywhere differentiable. Without restriction, we may thus reduce R such that $z(t)$ is differentiable in all $\bar{t} \in R$ and still $\mu^*(R) > 0$. Now let $\tilde{t} \in R$ be a density point of R w.r.t. μ^* and set $\tilde{z} = z(\tilde{t})$. Since $F(t,z) = \delta\zeta(t; z, u_0(z), u_1; 1, u_0'(z), 0)$ we obtain from (4.12) that F is continuously differentiable w.r.t. t, and since u_0 is C^2 in a neighborhood of $\tilde{z}$ we conclude by using Lemma 4.4 and applying Proposition 4.3 to (4.12) that $(t,z) \mapsto \zeta(t; z, u_0(z), u_1)$ is C^2 in a neighborhood Q of $(\tilde{t}, \tilde{z})$. Moreover, we know from section 4.3 that $\partial_t F(\tilde{t}, \tilde{z}) \neq 0$ and hence by (9.19) that $\partial_t F(\tilde{t}, \tilde{z}) < 0$; cf. also [8, Lem. 5.5]. Since $\tilde{t} \in R$ is a density point of R, there exists a strictly monotone increasing sequence $(t_k) \subset R$ with $t_k \nearrow \tilde{t}$. By the continuity of $t \mapsto y(t, \eta(t)-)$ in $\tilde{t}$ and the stability of genuine backward characteristics we have $z(t_k) \searrow z(\tilde{t}) = \tilde{z}$ as $k \to \infty$ and thus eventually $(t_k, z(t_k)) \in Q$. Now we get by (9.19) and $\partial_t F(\tilde{t}, \tilde{z}) < 0$

$$0 = \lim_{k\to\infty} \frac{F(t_k, z(t_k)) - F(\tilde{t}, z(\tilde{t}))}{t_k - \tilde{t}} = \partial_t F(\tilde{t}, \tilde{z}) + \partial_z F(\tilde{t}, \tilde{z})\dot{z}(\tilde{t}) < \partial_z F(\tilde{t}, \tilde{z})\dot{z}(\tilde{t})$$

and deduce from $\dot{z}(\tilde{t}) \leq 0$ that $\partial_z F(\tilde{t}, \tilde{z}) = \frac{d^2}{dz^2}\zeta(\tilde{t}; z, u_0(z), u_1)|_{z=\tilde{z}} < 0$. Hence, we obtain from $z(t_k) \searrow z(\tilde{t}) = \tilde{z}$ that for all sufficiently large k

$$\zeta(\tilde{t}; z(t_k), u_0(z(t_k)), u_1) < \zeta(\tilde{t}; z(\tilde{t}), u_0(z(\tilde{t})), u_1) = \eta(\tilde{t}),$$

i.e., the continuation of the genuine characteristic through $(t_k, \eta(t_k))$ intersects the genuine backward characteristic through $(\tilde{t}, \eta(\tilde{t}))$ at some time $t \in]t_k, \tilde{t}[$ and must thus intersect the shock η a second time on $]t_k, \tilde{t}[$. This is impossible, since the angle between the genuine backward characteristics and η is bounded away from zero independently of k, and also the curvature of the continuation of the characteristics is uniformly bounded. Hence, the assumption was wrong and the hypotheses of Lemma 6.2 cannot be violated on a set of nonzero measure. Applying a similar argument in the case of a rarefaction wave as left or right state concludes the proof. ∎

10. Conclusions and future work. We have presented a sensitivity calculus for entropy solutions of hyperbolic conservation laws with source terms that is based on a first order approximation by shift-variations. The obtained shift-differentiability result for the control-to-state mapping implies differentiability properties for a large class of tracking-type functionals. For this class of functionals we have derived a gradient representation by using the adjoint state. Hereby, the adjoint state is the unique reversible solution of a transport equation with discontinuous coefficient that guarantees uniqueness only for the class of reversible solutions. These results can be used to state optimality conditions for the optimal control of flows with shocks and provide an analytical justification for the use of gradient-based methods. In particular, it turns out that the numerical scheme used to compute the adjoint state should converge to the reversible solution in order to be consistent with the original problem. Since we allow shift-variations of the initial data, the shift-differentiability result can be repeatedly used over time slabs. We plan to exploit this in the design and analysis of a class of SQP methods with time domain decomposition.

Our analysis uses structural results for the state and the stability of reversible solutions of the adjoint equation with respect to its coefficients. This approach can formally be extended to systems and multidimensional problems and yields a correct sensitivity and adjoint calculus if the necessary stability properties of state and adjoint state actually hold.

The results of section 7 on the adjoint equation are discussed in detail in the follow-up paper [27]; see also the author's habilitation thesis [26]. In [26] we also extend the results of [1] on duality solutions for the corresponding forward problem to the case with source term, which yields the correct linearization of the state equation.

We plan to use the adjoint calculus of this paper for the formulation of optimality conditions and the analysis of gradient-based methods for the optimal control of flows with shocks. For the numerical approximation we have started to analyze which schemes for the discretization of the state equation lead to numerical adjoint schemes that converge to the correct reversible adjoint state, thus yielding consistent gradient approximations. First results on the discretization of backward transport equations with discontinuous coefficients and the corresponding conservative forward problem were obtained for the case without source term by Gosse and James in the very recent paper [15]. By extending these results, we have recently [26] obtained convergence results for discrete adjoints and the corresponding gradient approximations.

Appendix. Analysis of the adjoint equation. For convenience we present proofs of the existence and stability results for the adjoint equation that we stated in section 7. For a detailed analysis we refer to our recent works [27, 26]. In [26] we consider also measure solutions for the corresponding conservative forward problem with source term.

One can show [1] that the generalized backward flow X satisfies

(A.1)
$$\|\partial_s X\|_{\infty,\overset{\circ}{D}_b\times\mathbb{R}} \le \|a\|_\infty, \quad \|\partial_t X\|_{\infty\times\mathbb{R},\overset{\circ}{D}_b} \le \|a\|_\infty e^{\int_0^T \alpha}, \quad \|\partial_x X(s;t,\cdot)\|_\infty \le e^{\int_t^s \alpha},$$

where $(s,t) \in D_b$ in the last inequality. Moreover,

$$\partial_s \partial_x X(s;t,x) \le \alpha(s)\partial_x X(s;t,x) \quad \text{for a.a. } s \in]0,T[\tag{A.2}$$

on $\overset{\circ}{D}_b \times\mathbb{R}$. Thus, for arbitrary $z_1 < z_2$ and $0 \le t \le \sigma \le s \le T$

$$X(s;t,z_2) - X(s;t,z_1) \le X(\sigma;t,z_2) - X(\sigma;t,z_1) + \int_\sigma^s \alpha(\tau)(X(\tau;t,z_2) - X(\tau;t,z_1))d\tau,$$

and hence

$$X(s;t,z_2) - X(s;t,z_1) \le (X(\sigma;t,z_2) - X(\sigma;t,z_1))e^{\int_\sigma^s \alpha} \quad \forall\, t \le \sigma \le s \le T. \tag{A.3}$$

Proof of Theorem 7.7. We show first that p is well defined. Let $(t,x) \in \Omega_T$ be arbitrary. Since $z \longmapsto X(t;0,z)$ is surjective, there is $z \in \mathbb{R}$ with $x = X(t;0,z)$. Now (7.7) defines the values of p on the curve $(s, X(s;0,z))$, $t \le s \le T$. If z is not unique, then we get for all $\tilde{z}$ with $x = X(t;0,\tilde{z})$ by (A.3) $X(s;0,\tilde{z}) = X(s;0,z)$ for all $s \in [t,T]$. Hence, the definition does not depend on the choice of z. As a simple consequence of (A.1) and (7.7) we get (7.9). Hence, there exists a unique $p \in L^\infty(\Omega_T)$ satisfying (7.7). We show that p is Lipschitz continuous. Let $z_1 < z_2$ be arbitrary. Then $\Delta p(t) \overset{\text{def}}{=} p(t, X(t;0,z_2)) - p(t, X(t;0,z_1))$ satisfies $\Delta p(T) = p^T(X(T;0,z_2)) - p^T(X(T;0,z_1))$ and

$$\tfrac{d}{dt}\Delta p(t) = -b(t, X(t;0,z_2))\Delta p(t) - (b(t, X(t;0,z_2)) - b(t, X(t;0,z_1)))p(t, X(t;0,z_1)).$$

Thus, setting $I(t) = [X(t;0,z_1), X(t;0,z_2)]$ we get for all $\tau \in [0,T]$

$$\begin{aligned}|\Delta p(\tau)| \le{}& \int_\tau^T \|b(s)\|_{\infty,I(s)} |\Delta p(s)|\,ds + |p^T(X(T;0,z_2)) - p^T(X(T;0,z_1))| \\ &+ \int_\tau^T \|p(s)\|_{\infty,I(s)} |b(s,X(s;0,z_2)) - b(s,X(s;0,z_1))|\,ds.\end{aligned} \tag{A.4}$$

Hence, we have by (A.3) with $\Delta X(t) \stackrel{\text{def}}{=} X(t;0,z_2) - X(t;0,z_1)$

$$|\Delta p(\tau)| \le \left(\|\partial_x p^T\|_\infty + \|b\|_{L^1(0,T;C^{0,1})}\|p\|_\infty\right) e^{\int_\tau^T \alpha} \Delta X(\tau) + \|b\|_\infty \int_\tau^T |\Delta p(s)|\,ds$$

and by the Gronwall lemma for all $t \in [0,T]$

$$|\Delta p(t)| \le \Delta X(t) \left(\|\partial_x p^T\|_\infty + \|b\|_{L^1(0,T;C^{0,1})}\|p\|_\infty\right) e^{(T-t)\|b\|_\infty + \int_t^T \alpha}.$$

This yields a uniform bound for $\|\partial_x p\|_\infty$. Finally, let $z \in \mathbb{R}$ and $t_1, t_2 \in [0,T]$, $t_1 < t_2$, be arbitrary; then by (A.1)

$$\begin{aligned}&|p(t_2, X(t_1;0,z)) - p(t_1, X(t_1;0,z))| \\ &\le |p(t_2, X(t_2;0,z)) - p(t_1, X(t_1;0,z))| + |p(t_2, X(t_2;0,z)) - p(t_2, X(t_1;0,z))| \\ &\le (\|b\|_\infty \|p\|_\infty + \|a\|_\infty \|\partial_x p\|_\infty)|t_2 - t_1|.\end{aligned} \tag{A.5}$$

Hence, $p \in C^{0,1}(\Omega_T)$. Finally, p solves (7.1) a.e. in Ω_T. In fact, for a.a. $(t,x) = (t, X(t;0,z))$ in Ω_T the Lipschitz-function p is differentiable. Moreover, since $a(t,\cdot) \in BV_{loc}(\mathbb{R})$ for a.a. t by the one-sided Lipschitz condition, we have $a(t,x-) = a(t,x+)$ for a.a. $(t,x) \in \Omega_T$, and thus from (7.6) we have that

$$\partial_s X(t;0,z) = a(t, X(t;0,z))$$

for a.a. $(t,x) = (t, X(t;0,z))$. Now the chain rule yields with (7.7) that (7.1) is satisfied for all of these (t,x).

Since for $w \in C^{0,1}(\mathbb{R})$, $|w|_{var} = \|\partial_x w\|_1$ holds, we obtain by summing (A.4) for suitable $z_1^0 < z_2^0 = z_1^1 < z_2^1 = \cdots$ that for all $0 \le t \le \tau \le T$

$$\begin{aligned}\|\partial_x p(\tau)\|_{1,I(\tau)} \le{}& \int_\tau^T \|b(s)\|_{\infty,I(s)} \|\partial_x p(s)\|_{1,I(s)}\,ds + \|\partial_x p^T\|_{1,I(T)} \\ &+ \int_\tau^T \|p(s)\|_{\infty,I(s)} \|\partial_x b(s)\|_{1,I(s)}\,ds.\end{aligned}$$

With $I = I(t)$ and $J = [X(t;0,z_1) - \|a\|_\infty (T-t), X(t;0,z_2) + \|a\|_\infty (T-t)]$, we have by (A.1) that $I(s) \subset J$, $t \le s \le T$, and by the Gronwall lemma it follows that

$$\|\partial_x p(t)\|_{1,I} \le (\|\partial_x p^T\|_{1,J} + \|\partial_x b\|_{1,[t,T]\times J} \|p\|_{\infty,[t,T]\times J}) e^{(T-t)\|b\|_{\infty,[0,T]\times J}}. \tag{A.6}$$

Since p solves (7.1) a.e. in Ω_T, we get for arbitrary $0 < t_1 < t_2 < T$ and with $|I|$ denoting the length of I

$$\begin{aligned}\|\partial_t p\|_{1,[t_1,t_2]\times I} &\le \|bp\|_{1,[t_1,t_2]\times I} + \|a\|_{\infty,[t_1,t_2]\times I} \|\partial_x p\|_{1,[t_1,t_2]\times I} \\ &\le (t_2 - t_1)(|I| \|bp\|_{\infty,[t_1,t_2]\times I} + \|a\|_{\infty,[t_1,t_2]\times I} \|\partial_x p\|_{L^\infty(t_1,t_2;L^1(I))}).\end{aligned} \tag{A.7}$$

Now (7.10) and the asserted properties of the constant C follow directly from (7.9), (A.6), and (A.7). Moreover, we see that $p \in H^{1,1}(]0,T[\times I)$ with norm not depending on α. ∎

Proof of Theorem 7.8. Denote by X and X_n the backward flows according to Definition 7.3 for a and a_n, respectively. By [1] it holds that $X_n \to X$ in $C(D_b \times [-R,R])$ for any $R > 0$. By the definition of reversible solutions we have for all $z \in \mathbb{R}$

$$\begin{aligned} p_n(T, X_n(T;0,z)) &= p_n^T(X_n(T;0,z)), \\ \tfrac{d}{dt} p_n(t, X_n(t;0,z)) &= -(b_n p_n)(t, X_n(t;0,z)). \end{aligned}$$

For the reversible solution p of (7.1), equation (7.7) holds. Fix some $R > 0$ and consider an arbitrary $(t,x) \in [0,T] \times [-R,R]$. Then there exist $z, z_n \in \mathbb{R}$ with $x = X(t;0,z) = X_n(t;0,z_n)$, and we have $X(s;0,z), X_n(s;0,z_n) \in [R - M_a T, R + M_a T] \stackrel{\text{def}}{=} J$ according to (A.1) for all $s \in [t,T]$ with an upper bound M_a for $\|a_n\|_\infty$ and $\|a\|_\infty$. Since $X(s;t,x) = X(s;0,z)$ and $X_n(s;t,x) = X_n(s;0,z)$ by (7.5), we have

$$\begin{aligned} p_n(T, X_n(T;t,x)) &= p_n^T(X_n(T;t,x)), \\ \tfrac{d}{ds} p_n(s, X_n(s;t,x)) &= -(b_n p_n)(s, X_n(s;t,x)), \quad s \in]t,T[, \end{aligned}$$

and the same holds with p, X, p^T, b instead of p_n, X_n, p_n^T, b_n. Therefore, the difference $\Delta p_n(s) \stackrel{\text{def}}{=} p(s, X(s;t,x)) - p_n(s, X_n(s;t,x))$ satisfies

$$\begin{aligned} |\Delta p_n(T)| &= |p^T(X(T;t,x)) - p_n^T(X_n(T;t,x))| \\ &\le \|p^T - p_n^T\|_{C(J)} + \|\partial_x p^T\|_{\infty,J} \|X - X_n\|_{C(D_b \times J)} \end{aligned}$$

and for $s \in]t,T[$

$$\begin{aligned} \tfrac{d}{ds} \Delta p_n(s) = {} & (b_n(s, X_n(s;t,x)) - b(s, X(s;t,x))) p(s, X(s;t,x)) \\ & - b_n(s, X_n(s;t,x)) \Delta p_n(s). \end{aligned}$$

Thus, we get with $J_T \stackrel{\text{def}}{=} [0,T] \times J$

$$\begin{aligned} |\Delta p_n(s)| \le {} & T\|p\|_{C(J_T)} (\|b_n - b\|_{L^\infty(0,T;C(J))} + \|\partial_x b\|_{\infty,J_T} \|X - X_n\|_{C(D_b \times J)}) \\ & + |\Delta p_n(T)| + \int_s^T \|b_n\|_{L^\infty(0,T;C(J))} |\Delta p_n(\tau)| \, d\tau, \end{aligned}$$

and we deduce by the Gronwall lemma that

$$\lim_{n \to \infty} \|p - p_n\|_{C([0,T] \times [-R,R])} = 0. \qquad \blacksquare$$

Acknowledgments. The author would like to thank two anonymous referees for their insightful comments that helped to improve the presentation of the paper.

A part of this research was done while the author was visiting the Department of Computational and Applied Mathematics and the Center for Research on Parallel Computation at Rice University. The author would like to thank John E. Dennis and Matthias Heinkenschloss for their hospitality and support.

The author would like to thank his brother Michael (Technische Universität München) for helpful comments on an earlier draft of the paper and Matthias Heinkenschloss (Rice University) for interesting discussions on the material of the paper. Moreover, he is indebted to Günter Leugering (Technische Universität Darmstadt) for his interest and support.

REFERENCES

[1] F. Bouchut and F. James, *One-dimensional transport equations with discontinuous coefficients*, Nonlinear Anal., 32 (1998), pp. 891–933.

[2] F. Bouchut and F. James, *Differentiability with respect to initial data for a scalar conservation law*, in Hyperbolic Problems: Theory, Numerics, Applications, Vol. I (Zürich, 1998), Internat. Ser. Numer. Math. 129, Birkhäuser, Basel, 1999, pp. 113–118.

[3] A. Bressan and G. Guerra, *Shift-differentiability of the flow generated by a conservation law*, Discrete Contin. Dynam. Systems, 3 (1997), pp. 35–58.

[4] A. Bressan and A. Marson, *A variational calculus for discontinuous solutions of systems of conservation laws*, Comm. Partial Differential Equations, 20 (1995), pp. 1491–1552.

[5] E. M. Cliff, M. Heinkenschloss, and A. R. Shenoy, *An optimal control problem for flows with discontinuities*, J. Optim. Theory Appl., 94 (1997), pp. 273–309.

[6] E. M. Cliff, M. Heinkenschloss, and A. R. Shenoy, *Adjoint-based methods in aerodynamic design-optimization*, in Computational Methods for Optimal Design and Control (Arlington, VA, 1997), Progr. Systems Control Theory 24, Birkhäuser Boston, Boston, 1998, pp. 91–112.

[7] E. D. Conway, *Generalized solutions of linear differential equations with discontinuous coefficients and the uniqueness question for multidimensional quasilinear conservation laws*, J. Math. Anal. Appl., 18 (1967), pp. 238–251.

[8] C. M. Dafermos, *Generalized characteristics and the structure of solutions of hyperbolic conservation laws*, Indiana Univ. Math. J., 26 (1977), pp. 1097–1119.

[9] A. F. Filippov, *Differential equations with discontinuous right-hand side*, Amer. Math. Soc. Transl. (2), 42 (1964), pp. 199–231.

[10] P. D. Frank and G. R. Shubin, *A comparison of optimization-based approaches for a model computational aerodynamics design problem*, J. Comput. Phys., 98 (1992), pp. 74–89.

[11] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser-Verlag, Basel, 1984.

[12] E. Godlewski and P.-A. Raviart, *The linearized stability of solutions of nonlinear hyperbolic systems of conservation laws. A general numerical approach*, Math. Comput. Simulation, 50 (1999), pp. 77–95.

[13] E. Godlewski, M. Olazabal, and P.-A. Raviart, *On the linearization of hyperbolic systems of conservation laws. Application to stability*, in Équations aux dérivées partielles et applications, Gauthier–Villars, Éd. Sci. Méd. Elsevier, Paris, 1998, pp. 549–570.

[14] M. Golubitsky and D. G. Schaeffer, *Stability of shock waves for a single conservation law*, Adv. Math., 16 (1975), pp. 65–71.

[15] L. Gosse and F. James, *Numerical approximations of one-dimensional linear conservation equations with discontinuous coefficients*, Math. Comp., 69 (2000), pp. 987–1015.

[16] D. Hoff, *The sharp form of Oleĭnik's entropy condition in several space variables*, Trans. Amer. Math. Soc., 276 (1983), pp. 707–714.

[17] F. James and M. Sepúlveda, *Convergence results for the flux identification in a scalar conservation law*, SIAM J. Control Optim., 37 (1999), pp. 869–891.

[18] S. N. Kružkov, *First order quasilinear equations in several independent variables*, Math. USSR Sbornik, 10 (1970), pp. 217–243.

[19] P. D. Lax, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[20] Ph. LeFloch and Z. P. Xin, *Uniqueness via the adjoint problems for systems of conservation laws*, Comm. Pure Appl. Math., 46 (1993), pp. 1499–1533.

[21] O. A. Oleinik, *Discontinuous solutions of nonlinear differential equations*, Amer. Math. Soc. Transl. (2), 26 (1963), pp. 95–172.

[22] E. Polak, *Optimization. Algorithms and Consistent Approximations*, Springer-Verlag, New York, 1997.

[23] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, Berlin, 1983.

[24] E. Tadmor, *Local error estimates for discontinuous solutions of nonlinear hyperbolic equations*, SIAM J. Numer. Anal., 28 (1991), pp. 891–906.

[25] S. Ulbrich, *On the existence and approximation of solutions for the optimal control of nonlinear hyperbolic conservation laws*, in Optimal Control of Partial Differential Equations (Chemnitz, 1998), Internat. Ser. Numer. Math. 133, Birkhäuser, Basel, 1999, pp. 287–299.

[26] S. Ulbrich, *Optimal Control of Nonlinear Hyperbolic Conservation Laws with Source Terms*, Habilitation Thesis, Zentrum Mathematik, Technische Universität München, Germany, 2001.

[27] S. Ulbrich, *Adjoint-based derivative computations for the optimal control of discontinuous solutions of hyperbolic conservation laws*, Systems Control Lett., to appear.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 3, pp. 798–819

ON THE CONTROLLABILITY OF PARABOLIC SYSTEMS WITH A NONLINEAR TERM INVOLVING THE STATE AND THE GRADIENT*

A. DOUBOVA†, E. FERNÁNDEZ-CARA†, M. GONZÁLEZ-BURGOS†, AND E. ZUAZUA‡

Abstract. We present some results concerning the controllability of a quasi-linear parabolic equation (with linear principal part) in a bounded domain of $\mathbb{R}^N$ with Dirichlet boundary conditions. We analyze the controllability problem with distributed controls (supported on a small open subset) and boundary controls (supported on a small part of the boundary). We prove that the system is null and approximately controllable at any time if the nonlinear term $f(y, \nabla y)$ grows slower than $|y| \log^{3/2}(1+|y|+|\nabla y|)+|\nabla y| \log^{1/2}(1+|y|+|\nabla y|)$ at infinity (generally, in this case, in the absence of control, blow-up occurs). The proofs use global Carleman estimates, parabolic regularity, and the fixed point method.

Key words. controllability, parabolic equations, nonlinear gradient terms

AMS subject classifications. 93B05, 35K55, 35K05

PII. S0363012901386465

1. Introduction and main results. Let $\Omega \subset \mathbb{R}^N$ be a bounded connected open set with boundary $\partial\Omega$ of class C^2. Let $\mathcal{O} \subset \Omega$ be a nonempty open subset, let $\gamma \subset \partial\Omega$ be a nonempty relative open subset of the boundary, and assume that $T > 0$. We will use the following notation: $Q = \Omega \times (0,T)$, $\Sigma = \partial\Omega \times (0,T)$. For any $p \in [1, +\infty]$, we will denote by $\|\cdot\|_p$ the usual norm in $L^p(Q)$.

We will consider parabolic systems of the form

$$\left\{ \begin{array}{lll} \partial_t y - \Delta y + f(y, \nabla y) = v 1_{\mathcal{O}} & \text{in} & Q, \\ y = 0 & \text{on} & \Sigma, \\ y(x,0) = y_0(x) & \text{in} & \Omega \end{array} \right. \tag{1}$$

and

$$\left\{ \begin{array}{lll} \partial_t y - \Delta y + f(y, \nabla y) = 0 & \text{in} & Q, \\ y = v 1_{\gamma} & \text{on} & \Sigma, \\ y(x,0) = y_0(x) & \text{in} & \Omega, \end{array} \right. \tag{2}$$

where y_0 and v are given in appropriate spaces. In (1) and (2),

$$f : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}$$

is a locally Lipschitz-continuous function and $1_{\mathcal{O}}$ and 1_{γ} denote the characteristic functions of the sets $\mathcal{O}$ and γ, respectively. We will assume that $y_0 \in W^{1,\infty}(\Omega) \cap H_0^1(\Omega)$ (for simplicity), $v \in L^\infty(\mathcal{O} \times (0,T))$ in (1), and $v \in L^\infty(\gamma \times (0,T))$ in (2).

*Received by the editors March 14, 2001; accepted for publication (in revised form) January 17, 2002; published electronically August 8, 2002. This work was supported by grants PB96-0663, PB98-1134, and BFM2000-1317 of the DGES (Spain).

http://www.siam.org/journals/sicon/41-3/38646.html

†Dpto. Ecuaciones Diferenciales y Análisis Numérico, Universidad de Sevilla, Aptdo. 1160, 41080 Sevilla, Spain (dubova@numer.us.es, cara@numer.us.es, burgos@numer.us.es).

‡Dpto. Matemática Aplicada, Universidad Complutense, 28040 Madrid, Spain (zuazua@eucmax.sim.ucm.es).

The main goal of this paper is to analyze the controllability properties of (1) and (2). It will be said that (1) (resp., (2)) is null-controllable at time T if, for each $y_0 \in W^{1,\infty}(\Omega) \cap H_0^1(\Omega)$ (resp., $y_0 \in W^{1,\infty}(\Omega) \cap V$, where V is given below by (10)), there exists $v \in L^\infty(\mathcal{O} \times (0,T))$ (resp., $v \in L^\infty(\gamma \times (0,T))$) such that the corresponding initial boundary problem (1) (resp., (2)) admits a solution $y \in C^0([0,T]; L^2(\Omega))$ satisfying

$$y(x,T) = 0 \quad \text{in} \quad \Omega. \tag{3}$$

On the other hand, it will be said that (1) (resp., (2)) is approximately controllable in $L^2(\Omega)$ at time T if, for any $y_0 \in W^{1,\infty}(\Omega) \cap H_0^1(\Omega)$ (resp., $y_0 \in W^{1,\infty}(\Omega) \cap V$), any $y_d \in L^2(\Omega)$, and any $\varepsilon > 0$, there exists a control $v \in L^\infty(\mathcal{O} \times (0,T))$ (resp., $v \in L^\infty(\gamma \times (0,T))$) such that the corresponding initial boundary problem (1) (resp., (2)) possesses a solution $y \in C^0([0,T]; L^2(\Omega))$, with

$$\|y(\cdot,T) - y_d\|_{L^2} \le \varepsilon. \tag{4}$$

The controllability of linear and semilinear parabolic systems has been analyzed in several recent papers. Among them, let us mention [I], [FI], [F], [B], [AB], and [FZ2] in what concerns null controllability and [FPZ], [Z2], and [FZ2] for approximate controllability.

This paper generalizes all previous results, in particular those in [FZ2], where the nonlinear term is assumed to be of the form $f(y)$.

Notice that, under the hypothesis above, we can write

$$f(s,p) = f(0,0) + g(s,p)s + G(s,p) \cdot p \qquad \forall (s,p) \in \mathbb{R} \times \mathbb{R}^N \tag{5}$$

for some L^∞_{loc} functions g and G. These are respectively given by

$$g(s,p) = \int_0^1 \frac{\partial f}{\partial s}(\lambda s, \lambda p)\, d\lambda, \qquad G_i(s,p) = \int_0^1 \frac{\partial f}{\partial p_i}(\lambda s, \lambda p)\, d\lambda \quad \text{for } 1 \le i \le N.$$

Our first result is the following one.

THEOREM 1.1. *Assume that f is locally Lipschitz-continuous, $f(0,0) = 0$ and*

$$\lim_{|(s,p)|\to\infty} \frac{|g(s,p)|}{\log^{3/2}(1+|s|+|p|)} = 0, \quad \lim_{|(s,p)|\to\infty} \frac{|G(s,p)|}{\log^{1/2}(1+|s|+|p|)} = 0. \tag{6}$$

Then (1) is null-controllable at any time $T > 0$.

REMARK 1.1. *This result generalizes at least two cases that have been studied exhaustively before. First, the case of a globally Lipschitz-continuous function f, i.e., when $g \in L^\infty(\mathbb{R} \times \mathbb{R}^N)$ and $G \in L^\infty(\mathbb{R} \times \mathbb{R}^N)^N$. In this case, f is a function with sublinear behavior at infinity, and the proof of the corresponding controllability result is easier (cf.* [IY]*). Second, the case where $G \equiv 0$ and $g = g(s)$ satisfies $g(0) = 0$ and*

$$\lim_{|s|\to\infty} \frac{|g(s)|}{\log^{3/2}(1+|s|)} = 0. \tag{7}$$

The proof is again easier (cf. [FZ2]*).*

REMARK 1.2. *In* [FZ2]*, it is proved that, for each $\beta > 2$, there exist functions $f = f(s)$ with $f(0) = 0$ and*

$$|f(s)| \sim |s| \log^\beta (1+|s|) \quad \text{as} \quad |s| \to \infty \tag{8}$$

such that (1) *is not null-controllable for all* $T > 0$. *In view of Theorem* 1.1, *we see that when* f *satisfies* (8) *with* $3/2 \le \beta \le 2$, *the null controllability problem of* (1) *is an open question.*

REMARK 1.3. *Theorem* 1.1 *says in particular that, under assumption* (6), *for each* y_0 *there exists a control* v *such that* (1) *possesses a solution globally defined in* $[0,T]$. *This claim is not true for any right-hand side and any* $y_0 \in W^{1,\infty}(\Omega) \cap H_0^1(\Omega)$, *since we are in the range in which blow-up may occur (for instance, see* [CH]*).*

A consequence of Theorem 1.1 is the approximate controllability of (1). In this case, $f(0,0)$ will not be necessarily 0 and we will assume that f verifies (5) and (9), a condition slightly different from (6). Thus, our second main result is the following.

THEOREM 1.2. *Let* $T > 0$. *Assume that* $f : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}$ *is locally Lipschitz-continuous and verifies*

$$\begin{cases} \displaystyle\lim_{|(s,p)|\to\infty} \frac{1}{\log^{3/2}(1+|s|+|p|)} \left| \int_0^1 \frac{\partial f}{\partial s}(s_0+\lambda s, p_0+\lambda p)\, d\lambda \right| = 0, \\ \displaystyle\lim_{|(s,p)|\to\infty} \frac{1}{\log^{1/2}(1+|s|+|p|)} \left| \int_0^1 \frac{\partial f}{\partial p_i}(s_0+\lambda s, p_0+\lambda p)\, d\lambda \right| = 0 \end{cases} \tag{9}$$

uniformly in $(s_0, p_0) \in K$ *for every compact set* $K \subset \mathbb{R} \times \mathbb{R}^N$. *Then* (1) *is approximately controllable at time* T.

REMARK 1.4. *It will be seen in section* 4 *that, for systems like* (1), *the approximate controllability result is actually a consequence of the exact controllability to the trajectories in* $C^0([0,T]; W^{1,\infty}(\Omega))$.

REMARK 1.5. *Again, Theorem* 1.2 *generalizes two known results. First, the case where* f *is globally Lipschitz-continuous, i.e., the case in which* $\partial f/\partial s$ *and* $\partial f/\partial p_i$ $(1 \le i \le N)$ *are uniformly bounded (cf.* [Z2]*). On the other hand, Theorem* 1.2 *is also a generalization of the approximate controllability result in* [FZ2], *where* $G \equiv 0$, $g = g(s)$, *and* (7) *is satisfied.*

In the first draft of this paper (and also in the approximate controllability results in [FZ2]*), an additional assumption was imposed in Theorem* 1.2, *namely, the existence of a globally defined solution* y^* *corresponding to appropriate data* y_0^* *and* v^*. *But one of the referees provided an argument that shows that this hypothesis is in fact unnecessary (see the proof of Theorem* 1.2 *in section* 4*).*

REMARK 1.6. *In particular,* (9) *holds whenever* $\partial f/\partial s$ *and* $\partial f/\partial p_i$ $(1 \le i \le N)$ *satisfy*

$$\lim_{|(s,p)|\to\infty} \frac{\left|\dfrac{\partial f}{\partial s}(s,p)\right|}{\log^{3/2}(1+|s|+|p|)} = 0, \qquad \lim_{|(s,p)|\to\infty} \frac{\left|\dfrac{\partial f}{\partial p_i}(s,p)\right|}{\log^{1/2}(1+|s|+|p|)} = 0.$$

On the other hand, the assumptions (9) *can be easily interpreted when* $f = f(s)$. *Indeed, in this case they simply read as follows:*

$$\begin{cases} \displaystyle\lim_{|s|\to\infty} \frac{1}{\log^{3/2}(1+|s|)} \left| \int_0^1 f'(s_0+\lambda s)\, d\lambda \right| = 0 \\ \text{uniformly in } s_0 \in K \text{ for every compact set } K \subset \mathbb{R}. \end{cases}$$

The arguments in [FZ2] *show that this is equivalent to* (7) *and also to*

$$\lim_{|s|\to\infty} \frac{1}{\log^{3/2}(1+|s|)} \int_0^1 f'(\lambda s)\, d\lambda = 0.$$

REMARK 1.7. *It is proved in* [FZ2] *that, for each* $\beta > 2$*, there exists a function* $f = f(s)$ *satisfying* (8) *such that the corresponding system* (1) *is not approximately controllable for all* $T > 0$*. As in the case of null controllability, for* f *satisfying* (8) *with* $3/2 \leq \beta \leq 2$*, the approximate controllability of* (1) *is an open question.*

We can establish similar results for (2) under hypotheses of the same kind for f and y_0. More precisely, let us introduce the Hilbert space

$$V = \{\, z \in H^1(\Omega) \,:\, z = 0 \text{ on } \partial\Omega \setminus \gamma \,\}. \tag{10}$$

One has the following.

THEOREM 1.3. *Let* $T > 0$*. Assume that the assumptions in Theorem* 1.1 *are satisfied. Then* (2) *is null-controllable at any time* $T > 0$*.*

THEOREM 1.4. *Let* $T > 0$*. Assume that* f *is locally Lipschitz-continuous and verifies* (9)*. Then* (2) *is approximately controllable at time* T*.*

REMARK 1.8. *In the proofs of the previous controllability results, we will construct controls satisfying the appropriate properties. These controls are smooth. In particular, they will be such that the associated solutions of* (1) *and* (2) *belong to* $C^0([0,T]; W^{1,\infty}(\Omega))$*, a space where we can ensure uniqueness.*

The rest of this paper is organized as follows. Section 2 is devoted to proving some technical lemmas we will use below. In section 3, we will prove Theorem 1.1. In section 4, we will give the proof of the approximate controllability result for system (1) (Theorem 1.2). Finally, the proofs of Theorems 1.3 and 1.4 will be sketched in section 5.

2. Some technical results. Before giving the proofs of the theorems above, we have to present some technical results.

Let us consider the linear problem

$$\begin{cases} \partial_t y - \Delta y + B \cdot \nabla y + ay = F & \text{in} \quad Q, \\ y = 0 & \text{on} \quad \Sigma, \\ y(x,0) = y_0(x) & \text{in} \quad \Omega, \end{cases} \tag{11}$$

where y_0 and F are given, $a \in L^\infty(Q)$, and $B \in L^\infty(Q)^N$. One has the following lemma, whose proof is essentially given in [LSU].

LEMMA 2.1. *Assume that* $F \in L^q(Q)$ *with* $q > N + 2$*,* $y_0 \in W^{2,p}(\Omega) \cap H_0^1(\Omega)$ *with* $p > N$*,* $a \in L^\infty(Q)$*, and* $B \in L^\infty(Q)^N$*. Then the solution* y *of* (11) *satisfies*

$$\begin{cases} y \in L^q(0,T; W^{2,\beta}(\Omega)),\ \partial_t y \in L^q(0,T; L^\beta(\Omega)), \\ \textit{with} \quad \beta = \min(p,q) > N \end{cases} \tag{12}$$

and

$$\begin{cases} \|y\|_{L^q(0,T;W^{2,\beta})} + \|\partial_t y\|_{L^q(0,T;L^\beta)} \\ \qquad \leq C(\Omega, T, \|a\|_\infty, \|B\|_\infty)\,(\|y_0\|_{W^{2,p}} + \|F\|_q)\,. \end{cases} \tag{13}$$

Furthermore, we also have $y \in C^0([0,T]; W^{1,\infty}(\Omega))$ *and*

$$\|y\|_{C^0([0,T];W^{1,\infty})} \leq M(\Omega, T, \|a\|_\infty, \|B\|_\infty)\,(\|y_0\|_{W^{2,p}} + \|F\|_q)\,, \tag{14}$$

where

$$\begin{cases} M(\Omega, T, \|a\|_\infty, \|B\|_\infty) \\ \qquad = \exp\left[M_0\left(1 + T + (T + T^{1/2})\|a\|_\infty + (T + T^{1/2})\|B\|_\infty^2\right)\right] \end{cases} \tag{15}$$

and M_0 is a positive constant depending only on Ω.

For the reader's convenience, we have sketched the proof of this result in the appendix.

We will also recall a global Carleman inequality from [IY] for the linear problem

$$\begin{cases} -\partial_t \varphi - \Delta \varphi = F_0 + \sum_{i=1}^{N} \dfrac{\partial F_i}{\partial x_i} & \text{in } Q, \\ \varphi = 0 & \text{on } \Sigma, \\ \varphi(x,T) = \varphi_T(x) & \text{in } \Omega, \end{cases} \tag{16}$$

where F_0, $F_i \in L^2(Q)$ $(1 \le i \le N)$ and $\varphi_T \in L^2(\Omega)$. One has the following.

LEMMA 2.2. *There exists a smooth function $\alpha_0 = \alpha_0(x)$ that is defined and strictly positive for $x \in \overline{\Omega}$, and there exist positive constants C_0 and σ_0 (only depending on Ω and $\mathcal{O}$) such that*

$$\begin{aligned} s^3 \iint_Q e^{-2s\alpha} t^{-3} (T-t)^{-3} |\varphi|^2 \le C_0 \Bigg(& s^3 \iint_{\mathcal{O}\times(0,T)} e^{-2s\alpha} t^{-3} (T-t)^{-3} |\varphi|^2 \\ & + \iint_Q e^{-2s\alpha} |F_0|^2 + s^2 \sum_{i=1}^{N} \iint_Q e^{-2s\alpha} t^{-2} (T-t)^{-2} |F_i|^2 \Bigg) \end{aligned} \tag{17}$$

for all $s \ge s_0 = \sigma_0(T+T^2)$, where φ is the solution of (16) *associated to $\varphi_T \in L^2(\Omega)$. In* (17)*, the function $\alpha = \alpha(x,t)$ is given by*

$$\alpha(x,t) = \frac{\alpha_0(x)}{t(T-t)}\,.$$

REMARK 2.1. *The inequality* (17) *is based on a similar Carleman inequality for the heat equation with a right-hand side in $L^2(Q)$. The precise way s_0 depends on T has been analyzed in* [FZ1] *and is essential in our analysis.*

In what follows, unless otherwise specified, C will stand for a generic positive constant depending only on Ω and $\mathcal{O}$, whose value can change from line to line. Let us introduce the following (adjoint) system:

$$\begin{cases} -\partial_t q - \Delta q - \nabla \cdot (qB) + aq = 0 & \text{in } Q, \\ q = 0 & \text{on } \Sigma, \\ q(x,T) = q_T(x) & \text{in } \Omega, \end{cases} \tag{18}$$

where $q_T \in L^2(\Omega)$. Arguing as in [FZ1], we can deduce from the Carleman estimates (17) an observability inequality for (18), as follows.

THEOREM 2.3. *For any $a \in L^\infty(Q)$, $B \in L^\infty(Q)^N$, and $q_T \in L^2(\Omega)$, one has*

$$\|q(\cdot,0)\|_{L^2}^2 \le \exp\left[C\, B(T, \|a\|_\infty, \|B\|_\infty)\right] \iint_{\mathcal{O}\times(0,T)} |q|^2, \tag{19}$$

where

$$B(T, \|a\|_\infty, \|B\|_\infty) = 1 + \frac{1}{T} + T\|a\|_\infty + \|a\|_\infty^{2/3} + (1+T)\|B\|_\infty^2$$

and q is the solution to the corresponding system (18).

Proof. Let a, B, and q_T be given and let q be the solution to (18). Let us first see that

$$\iint_{\Omega\times(T/4,3T/4)} |q|^2 \le \exp\left[C\left(1+\frac{1}{T}+\|a\|_\infty^{2/3}+\|B\|_\infty^2\right)\right]\iint_{\mathcal{O}\times(0,T)}|q|^2\,. \tag{20}$$

We can write (17) for $\varphi = q$. This gives

$$\begin{aligned} s^3\iint_Q e^{-2s\alpha}t^{-3}(T-t)^{-3}|q|^2 \le C_0\Bigg(&s^3\iint_{\mathcal{O}\times(0,T)} e^{-2s\alpha}t^{-3}(T-t)^{-3}|q|^2 \\ &+\iint_Q e^{-2s\alpha}|aq|^2+s^2\iint_Q e^{-2s\alpha}t^{-2}(T-t)^{-2}|Bq|^2\Bigg) \end{aligned} \tag{21}$$

for all $s \ge s_0$. We can estimate the terms on the right as follows:

$$\iint_Q e^{-2s\alpha}|aq|^2 \le 2^{-6}T^6\|a\|_\infty^2\iint_Q e^{-2s\alpha}t^{-3}(T-t)^{-3}|q|^2$$

and

$$\iint_Q e^{-2s\alpha}t^{-2}(T-t)^{-2}|Bq|^2 \le 2^{-2}T^2\|B\|_\infty^2\iint_Q e^{-2s\alpha}t^{-3}(T-t)^{-3}|q|^2.$$

Thus, we deduce from (21) that

$$\iint_Q e^{-2s\alpha}t^{-3}(T-t)^{-3}|q|^2 \le C\iint_{\mathcal{O}\times(0,T)} e^{-2s\alpha}t^{-3}(T-t)^{-3}|q|^2, \tag{22}$$

provided

$$s \ge s_1 = \max\left(s_0, 2^{-4/3}C_0^{1/3}T^2\|a\|_\infty^{2/3}, C_0T^2\|B\|_\infty^2\right).$$

On the other hand, it can be easily verified that

$$e^{-2s\alpha}t^{-3}(T-t)^{-3} \le 2^6T^{-6}\exp\left(-CsT^{-2}\right) \quad \forall (x,t)\in\overline{Q} \tag{23}$$

and

$$e^{-2s\alpha}t^{-3}(T-t)^{-3} \ge \left(\frac{16}{3}\right)^3 T^{-6}\exp\left(-CsT^{-2}\right) \quad \forall (x,t)\in\overline{\Omega}\times[T/4,3T/4] \tag{24}$$

whenever

$$s \ge s_2 = \max\left(s_1, 3T^2\left(8\min_{x\in\overline{\Omega}}\alpha_0(x)\right)^{-1}\right).$$

Analyzing the definitions of s_1 and s_2, we see that $s_2 \le s_3$, where s_3 is of the form

$$s_3 = \sigma_3\left(T+T^2+T^2\|a\|_\infty^{2/3}+T^2\|B\|_\infty^2\right)$$

and σ_3 depends only on Ω and $\mathcal{O}$. From now on, we fix s, with $s = s_3$. Taking into account (23) and (24) and coming back to (22) (written for $s = s_3$), we deduce that (20) is satisfied for any solution q of (18).

Let us now prove that

$$\|q(\cdot,T/4)\|_2^2 \le \exp\left[C\left(\frac{1}{T}+T\|a\|_\infty+T\|B\|_\infty^2\right)\right]\iint_{\Omega\times(T/4,3T/4)}|q|^2. \tag{25}$$

Multiplying (18) by q and integrating in Ω, we obtain

$$-\frac{1}{2}\frac{d}{dt}\int_\Omega |q|^2\,dx+\int_\Omega|\nabla q|^2\,dx=-\int_\Omega qB\cdot\nabla q\,dx-\int_\Omega a|q|^2\,dx \qquad \forall t\ge 0.$$

Thus,

$$-\frac{d}{dt}\int_\Omega |q|^2\,dx+\int_\Omega|\nabla q|^2\,dx\le\left(\|B\|_\infty^2+2\|a\|_\infty\right)\int_\Omega|q|^2\,dx$$

and

$$\frac{d}{dt}\left(\exp\left((2\|a\|_\infty+\|B\|_\infty^2)t\right)\int_\Omega|q|^2\,dx\right)\ge 0 \tag{26}$$

for all $t\ge 0$. Integrating this inequality with respect to the time variable in $[T/4,t]$, where $t\in[T/4,3T/4]$, we obtain

$$\begin{cases}\displaystyle\int_\Omega|q(x,t)|^2\,dx\ge\exp\left[\left(2\|a\|_\infty+\|B\|_\infty^2\right)(T/4-t)\right]\int_\Omega|q(x,T/4)|^2\,dx\\ \qquad\qquad\displaystyle\ge\exp\left[-\left(\|a\|_\infty+\frac{1}{2}\|B\|_\infty^2\right)T\right]\int_\Omega|q(x,T/4)|^2\,dx\end{cases} \tag{27}$$

for all $t\in[T/4,3T/4]$. Integrating (27) again with respect to t, we find that

$$\frac{T}{2}\int_\Omega|q(x,T/4)|^2\,dx\le\exp\left[\left(\|a\|_\infty+\frac{1}{2}\|B\|_\infty^2\right)T\right]\iint_{\Omega\times(T/4,3T/4)}|q(x,t)|^2, \tag{28}$$

whence we easily deduce (25).

Finally, let us prove that

$$\int_\Omega|q(x,0)|^2\,dx\le\exp\left[CT\left(\|a\|_\infty+\|B\|_\infty^2\right)\right]\int_\Omega|q(x,T/4)|^2\,dx. \tag{29}$$

This, together with (25) and (20), will lead to the desired observability estimate (19).

To prove (29), it suffices to integrate (26) in the time interval $[0,T/4]$. Indeed, we find at once that

$$\int_\Omega|q(x,0)|^2\,dx\le\exp\left[\left(2\|a\|_\infty+\|B\|_\infty^2\right)\frac{T}{4}\right]\int_\Omega|q(x,T/4)|^2\,dx$$

and thus (29) holds. This completes the proof of Theorem 2.3. □

In fact, for the analysis of the controllability of (1) and (2), where f is not necessarily globally Lipschitz-continuous, we need a refined version of the observability inequality (19). This is furnished by the following result.

THEOREM 2.4. *For any $a\in L^\infty(\Omega)$, $B\in L^\infty(\Omega)^N$, and $q_T\in L^2(\Omega)$, one has*

$$\|q(\cdot,0)\|_{L^2}^2\le\exp\left[C\,K\left(T,\|a\|_\infty,\|B\|_\infty\right)\right]\left(\iint_{\mathcal{O}\times(0,T)}|q|\right)^2, \tag{30}$$

where

$$K\,(T,\|a\|_\infty,\|B\|_\infty)=1+\frac{1}{T}+T+(T+T^{1/2})\|a\|_\infty+\|a\|_\infty^{2/3}+(1+T)\|B\|_\infty^2. \tag{31}$$

Proof. Let $\mathcal{O}'$ be a nonempty open set such that $\mathcal{O}'\subset\subset\mathcal{O}$. From Theorem 2.3 applied to $\mathcal{O}'$ and the time interval $[T/4,3T/4]$, we deduce that

$$\|q(\cdot,T/4)\|_{L^2}^2\le\exp\left[CK'(T,\|a\|_\infty,\|B\|_\infty)\right]\iint_{\mathcal{O}'\times(T/4,3T/4)}|q|^2, \tag{32}$$

where q is the solution of (18) associated to $q_T\in L^2(\Omega)$, $K'(T,\|a\|_\infty,\|B\|_\infty)$ is given by

$$K'(T,\|a\|_\infty,\|B\|_\infty)=1+\frac{1}{T}+T\|a\|_\infty+\|a\|_\infty^{2/3}+(1+T)\,\|B\|_\infty^2,$$

and C is a new positive constant depending only on $\mathcal{O}'$ (i.e., on $\mathcal{O}$) and Ω. Using (26), we obtain

$$\int_\Omega|q(x,0)|^2\,dx\le\exp\left[\frac{T}{4}\left(2\|a\|_\infty+\|B\|_\infty^2\right)\right]\int_\Omega|q(x,T/4)|^2\,dx,$$

and combining this with (32), we find that

$$\|q(\cdot,0)\|_{L^2}^2\le\exp\left[CK'(T,\|a\|_\infty,\|B\|_\infty)\right]\iint_{\mathcal{O}'\times(T/4,3T/4)}|q|^2. \tag{33}$$

At this point, we are going to use a technical result, related to the regularizing effect of the heat equation, whose proof will be given below.

LEMMA 2.5. *Let $\mathcal{O}_i$, T_i, r_i, and γ_i $(i=0,1)$ be given, with*

$$\begin{cases}\mathcal{O}'\subset\mathcal{O}_0\subset\subset\mathcal{O}_1\subset\mathcal{O},\quad 0\le T_1<T_0<T/2,\quad 1\le r_1<r_0<\infty,\\ 1\le\gamma_1<\gamma_0<\infty,\qquad \dfrac{1}{\gamma_1}-\dfrac{1}{\gamma_0}+\dfrac{N}{2}\left(\dfrac{1}{r_1}-\dfrac{1}{r_0}\right)<\dfrac{1}{2}\,.\end{cases}$$

Then

$$\begin{cases}\left(\displaystyle\int_{T_0}^{T-T_0}\left(\int_{\mathcal{O}_0}|q|^{r_0}\,dx\right)^{\gamma_0/r_0}dt\right)^{1/\gamma_0}\\ \qquad\le C\,T^\lambda H(T,T_0,T_1,\|a\|_\infty,\|B\|_\infty)\left(\displaystyle\int_{T_1}^{T-T_1}\left(\int_{\mathcal{O}_1}|q|^{r_1}\,dx\right)^{\gamma_1/r_1}dt\right)^{1/\gamma_1}\end{cases} \tag{34}$$

for all $q_T\in L^2(\Omega)$, with $C=C(\Omega,\mathcal{O}_i,r_i,\gamma_i,N)$, $\lambda=\lambda(r_i,\gamma_i,N)$, and

$$\begin{cases}H(T,T_0,T_1,\|a\|_\infty,\|B\|_\infty)\\ \qquad=1+\dfrac{T^{1/2}}{T_0-T_1}+T^{1/2}(1+\|a\|_\infty)+(1+T^{1/2})\|B\|_\infty.\end{cases} \tag{35}$$

We will now apply this lemma together with (33). To this end, let us set $r_0=\gamma_0=2$ and let us introduce the numbers γ_i and r_i, given by the equalities

$$\frac{1}{\gamma_i}=\frac{1}{r_i}=\frac{1}{2}+\frac{i}{2(N+2)},\qquad 1\le i\le N+2.$$

It is immediate that $\gamma_{N+1} > 1$, $r_{N+1} > 1$, and $\gamma_{N+2} = r_{N+2} = 1$. Now, let us set $\delta = T/4(N+2)$. Accordingly,

$$[T/4 - (N+2)\delta, 3T/4 + (N+2)\delta] = [0, T].$$

Let us also introduce a family of open sets $\mathcal{O}_i$ such that

$$\mathcal{O}' = \mathcal{O}_0 \subset\subset \mathcal{O}_1 \subset\subset \mathcal{O}_2 \subset\subset \cdots \subset\subset \mathcal{O}_{N+1} \subset\subset \mathcal{O}_{N+2} = \mathcal{O}.$$

For $0 \le i \le N+1$, we can use inequality (34) with $\mathcal{O}_0$, $\mathcal{O}_1$, T_0, T_1, r_0, r_1, γ_0, and γ_1, respectively, replaced by $\mathcal{O}_i, \mathcal{O}_{i+1}$, $T/4 - i\delta$, $T/4 - (i+1)\delta$, r_i, r_{i+1}, γ_i, and γ_{i+1}. The whole set of these inequalities gives

$$\left(\iint_{\mathcal{O}' \times (T/4, 3T/4)} |q|^2 \right)^{1/2} \le C T^{\alpha} H(T, \|a\|_\infty, \|B\|_\infty)^{\beta} \left(\iint_{\mathcal{O} \times (0,T)} |q| \right), \tag{36}$$

where $\beta = N+2$ and α is the sum of the exponents λ_i. If we now combine the inequalities (33) and (36), we obtain (30). This completes the proof of Theorem 2.4. □

Proof of Lemma 2.5. Let ρ_1 and ρ_2 be functions in $\mathcal{D}(\mathcal{O}_1)$ and $\mathcal{D}((T_1, T - T_1))$, respectively, such that

$$\rho_1 \equiv 1 \text{ in } \mathcal{O}_0, \qquad \rho_2 \equiv 1 \text{ in } (T_0, T - T_0),$$

and $0 \le \rho_1, \rho_2 \le 1$. Let us put $\rho(x,t) = \rho_1(x)\rho_2(t)$ and $u = \rho q$, where q is the solution to (18) associated to $q_T \in L^2(\Omega)$. Obviously,

$$\operatorname{supp} u \subset \mathcal{O}_1 \times (T_1, T - T_1)$$

and

$$\begin{cases} -\partial_t u - \Delta u = -a\rho q + \nabla \cdot (\rho q B) - (\partial_t \rho + \Delta \rho) q - 2\nabla \rho \cdot \nabla q - (\nabla \rho \cdot B) q & \text{in } Q, \\ u = 0 & \text{on } \Sigma, \\ u(x,T) = 0 & \text{in } \Omega. \end{cases}$$

In order to clarify the computations, let us put $\widetilde{u}(x,t) = u(x, T-t)$ for $(x,t) \in Q$. In a similar way, let us introduce the functions $\widetilde{a}$, $\widetilde{B}$, $\widetilde{\rho}$, and $\widetilde{q}$. We then have

$$\begin{cases} \partial_t \widetilde{u} - \Delta \widetilde{u} = F & \text{in } Q, \\ \widetilde{u} = 0 & \text{on } \Sigma, \\ \widetilde{u}(x,0) = 0 & \text{in } \Omega, \end{cases}$$

where F is given by

$$F = -\widetilde{a}\widetilde{\rho}\widetilde{q} + \nabla \cdot (\widetilde{\rho}\widetilde{q}\widetilde{B}) + (\partial_t \widetilde{\rho} - \Delta \widetilde{\rho})\widetilde{q} - 2\nabla \widetilde{\rho} \cdot \nabla \widetilde{q} - (\nabla \widetilde{\rho} \cdot \widetilde{B})\widetilde{q}.$$

Let us denote by $\{S(t) : t \ge 0\}$ the semigroup generated by the heat equation with Dirichlet boundary conditions. Then one has

$$\widetilde{u}(\cdot, t) = \int_0^t S(t-s) F(\cdot, s)\, ds, \tag{37}$$

where the integral can be understood, for instance, in $L^{r_0}(\Omega)$.

Thanks to the regularizing effect of the heat equation, taking L^{r_0}-norms in (37), we obtain the following for $t \in (T_1, T - T_1)$:

$$\begin{aligned}\|\widetilde{u}(\cdot,t)\|_{L^{r_0}} \le C \Bigg[& (\|B\|_\infty + 1) \int_{T_1}^t (t-s)^{-\frac{N}{2}\left(\frac{1}{r_1}-\frac{1}{r_0}\right)-\frac{1}{2}} \|\widetilde{q}(\cdot,s)\|_{L^{r_1}(\mathcal{O}_1)} ds \\ & + \left(1 + \frac{1}{T_0 - T_1} + \|a\|_\infty + \|B\|_\infty\right) \int_{T_1}^t (t-s)^{-\frac{N}{2}\left(\frac{1}{r_1}-\frac{1}{r_0}\right)} \|\widetilde{q}(\cdot,s)\|_{L^{r_1}(\mathcal{O}_1)} ds \Bigg].\end{aligned} \tag{38}$$

Here, C is a positive constant depending on $\mathcal{O}_0$ and $\mathcal{O}_1$. This gives

$$\|\widetilde{u}(\cdot,t)\|_{L^{r_0}} \le C\, H \int_{T_1}^t (t-s)^{-\frac{N}{2}\left(\frac{1}{r_1}-\frac{1}{r_0}\right)-\frac{1}{2}} \|\widetilde{q}(\cdot,s)\|_{L^{r_1}(\mathcal{O}_1)} ds \tag{39}$$

for all $t \in (T_1, T - T_1)$, where $H = H(T, T_0, T_1, \|a\|_\infty, \|B\|_\infty)$ is given by (35). Due to the assumption

$$\frac{N}{2}\left(\frac{1}{r_1} - \frac{1}{r_0}\right) + \frac{1}{\gamma_1} - \frac{1}{\gamma_0} < \frac{1}{2}$$

we can apply Young's inequality to (39) and estimate the $L^{\gamma_0}(0,T;L^{r_0}(\Omega))$-norm of $\widetilde{u}$ as follows:

$$\left(\int_{T_1}^{T-T_1} \|\widetilde{u}(\cdot,t)\|_{L^{r_0}}^{\gamma_0} dt\right)^{1/\gamma_0} \le C\, H T^\lambda \left(\int_{T_1}^{T-T_1} \|\widetilde{u}(\cdot,t)\|_{L^{r_1}(\mathcal{O}_1)}^{\gamma_1} dt\right)^{1/\gamma_1}. \tag{40}$$

Here, C is a new positive constant only depending on Ω, $\mathcal{O}_i$, r_i, and γ_i, and N and H are given by (35) and

$$\lambda = -\left[\frac{N}{2}\left(\frac{1}{r_1} - \frac{1}{r_0}\right) + \frac{1}{\gamma_1} - \frac{1}{\gamma_0}\right] + \frac{1}{2}.$$

Inequality (34) is directly obtained from (40). This completes the proof of Lemma 2.5. □

REMARK 2.2. *As an easy consequence of Theorem* 2.4 *and* (30)*, we can also deduce for each* $r \in (1,\infty)$ *an observability inequality in* $L^r(\mathcal{O} \times (0,T))$*:*

$$\|q(\cdot,0)\|_{L^2}^2 \le \exp\left[C_r\, K\left(T, \|a\|_\infty, \|B\|_\infty\right)\right] \left(\iint_{\mathcal{O}\times(0,T)} |q|^r\right)^{\frac{2}{r}} \tag{41}$$

for any $a \in L^\infty(\Omega)$, $B \in L^\infty(\Omega)^N$, *and* $q_T \in L^2(\Omega)$. *In* (41), $K\left(T, \|a\|_\infty, \|B\|_\infty\right)$ *is given by* (31) *and* C_r *only depends on* Ω, $\mathcal{O}$, *and* r.

3. Proof of the null controllability result. This section is devoted to proving Theorem 1.1. Using Theorem 2.4, we will first establish a null controllability result for a similar linear heat equation with controls in $L^\infty(\mathcal{O} \times (0,T))$. We will then apply a fixed point argument to obtain the desired result. The structure of the proof (the controllability of a similar linear system together with a fixed point argument) is rather general. It was introduced in [Z1] in the context of the boundary controllability of the semilinear wave equation. For other results proved in a similar way, see, for instance, [FPZ] and [FI].

3.1. A null controllability result for a linear problem. We will consider the linear system

$$\left\{\begin{array}{lll} \partial_t y - \Delta y + B\cdot\nabla y + ay = v1_{\mathcal{O}} & \text{in} & Q,\\ y=0 & \text{on} & \Sigma,\\ y(x,0)=y_0(x) & \text{in} & \Omega, \end{array}\right. \tag{42}$$

where $a\in L^\infty(Q)$, $B\in L^\infty(Q)^N$, and $y_0\in L^2(\Omega)$ are given. The following holds.

THEOREM 3.1. *Assume that $T>0$, $a\in L^\infty(Q)$, $B\in L^\infty(Q)^N$, and $y_0\in L^2(\Omega)$. Then there exists a control $\widehat{v}\in L^\infty(\mathcal{O}\times(0,T))$ such that the corresponding solution of* (42) *satisfies*

$$\widehat{y}(x,T)=0 \quad \textit{in } \Omega. \tag{43}$$

Furthermore, $\widehat{v}$ can be chosen in such a way that

$$\|\widehat{v}\|_{L^\infty(\mathcal{O}\times(0,T))}\le \exp\left[C\,K\left(T,\|a\|_\infty,\|B\|_\infty\right)\right]\|y_0\|_{L^2}, \tag{44}$$

where $K(T,\|a\|_\infty,\|B\|_\infty)$ is given by (31).

Proof. For every $\varepsilon>0$, let us consider the functional J_ε, with

$$J_\varepsilon(q_T)=\frac{1}{2}\left(\iint_{\mathcal{O}\times(0,T)}|q|\right)^2+\varepsilon\|q_T\|_{L^2}+\int_\Omega q(x,0)\,y_0(x)\,dx \qquad \forall q_T\in L^2(\Omega). \tag{45}$$

Here, q is the solution of (18) associated to $q_T\in L^2(\Omega)$.

It is easy to see that J_ε is a continuous and strictly convex functional in $L^2(\Omega)$. Furthermore, from (22), it is immediate to deduce the following unique continuation property for (18): *If $q=0$ in $\mathcal{O}\times(0,T)$, then $q\equiv 0$.*

Thus, arguing as in [FPZ], we also see that

$$\liminf_{\|q_T\|_{L^2}\to\infty}\frac{J_\varepsilon(q_T)}{\|q_T\|_{L^2}}\ge\varepsilon$$

and, therefore, J_ε achieves its minimum at a unique point $\widehat{q}_T^\varepsilon\in L^2(\Omega)$.

Let $\widehat{q}_\varepsilon$ be the solution of (18) associated to $\widehat{q}_T^\varepsilon$. Taking $v=\widehat{v}_\varepsilon$ in (42) with

$$\widehat{v}_\varepsilon\in(\operatorname{sgn}\widehat{q}_\varepsilon)\left(\int_{\mathcal{O}\times(0,T)}|\widehat{q}_\varepsilon|\right)1_{\mathcal{O}} \tag{46}$$

and arguing as in [FPZ], we see that the associated solution $\widehat{y}_\varepsilon$ satisfies

$$\|\widehat{y}_\varepsilon(\cdot,T)\|_{L^2}\le\varepsilon. \tag{47}$$

It is not difficult to see that

$$\|\widehat{v}_\varepsilon\|_{L^\infty(\mathcal{O}\times(0,T))}=\iint_{\mathcal{O}\times(0,T)}|\widehat{q}_\varepsilon|\le\exp\left[C\,K\left(T,\|a\|_\infty,\|B\|_\infty\right)\right]\|y_0\|_{L^2} \tag{48}$$

for all $\varepsilon>0$. Indeed, the fact that

$$\|\widehat{v}_\varepsilon\|_{L^\infty(\mathcal{O}\times(0,T))}=\iint_{\mathcal{O}\times(0,T)}|\widehat{q}_\varepsilon|$$

is implied by (46). On the other hand, since

$$J_\varepsilon(\widehat{q}_T^\varepsilon) \le J_\varepsilon(0) = 0,$$

we see from (45) that

$$\frac{1}{2}\left(\iint_{\mathcal{O}\times(0,T)} |\widehat{q}_\varepsilon|\right)^2 \le -\int_\Omega \widehat{q}_\varepsilon(x,0)y_0(x)\,dx \le \|\widehat{q}_\varepsilon(\cdot,0)\|_{L^2}\|y_0\|_{L^2}.$$

In view of (30), (48) holds.

Since $\widehat{v}_\varepsilon$ is uniformly bounded in $L^\infty(\mathcal{O}\times(0,T))$, at least for an appropriate subsequence we must have

$$\widehat{v}_\varepsilon \to \widehat{v} \quad \text{weakly-}* \text{ in } L^\infty(\mathcal{O}\times(0,T)), \tag{49}$$

where $\widehat{v} \in L^\infty(\mathcal{O}\times(0,T))$ satisfies (44). Accordingly,

$$\widehat{y}_\varepsilon(T) \to \widehat{y}(T) \quad \text{in } L^2(\Omega),$$

where $\widehat{y}$ is the solution of (42) associated to $\widehat{v}$. Since we have (47) for all $\varepsilon > 0$, (43) is satisfied. This ends the proof. □

3.2. Proof of Theorem 1.1. We are now ready to prove Theorem 1.1. First, observe that we can assume in this theorem that $y_0 \in W^{2,p}(\Omega)\cap H_0^1(\Omega)$, with $p > N$. Indeed, it suffices to set $v = 0$ for $t \in [0,\delta]$ and to work in the time interval $[\delta, T]$, looking at $y(\cdot,\delta)$ as the initial state.

As we said above, a fixed point argument will be used. For convenience, it will be assumed in a first step that g and G are continuous.

3.2.1. The case in which g and G are continuous. Let y_0 be given in $W^{2,p}(\Omega)\cap H_0^1(\Omega)$ with $p > N$. We will assume that

$$g \in C^0(\mathbb{R}\times\mathbb{R}^N), \qquad G \in C^0(\mathbb{R}\times\mathbb{R}^N)^N, \tag{50}$$

and (6) is satisfied. It is then clear that, for each $\varepsilon > 0$, there exists $C_\varepsilon > 0$ such that

$$|g(s,p)|^{2/3} + |G(s,p)|^2 \le C_\varepsilon + \varepsilon\log(1+|s|+|p|) \quad \forall (s,p) \in \mathbb{R}\times\mathbb{R}^N. \tag{51}$$

Let us set $Z = C^0([0,T];W^{1,\infty}(\Omega))$ and let $R > 0$ be a constant whose value will be determined below. We will use the truncation functions $\mathrm{T}_R : \mathbb{R} \mapsto \mathbb{R}$ and $\mathbf{T}_R : \mathbb{R}^N \mapsto \mathbb{R}^N$, given as follows:

$$\mathrm{T}_R(s) = \begin{cases} s & \text{if } |s| \le R, \\ R\,\mathrm{sgn}\,(s) & \text{otherwise} \end{cases}$$

and

$$\mathbf{T}_R(p) = (\mathrm{T}_R(p_i))_{1\le i\le N} \quad \forall p \in \mathbb{R}^N.$$

For each $z \in Z$, we will consider the corresponding linear systems

$$\begin{cases} \partial_t y - \Delta y + G\left(\mathrm{T}_R(z), \mathbf{T}_R(\nabla z)\right)\cdot\nabla y + g\left(\mathrm{T}_R(z), \mathbf{T}_R(\nabla z)\right) y = v1_{\mathcal{O}} & \text{in } Q, \\ y = 0 & \text{on } \Sigma, \\ y(x,0) = y_0(x) & \text{in } \Omega. \end{cases} \tag{52}$$

We are going to associate to z a family $U(z)$ of L^∞-controls which serve to drive the solutions to zero. Observe that (52) is of the form (42) with

$$\left\{\begin{array}{l} a = a_z = g\left(\mathrm{T}_R(z), \mathbf{T}_R(\nabla z)\right) \in L^\infty(Q), \\ B = B_z = G\left(\mathrm{T}_R(z), \mathbf{T}_R(\nabla z)\right) \in L^\infty(Q)^N. \end{array}\right. \tag{53}$$

Consequently, we can apply Theorem 3.1 to (52). In fact, we are going to apply this result in an adequate (eventually smaller) time interval $(0, T_z)$, where

$$T_z = \min\left\{T, \|g\left(\mathrm{T}_R(z), \mathbf{T}_R(\nabla z)\right)\|_\infty^{-2/3}, \|g\left(\mathrm{T}_R(z), \mathbf{T}_R(\nabla z)\right)\|_\infty^{-1/3}\right\}. \tag{54}$$

This is a key point in our proof that will lead to appropriate estimates (this idea is taken from [FZ2]).

From Theorem 3.1, we directly deduce the existence of a control $\widehat{v}_z \in L^\infty(\mathcal{O} \times (0, T_z))$ such that the solution of (52) in $\Omega \times (0, T_z)$ with $v = \widehat{v}_z$ satisfies

$$\widehat{y}_z(x, T_z) = 0 \quad \text{in } \Omega$$

and, moreover,

$$\|\widehat{v}_z\|_{L^\infty(\mathcal{O}\times(0,T_z))} \leq \exp\left[C\, K(T_z, \|a_z\|_\infty, \|B_z\|_\infty)\right] \|y_0\|_{L^2}.$$

(K is given by (31) and a_z and B_z are given by (53).)

Let $\widetilde{v}_z$ and $\widetilde{y}_z$ be the extensions by zero of $\widehat{v}_z$ and $\widehat{y}_z$ to the whole cylinder $Q = \Omega \times (0, T)$. It is clear that $\widetilde{y}_z$ is the corresponding solution of (52) associated to $\widetilde{v}_z$ and

$$\widetilde{y}_z(x, T) = 0 \quad \text{in } \Omega. \tag{55}$$

From the definition of T_z, we see that

$$\|\widetilde{v}_z\|_{L^\infty(\mathcal{O}\times(0,T))} \leq \exp\left[C\left(1 + \|a_z\|_\infty^{2/3} + \|B_z\|_\infty^2\right)\right] \|y_0\|_{L^2}, \tag{56}$$

where the positive constant C now depends on Ω, $\mathcal{O}$, and T.

On the other hand, from (50) and Lemma 2.1, we obtain that

$$\widehat{y}_z \in C^0([0, T_z]; W^{1,\infty}(\Omega))$$

and

$$\|\widehat{y}_z\|_{C^0([0,T_z];W^{1,\infty})} \leq M(\Omega, T_z, \|a_z\|_\infty, \|B_z\|_\infty)\left(\|y_0\|_{W^{2,p}} + \|\widehat{v}_z\|_{L^\infty(\mathcal{O}\times(0,T_z))}\right)$$

(M is given by (15)). Taking into account once again the definition of T_z, the estimate (56), and the definition of $\widetilde{y}_z$, we find that $\widetilde{y}_z \in Z$ and

$$\|\widetilde{y}_z\|_Z \leq \exp\left[C\left(1 + \|a_z\|_\infty^{2/3} + \|B_z\|_\infty^2\right)\right] \|y_0\|_{W^{2,p}}, \tag{57}$$

where (again) $C = C(\Omega, \mathcal{O}, T)$.

The estimates (56) and (57) can be written in the form

$$\|\widetilde{v}_z\|_{L^\infty(\mathcal{O}\times(0,T))} \leq C_1(\Omega, \mathcal{O}, T, z)\|y_0\|_{L^2} \tag{58}$$

and

$$\|\widetilde{y}_z\|_Z \le C_1(\Omega, \mathcal{O}, T, z)\|y_0\|_{W^{2,p}}, \tag{59}$$

where

$$C_1(\Omega, \mathcal{O}, T, z) = \exp\left[C\left(1 + \|a_z\|_\infty^{2/3} + \|B_z\|_\infty^2\right)\right]. \tag{60}$$

For any given $v \in L^\infty(\mathcal{O} \times (0,T))$, let $y_v \in Z$ be the solution of (52) in Q with right-hand side v. (In order to simplify the notation, we omit the dependence on z.) With this notation in mind, let us now set for each $z \in Z$

$$U(z) = \left\{v \in L^\infty(\mathcal{O} \times (0,T)) \,:\, y_v(T) = 0, \quad \|v\|_{L^\infty(\mathcal{O}\times(0,T))} \le C_1(\Omega, \mathcal{O}, T, z)\|y_0\|_{L^2}\right\}$$

and

$$\Lambda(z) = \{y_v \,:\, v \in U(z), \quad \|y_v\|_Z \le C_1(\Omega, \mathcal{O}, T, z)\|y_0\|_{W^{2,p}}\}. \tag{61}$$

In this way, we have been able to introduce a set-valued mapping on Z

$$z \mapsto \Lambda(z).$$

We will prove that this mapping possesses at least one fixed point y. We will also prove that, for some R, every fixed point of Λ verifies

$$\|y\|_Z \le R. \tag{62}$$

Of course, this will imply the existence of a control $v \in L^\infty(\mathcal{O} \times (0,T))$ such that (1) has a solution satisfying (3).

Let us see that Kakutani's fixed point theorem can be applied to Λ. (For the statement and proof of this result, see [A, Chapter 9, pp. 119–126].) First, from (58) and (59), we deduce that $\Lambda(z)$ is, for every $z \in Z$, a nonempty set. Moreover, it is easy to check that $\Lambda(z)$ is a uniformly bounded closed convex subset of Z. Owing to the regularity hypothesis on y_0 and Lemma 2.1, we have (12) (here $\beta = p$) and the estimate

$$\|y\|_{L^\infty(0,T;W^{2,p})} + \|\partial_t y\|_{L^\infty(0,T;L^p)} \le C(\Omega, \mathcal{O}, T, R, \|y_0\|_{W^{2,p}})$$

(where $C(\Omega, \mathcal{O}, T, R, \|y_0\|_{W^{2,p}})$ is independent of z) for any $y \in \Lambda(z)$. Since $p > N$, we can apply well-known compactness results and conclude that there exists a compact set $K \subset Z$ (which depends on R) such that

$$\Lambda(z) \subset K \quad \forall z \in Z \tag{63}$$

(for instance, see [S]).

Let us now prove that the mapping $z \mapsto \Lambda(z)$ is *upper hemicontinuous*, i.e., that the real-valued function

$$z \in Z \mapsto \sup_{y \in \Lambda(z)} \langle \mu, y \rangle$$

is *upper semicontinuous* for each bounded linear form $\mu \in Z'$. In other words, let us see that

$$B_{\alpha,\mu} = \left\{ z \in Z \,:\, \sup_{y \in \Lambda(z)} \langle \mu, y \rangle \ge \alpha \right\}$$

is a closed set of Z for every $\alpha \in \mathbb{R}$ and every $\mu \in Z'$. Thus, let $\{z_n\}$ be a sequence in $B_{\alpha,\mu}$ such that $z_n \to z$ in Z. Our aim is to prove that $z \in B_{\alpha,\mu}$. In view of the continuity hypothesis on g and G, we have

$$g(\mathrm{T}_R(z_n), \mathbf{T}_R(\nabla z_n)) \to g(\mathrm{T}_R(z), \mathbf{T}_R(\nabla z)) \quad \text{in} \quad L^\infty(Q)$$

and

$$G(\mathrm{T}_R(z_n), \mathbf{T}_R(\nabla z_n)) \to G(\mathrm{T}_R(z), \mathbf{T}_R(\nabla z)) \quad \text{in} \quad L^\infty(Q)^N.$$

Since all sets $\Lambda(z_n)$ are compact and satisfy (63), we deduce that

$$\alpha \leq \sup_{y \in \Lambda(z_n)} \langle \mu, y \rangle = \langle \mu, y_n \rangle \tag{64}$$

for some $y_n \in \Lambda(z_n)$. From the definitions of $\Lambda(z_n)$ and $U(z_n)$, there must exist $v_n \in L^\infty(\mathcal{O} \times (0,T))$ such that

$$\partial_t y_n - \Delta y_n + G\left(\mathrm{T}_R(z_n), \mathbf{T}_R(\nabla z_n)\right) \cdot \nabla y_n + g\left(\mathrm{T}_R(z_n), \mathbf{T}_R(\nabla z_n)\right) y_n = v_n 1_{\mathcal{O}}$$

in Q. Furthermore,

$$\|v_n\|_{L^\infty(\mathcal{O}\times(0,T))} \leq C_1(\Omega, \mathcal{O}, T, z_n)\|y_0\|_{L^2}$$

and

$$\|y_n\|_Z \leq C_1(\Omega, \mathcal{O}, T, z_n)\|y_0\|_{W^{2,p}},$$

whence y_n (resp., v_n) is uniformly bounded in Z (resp., $L^\infty(\mathcal{O} \times (0,T))$). Therefore, we can write the following at least for a subsequence:

$$y_n \to \widehat{y} \quad \text{strongly in } Z$$

(recall that (63) is satisfied) and

$$v_n \to \widehat{v} \quad \text{weakly-}* \text{ in } L^\infty(\mathcal{O} \times (0,T)).$$

Now, it is not difficult to check that

$$\begin{cases} \partial_t \widehat{y} - \Delta \widehat{y} + G\left(\mathrm{T}_R(z), \mathbf{T}_R(\nabla z)\right) \cdot \nabla \widehat{y} + g\left(\mathrm{T}_R(z), \mathbf{T}_R(\nabla z)\right) \widehat{y} = \widehat{v} 1_{\mathcal{O}} & \text{in} \quad Q, \\ \widehat{y} = 0 & \text{on} \quad \Sigma, \\ \widehat{y}(x,0) = y_0(x), \quad \widehat{y}(x,T) = 0 & \text{in} \quad \Omega, \end{cases}$$

i.e., that $\widehat{v} \in U(z)$ and $\widehat{y} \in \Lambda(z)$. Consequently, we can take limits in (64) and deduce that

$$\alpha \leq \langle \mu, \widehat{y} \rangle \leq \sup_{y \in \Lambda(z)} \langle \mu, y \rangle,$$

that is to say, $z \in B_{\alpha,\mu}$. This proves that $z \mapsto \Lambda(z)$ is upper hemicontinuous.

As a consequence, for any fixed $R > 0$ Kakutani's theorem can be applied, ensuring the existence of a fixed point of Λ. As we said above, we will finish the proof by showing that we can choose $R > 0$ in such a way that any fixed point of Λ satisfies (62). It is just here where the assumptions (6) (in fact (51)) will be used.

Thus, let y be a fixed point of Λ associated to the control $v \in U(y)$. Then (59), (60), and (51) lead to the estimates

$$\begin{aligned}\|y\|_Z &\le \exp\left(C\left(1+\|g(\mathrm{T}_R(y),\mathbf{T}_R(\nabla y))\|_\infty^{2/3}+\|G(\mathrm{T}_R(y),\mathbf{T}_R(\nabla y))\|_\infty^2\right)\right)\|y_0\|_{W^{2,p}}\\ &\le \exp\left(C\left(1+C_\varepsilon+\varepsilon\log\left(1+2R\right)\right)\right)\|y_0\|_{W^{2,p}}\\ &= \exp\left(C\left(1+C_\varepsilon\right)\right)(1+2R)^{C\varepsilon}\|y_0\|_{W^{2,p}},\end{aligned}$$

where $C = C(\Omega, \mathcal{O}, T)$. Taking $\varepsilon = 1/(2C)$, we find that

$$\|y\|_Z \le C(1+2R)^{1/2}\|y_0\|_{W^{2,p}},$$

whence (62) holds whenever R is large enough (depending on Ω, $\mathcal{O}$, T, g, and G). We have then proved Theorem 1.1 in the case of smooth data.

3.2.2. The general case. Let us now suppose that f is a locally Lipschitz-continuous function satisfying assumption (5) (with $f(0,0) = 0$) and (6). Let us introduce a function $\rho \in \mathcal{D}(\mathbb{R}\times\mathbb{R}^N)$ such that $\rho \ge 0$ in $\mathbb{R}\times\mathbb{R}^N$, $\operatorname{supp}\rho \subset \overline{B}(0,1)$, and

$$\iint_{\mathbb{R}\times\mathbb{R}^N} \rho(s,p)\,ds\,dp = 1.$$

We consider the functions ρ_n, g_n, and G_n $(n \ge 1)$, with

$$\rho_n(s,p) = n^{N+1}\rho(ns,np) \quad \forall (s,p)\in\mathbb{R}\times\mathbb{R}^N,$$

$$g_n = \rho_n * g, \qquad G_n = \rho_n * G.$$

Then it is not difficult to check that the following properties of g_n and G_n hold:

1. $g_n \in C^0(\mathbb{R}\times\mathbb{R}^N)$ and $G_n \in C^0(\mathbb{R}\times\mathbb{R}^N)^N$ for all $n \ge 1$.
2. If we put $f_n(s,p) = g_n(s,p)s + G_n(s,p)\cdot p$ for all $(s,p)\in\mathbb{R}\times\mathbb{R}^N$, then
$$f_n \to f \quad \text{uniformly in the compact sets of } \mathbb{R}\times\mathbb{R}^N.$$
3. For any given $M > 0$, there exists $C(M) > 0$ such that
$$\sup_{|(s,p)|\le M}\left(|g_n(s,p)| + |G_n(s,p)|\right) \le C(M) \quad \forall n \ge 1.$$
4. The functions g_n and G_n verify (6) uniformly in n, that is to say, for any $\varepsilon > 0$, there exists $M(\varepsilon) > 0$ such that
$$\text{(65)} \qquad \begin{cases} |g_n(s,p)| \le \varepsilon\log^{3/2}(1+|s|+|p|),\\ |G_n(s,p)| \le \varepsilon\log^{1/2}(1+|s|+|p|)\end{cases}$$
whenever $|(s,p)| \ge M(\varepsilon)$ for all $n \ge 1$.

For every n, we can argue as in section 3.2.1 and find a control $v_n \in L^\infty(\mathcal{O}\times(0,T))$ such that the system

$$\text{(66)} \qquad \begin{cases} \partial_t y_n - \Delta y_n + f_n(y_n,\nabla y_n) = v_n 1_{\mathcal{O}} & \text{in } Q,\\ y_n = 0 & \text{on } \Sigma,\\ y_n(x,0) = y_0(x) & \text{in } \Omega\end{cases}$$

possesses at least one solution $y_n \in Z$ satisfying

$$y_n(x,T) = 0 \quad \text{in } \Omega.$$

From the properties satisfied by g_n and G_n, and thanks to the estimates obtained in section 3.2.1, we deduce that

$$\|v_n\|_{L^\infty(\mathcal{O}\times(0,T))} \le C \quad \text{and} \quad \|y_n\|_Z \le C$$

for all $n \ge 1$. In fact, in view of Lemma 2.1 we have $y_n \in K$ for all n, where K is a fixed compact set in Z. Accordingly, we can assume that, at least for a subsequence,

$$v_n \to v \quad \text{weakly-}* \text{ in } L^\infty(\mathcal{O}\times(0,T))$$

and

$$y_n \to y \quad \text{strongly in } Z.$$

Hence, passing to the limit in (66), we find a control $v \in L^\infty(\mathcal{O}\times(0,T))$ such that (1) possesses a solution y satisfying (3). This ends the proof of Theorem 1.1. ∎

REMARK 3.1. *Analyzing the proof of Theorem* 1.1*, we deduce that the null controllability result remains valid if we change* (6) *by the following assumptions:*

$$\limsup_{|(s,p)|\to\infty} \frac{|g(s,p)|}{\log^{3/2}(1+|s|+|p|)} \le l_1 < \infty, \qquad \limsup_{|(s,p)|\to\infty} \frac{|G(s,p)|}{\log^{1/2}(1+|s|+|p|)} \le l_2 < \infty,$$

where l_1 and l_2 are positive and sufficiently small (depending only on Ω and $\mathcal{O}$).

REMARK 3.2. *In Theorem* 1.1*, we can consider as well a more general nonlinear term of the form $f(x,t;s,p)$, with $(x,t) \in Q$ and $(s,p) \in \mathbb{R}\times\mathbb{R}^N$. The assumptions on f have to be the following in this case:*

1. *$f(x,t;0,0) = 0$ for all $(x,t) \in Q$,*
2. *$f(\cdot;s,p) \in L^\infty(Q)$ for all $(s,p) \in \mathbb{R}\times\mathbb{R}^N$,*
3. *$f(x,t;\cdot)$ is locally Lipschitz-continuous for (x,t) a.e. in Q, with Lipschitz constants independent of (x,t) in the bounded sets of $\mathbb{R}\times\mathbb{R}^N$,*
4. *$f(\cdot;s,p) = g(\cdot;s,p)s + G(\cdot;s,p)\cdot p$ for all $(s,p) \in \mathbb{R}\times\mathbb{R}^N$, with*

$$\lim_{|(s,p)|\to\infty} \frac{|g(x,t;s,p)|}{\log^{3/2}(1+|s|+|p|)} = 0, \qquad \lim_{|(s,p)|\to\infty} \frac{|G(x,t;s,p)|}{\log^{1/2}(1+|s|+|p|)} = 0$$

uniformly in $(x,t) \in Q$.

REMARK 3.3. *Adapting the arguments used in the proof of Theorem* 1.1*, we can deduce a local null controllability result for* (1) *with a general nonlinear term $f(s,p)$ satisfying $f(0,0) = 0$. To be precise, if f is given, there exists $\delta = \delta(\Omega,\mathcal{O},T,f) > 0$ such that for every $y_0 \in W^{2,p}(\Omega)\cap H_0^1(\Omega)$ $(p > N)$ with $\|y_0\|_{W^{2,p}} \le \delta$, a control $v \in L^\infty(\mathcal{O}\times(0,T))$ can be found such that the corresponding problem has a unique solution $y \in L^\infty(0,T;W^{1,\infty}(\Omega))$ which satisfies*

$$y(x,T) = 0 \quad \text{in } \Omega.$$

4. Proof of the approximate controllability result. In this section we will prove Theorem 1.2. Let us fix $T > 0$, $\varepsilon > 0$, $y_0 \in W^{1,\infty}(\Omega) \cap H_0^1(\Omega)$, and $y_d \in W^{2,p}(\Omega) \cap H_0^1(\Omega)$ with $p > N$ (for instance). Obviously, it will be sufficient to consider final data in $W^{2,p}(\Omega) \cap H_0^1(\Omega)$, since this space is dense in $L^2(\Omega)$. We will present the proof in several steps and start with a result concerning the exact controllability to the trajectories in $C^0([0,T]; W^{1,\infty}(\Omega))$.

LEMMA 4.1. *Assume the hypotheses on f in Theorem 1.2 are satisfied. Let $y_0 \in W^{1,\infty}(\Omega) \cap H_0^1(\Omega)$ be given and let y^* be a solution to (1) in $C^0([0,T]; W^{1,\infty}(\Omega))$ corresponding to the data*

$$y_0^* \in W^{1,\infty}(\Omega) \cap H_0^1(\Omega), \quad v^* \in L^\infty(\mathcal{O} \times (0,T)).$$

There exists a control $v \in L^\infty(\mathcal{O} \times (0,T))$ and a state $y \in C^0([0,T]; W^{1,\infty}(\Omega))$ associated to y_0 and v such that

$$y(x,T) = y^*(x,T) \quad \textit{in} \quad \Omega.$$

Proof. Let us put $y = y^* + w$. We will look for a control $u \in L^\infty(\mathcal{O} \times (0,T))$ such that the solution of

$$\text{(67)} \qquad \begin{cases} \partial_t w - \Delta w + F(x,t;w,\nabla w) = u 1_{\mathcal{O}} & \text{in} \quad Q, \\ w = 0 & \text{on} \quad \Sigma, \\ w(0) = y_0 - y_0^* & \text{in} \quad \Omega \end{cases}$$

satisfies

$$w(x,T) = 0 \quad \text{in} \quad \Omega.$$

Here, F is given by

$$F(x,t;s,p) = f(y^*(x,t) + s, \nabla y^*(x,t) + p) - f(y^*(x,t), \nabla y^*(x,t))$$

for all $(x,t) \in Q$ and $(s,p) \in \mathbb{R} \times \mathbb{R}^N$. The proof of this lemma will be achieved if we check that such a control u exists.

Notice that

$$F(x,t;s,p) = \widetilde{g}(x,t;s,p)s + \widetilde{G}(x,t;s,p) \cdot p,$$

where

$$\widetilde{g}(x,t;s,p) = \int_0^1 \frac{\partial f}{\partial s}(y^*(x,t) + \lambda s, \nabla y^*(x,t) + \lambda p)\, d\lambda$$

and

$$\widetilde{G}_i(x,t;s,p) = \int_0^1 \frac{\partial f}{\partial p_i}(y^*(x,t) + \lambda s, \nabla y^*(x,t) + \lambda p)\, d\lambda \qquad \text{for } 1 \le i \le N.$$

Thus, in view of (9) and the fact that $y^* \in C^0([0,T]; W^{1,\infty}(\Omega))$, it is clear that F satisfies the assumptions of Remark 3.2. This is sufficient to ensure that u exists. This completes the proof of this lemma. ∎

Now, we argue as follows:

• There exists $\delta_0 > 0$, depending only on Ω, y_d, and f, such that the system

$$\left\{\begin{array}{lll} \partial_t w - \Delta w + f(w, \nabla w) = 0 & \text{in} & \Omega \times (0, \delta_0), \\ w = 0 & \text{on} & \partial\Omega \times (0, \delta_0), \\ w(x, 0) = y_d(x) & \text{in} & \Omega \end{array}\right. \tag{68}$$

has exactly one solution $w \in C^0([0, \delta_0]; W^{1,\infty}(\Omega))$ also satisfying

$$w(\cdot, t) \in W^{2,p}(\Omega) \cap H_0^1(\Omega) \qquad \forall t \in [0, \delta_0]. \tag{69}$$

Obviously, we can associate to ε a parameter $\delta_1 \in (0, \delta_0]$ (small enough) such that

$$\|w(\cdot, t) - y_d\|_{L^2} \leq \varepsilon \qquad \forall t \in [0, \delta_1]. \tag{70}$$

In what follows, we fix δ_1 verifying (70).

• There exists $v_1 \in L^\infty(\mathcal{O} \times (0, \delta_1))$ such that the corresponding system

$$\left\{\begin{array}{lll} \partial_t y - \Delta y + f(y, \nabla y) = v_1 1_{\mathcal{O}} & \text{in} & \Omega \times (0, \delta_1), \\ y = 0 & \text{on} & \partial\Omega \times (0, \delta_1), \\ y(x, 0) = y_0(x) & \text{in} & \Omega \end{array}\right. \tag{71}$$

possesses exactly one solution $y_1 \in C^0([0, \delta_1]; W^{1,\infty}(\Omega))$, with

$$y_1(x, \delta_1) = w(x, \delta_1) \quad \text{in} \quad \Omega.$$

This is a consequence of Lemma 4.1.

• On the other hand, there exists $\widetilde{v} \in L^\infty(\mathcal{O} \times (0, \delta_1))$ such that the system

$$\left\{\begin{array}{lll} \partial_t y - \Delta y + f(y, \nabla y) = \widetilde{v} 1_{\mathcal{O}} & \text{in} & \Omega \times (0, \delta_1), \\ y = 0 & \text{on} & \partial\Omega \times (0, \delta_1), \\ y(x, 0) = w(x, \delta_1) & \text{in} & \Omega \end{array}\right. \tag{72}$$

possesses exactly one solution $\widetilde{y} \in C^0([0, \delta_1]; W^{1,\infty}(\Omega))$, with

$$\widetilde{y}(x, \delta_1) = w(x, \delta_1) \quad \text{in} \quad \Omega.$$

This is again a consequence of Lemma 4.1.

• Assume that $T = n\delta_1 + \delta$ for some integer $n \geq 0$ and some $\delta \in [0, \delta_1)$. Let us put $I_k = [k\delta_1, (k+1)\delta_1)$ for $0 \leq k \leq n-1$ and $I_n = [n\delta_1, T]$. We will construct the control v as follows.

For $t \in I_0$, we set $v(x, t) = v_1(x, t)$ a.e., where v_1 is the control arising in (71). Then, for $1 \leq k \leq n-1$ and $t \in I_k$, we set $v(x, t) = \widetilde{v}(x, t - k\delta_1)$, where $\widetilde{v}$ is the control in (72).

If $\delta = 0$, we have constructed in this way a control $v \in L^\infty(\mathcal{O} \times (0, T))$ such that the associate state y satisfies

$$y(x, T) = w(x, \delta_1) \quad \text{in} \quad \Omega. \tag{73}$$

In view of (70), (4) is satisfied.

If $\delta \in (0, \delta_1)$, then we complete the definition of v by setting $v(x, t) = \widehat{v}(x, t - n\delta_1)$ for all $t \in I_n$. Here, $\widehat{v} \in L^\infty(\mathcal{O} \times (0, \delta))$ is a control such that the system

$$\left\{\begin{array}{lll} \partial_t y - \Delta y + f(y, \nabla y) = \widehat{v} 1_{\mathcal{O}} & \text{in} & \Omega \times (0, \delta), \\ y = 0 & \text{on} & \partial\Omega \times (0, \delta), \\ y(x, 0) = w(x, \delta_1) & \text{in} & \Omega \end{array}\right. \tag{74}$$

possesses exactly one solution $\hat{y} \in C^0([0,\delta]; W^{1,\infty}(\Omega))$ satisfying

$$\hat{y}(x,\delta) = w(x,\delta) \quad \text{in} \quad \Omega.$$

(Once more, the existence of $\hat{v}$ is implied by Lemma 4.1.) Now, the state y associated to y_0 and v satisfies

$$y(x,T) = w(x,\delta) \quad \text{in} \quad \Omega. \tag{75}$$

Again, taking (70) into account, we see that (4) is satisfied in this case.

This completes the proof of Theorem 1.2.

5. Sketch of the proofs of the boundary controllability results. We devote this section to sketching briefly the proofs of Theorems 1.3 and 1.4. Both results are implied by the results established in the case of internal controllability.

For instance, let us refer to the proof of Theorem 1.3. Let us assume, for simplicity, that $y_0 \in W^{2,p}(\Omega) \cap V$ for some $p > N$ (recall that V is given by (10)). We have assumed that $f : \mathbb{R} \times \mathbb{R}^N \mapsto \mathbb{R}$ is a locally Lipschitz-continuous function that satisfies $f(0,0) = 0$ and (6). Let D be a bounded open set with boundary ∂D of class C^2 such that $\Omega \subset D$ and $\partial\Omega \cap D = \gamma$. Let $\mathcal{O}$ be an open subset of $D \setminus \overline{\Omega}$. There exists a function $\widetilde{y}_0 \in W^{2,p}(D) \cap H_0^1(D)$ such that $\widetilde{y}_0 = y_0$ in Ω and

$$\|\widetilde{y}_0\|_{W^{2,p}(D)} \leq C \, \|y_0\|_{W^{2,p}(\Omega)},$$

where C is a positive constant depending only on Ω and D.

Let $\widetilde{v} \in L^\infty(\mathcal{O} \times (0,T))$ be a control, furnished by Theorem 1.1, such that

$$\begin{cases} \partial_t \widetilde{y} - \Delta \widetilde{y} + f(\widetilde{y}, \nabla \widetilde{y}) = \widetilde{v} 1_{\mathcal{O}} & \text{in} \quad D \times (0,T), \\ \widetilde{y} = 0 & \text{on} \quad \partial D \times (0,T), \\ \widetilde{y}(x,0) = \widetilde{y}_0(x) & \text{in} \quad D \end{cases}$$

possesses exactly one solution $\widetilde{y} \in C^0([0,T]; W^{1,\infty}(D))$ with

$$\widetilde{y}(x,T) = 0 \qquad \text{in } D.$$

Let v be the trace of $\widetilde{y}$ on $\gamma \times (0,T)$. Then $v \in L^\infty(\gamma \times (0,T))$, and the restriction to $\Omega \times (0,T)$ of $\widetilde{y}$ solves the corresponding system (2). This proves Theorem 1.3.

In order to prove Theorem 1.4, it suffices to argue in a similar way.

Appendix. Proof of Lemma 2.1. The statement (12) and the inequality (13) are proved in [LSU, Theorem 9.1, p. 342]. The inequality (14) is not explicitly proved in [LSU], but it can be deduced (in several ways) from other results of this book. One of the arguments is as follows.

From Theorem 16.3 in [LSU] (p. 412), we deduce the inequalities

$$\|S(t)\varphi\|_{L^\gamma} \leq C t^{-\frac{N}{2}(\frac{1}{r}-\frac{1}{\gamma})} \|\varphi\|_{L^r}$$

and

$$\|S(t)\varphi\|_{W^{1,\gamma}} \leq C t^{-\frac{N}{2}(\frac{1}{r}-\frac{1}{\gamma})-\frac{1}{2}} \|\varphi\|_{L^r},$$

which hold for all $t > 0$ and $r, \gamma \in [1,\infty]$ with $\gamma \geq r$. Here, $S(t)$ is the semigroup generated by the heat equation with Dirichlet boundary conditions. With these inequalities in mind, one can prove the following.

LEMMA A.1. *Let y be the solution of (11). If $y \in L^\infty(0,T;W^{1,r}(\Omega))$ with $r \le 2N$, then*

$$y \in L^\infty(0,T;W^{1,\gamma}(\Omega)), \quad \text{where} \quad \gamma = \begin{cases} \left(\dfrac{1}{r} - \dfrac{1}{2N}\right)^{-1} & \text{if } r < 2N, \\ \infty & \text{if } r = 2N \end{cases} \tag{76}$$

and

$$\begin{cases} \|y\|_{L^\infty(0,T;W^{1,\gamma})} \le e^{C(1+T)} \left(\|y_0\|_{W^{2,p}} + \|F\|_q\right) \\ \qquad\qquad + e^{C(1+T^{1/2}\|B\|_\infty^2 + T^{1/2}\|a\|_\infty)} \|y\|_{L^\infty(0,T;W^{1,r})}. \end{cases}$$

Proof. The solution y of (11) can be written in the form $y(t) = z(t) - w(t)$, with

$$z(t) = S(t)y_0 + \int_0^t S(t-s)F(s)\,ds$$

and

$$w(t) = \int_0^t S(t-s)\left[ay + B\cdot\nabla y\right](s)\,ds.$$

Since $y_0 \in W^{2,p}(\Omega)$ with $p > N$ and $F \in L^q(Q)$ with $q > N+2$, it is not difficult to see that $z \in L^\infty(0,T;W^{1,\infty}(\Omega))$ and

$$\|z\|_{L^\infty(0,T;W^{1,\infty})} \le C\|y_0\|_{W^{2,p}} + \frac{C}{\alpha(q)} T^{\alpha(q)} \|F\|_q \le e^{C_q(1+T)} \left(\|y_0\|_{W^{2,p}} + \|F\|_q\right),$$

where

$$\alpha(q) = \frac{q-(N+2)}{2(q-1)}.$$

On the other hand, the usual Sobolev imbeddings give $y \in L^\infty(0,T;L^\gamma(\Omega))$ (γ is given in (76)). Moreover, we can write the following for all $t > 0$:

$$\begin{cases} \|w(\cdot,t)\|_{W^{1,\gamma}} \le C \displaystyle\int_0^t (t-s)^{-1/2} \|(ay)(\cdot,s)\|_{L^\gamma}\,ds \\ \qquad\qquad + C \displaystyle\int_0^t (t-s)^{-1/4} \|(B\cdot\nabla y)(\cdot,s)\|_{L^r}\,ds. \end{cases}$$

We can now apply Young's inequality to obtain

$$\|w\|_{L^\infty(W^{1,\gamma})} \le C\left(T^{1/2}\|a\|_\infty + T^{1/4}\|B\|_\infty\right) \|y\|_{L^\infty(0,T;W^{1,r})}.$$

This completes the proof of Lemma A.1. ∎

We are now ready to prove (14). Since $y_0 \in H_0^1(\Omega)$ and $F \in L^2(Q)$, the classical energy estimates give

$$y \in L^\infty(0,T;H_0^1(\Omega)) \cap L^2(0,T;H^2(\Omega)),$$

with

$$\|y\|_{L^\infty(0,T;H_0^1)} + \|y\|_{L^2(0,T;H^2)} \le e^{C(1+T+(T+T^{1/2})\|a\|_\infty + T\|B\|_\infty^2)} \left(\|y_0\|_{W^{2,p}} + \|F\|_q\right).$$

We can now apply Lemma A.1 with $r = 2$ and obtain $y \in L^\infty(0,T;W^{1,r_1}(\Omega))$, where

$$\frac{1}{r_1} = \frac{1}{2} - \frac{1}{2N}$$

and

$$\begin{cases} \|y\|_{L^\infty(0,T;W^{1,r_1})} \le e^{C(1+T)} \left(\|y_0\|_{W^{2,p}} + \|F\|_q\right) \\ \qquad\qquad + e^{C(1+T^{1/2}\|B\|_\infty^2 + T^{1/2}\|a\|_\infty)} \|y\|_{L^\infty(0,T;H_0^1)}. \end{cases}$$

Combining the last two inequalities, we obtain

$$\|y\|_{L^\infty(0,T;W^{1,r_1})} \le e^{C(1+T+(T+T^{1/2})\|a\|_\infty + (T+T^{1/2})\|B\|_\infty^2)} \left(\|y_0\|_{W^{2,p}} + \|F\|_q\right).$$

We can repeat this process for $i = 2, \dots, N$, with

$$r_i = \left(\frac{1}{2} - \frac{i}{2N}\right)^{-1} \quad \text{for } i \le N-1 \quad \text{and} \quad r_N = \infty.$$

Obviously, this leads to (14).

Acknowledgments. The authors thank the referees for their interesting comments and suggestions. In particular, they are indebted to referee no. 1 for having suggested the argument used in the proof of Theorem 1.2.

REFERENCES

[AB] S. Anita and V. Barbu, *Null controllability of nonlinear convective heat equation*, ESAIM Control Optim. Calc. Var., 5 (2000), pp. 157–173.

[A] J.P. Aubin, *L'Analyse non Linéaire et ses Motivations Économiques*, Masson, Paris, 1984.

[B] V. Barbu, *Exact controllability of the superlinear heat equation*, Appl. Math. Optim., 42 (2000), pp. 73–89.

[CH] Th. Cazenave and A. Haraux, *Équations d'évolution avec non-linéarité logarithmique*, Ann. Fac. Sci. Toulouse, 2 (1980), pp. 21–51.

[FPZ] C. Fabre, J.P. Puel, and E. Zuazua, *Approximate controllability of the semilinear heat equation*, Proc. Roy. Soc. Edinburgh Sect. A, 125 (1995), pp. 31–61.

[F] E. Fernández-Cara, *Null controllability of the semilinear heat equation*, ESAIM Control Optim. Calc. Var., 2 (1997), pp. 87–107.

[FZ1] E. Fernández-Cara and E. Zuazua, *The cost of approximate controllability for heat equations: The linear case*, Adv. Differential Equations, 5 (2000), pp. 465–514.

[FZ2] E. Fernández-Cara and E. Zuazua, *Null and approximate controllability for weakly blowing up semilinear heat equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 17 (2000), pp. 583–616.

[FI] A. Fursikov and O. Yu. Imanuvilov, *Controllability of Evolution Equations*, Lecture Notes Ser. 34, Seoul National University, Korea, 1996.

[I] O. Yu. Imanuvilov, *Controllability of parabolic equations*, Mat. Sb., 186 (1995), pp. 102–132.

[IY] O. Yu. Imanuvilov and M. Yamamoto, *Carleman estimate for a parabolic equation in a Sobolev space of negative order and its applications*, in Control of Nonlinear Distributed Parameter Systems, Lecture Notes in Pure and Appl. Math. 218, Dekker, New York, 2001, pp. 113–137.

[LSU] O.A. Ladyzenskaya, V.A. Solonnikov, and N.N. Uraltzeva, *Linear and Quasilinear Equations of Parabolic Type*, Nauka, Moscow, 1967.

[S] J. Simon, *Compact sets in the spaces $L^p(0,T;B)$*, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65–96.

[Z1] E. Zuazua, *Exact boundary controllability for the semilinear wave equation*, in Nonlinear Partial Differential Equations and Their Applications, Vol. 10, H. Brezis and J.L. Lions, eds., Longman Scientific and Technical, Harlow, 1991, pp. 357–391.

[Z2] E. Zuazua, *Approximate controllability for semilinear heat equations with globally Lipschitz nonlinearities*, Control Cybernet., 28 (1999), pp. 665–683.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 3, pp. 820–825

SOLUTION OF A FUNCTIONAL EQUATION ARISING IN CONTINUOUS GAMES: A DYNAMIC PROGRAMMING APPROACH*

K. D. SENAPATI† AND G. PANDA‡

Abstract. This paper deals with a functional equation in a zero sum continuous game. The existence of the solution for this equation is proved through a dynamic programming approach. Also, the authors have tried to prove the uniqueness and effectiveness of the solution.

Key words. continuous game, dynamic programming, fixed point

AMS subject classifications. 90C39, 91A20

PII. S0363012999357240

1. Introduction. In this paper, we have considered a multistage allocation process where some decisions are made to maximize and some to minimize. This type of situation occurs in the theory of games. Consider a two-person zero sum game in which both players A and B choose from a continuous domain. Suppose the real numbers x and y are treated as resource or state variables of A and B, respectively; A allocates a certain quantity u of his resource x, $0 \le u \le x$, and B allocates a certain quantity v of his resource y, $0 \le v \le y$. u and v are treated as decision variables. S is the state space, and D is the decision space.

As a result of the above allocations, A receives a payoff $R(x, y, u, v)$, and B receives a payoff $-R(x, y; u, v)$, where $R : S \times D \to R$ and R is the set of real numbers. In addition to these payoffs, there is an alternation to their resources. x becomes $T(x, y, u, v)$ and y becomes $T'(x, y; u, v)$, where T, $T' : S \times D \to R$. We assume that the R, T, and T' are continuous functions of x, y, u, and v. Bhakta and Mitra [2] proved some existence theorems in game theory through a dynamic programming approach. The functional equation (by Bellman [1, Chapter x]) we shall consider is

$$f(x,y) \;:\; \operatorname*{Max}_{G} \operatorname*{Min}_{G} \left\lfloor \int_u \int_v [R(x,y;u,v) + h(x,y;u,v) f(T,T')] dG(u) dG'(v) \right\rfloor$$

$$= \operatorname*{Max}_{G} \operatorname*{Min}_{G} [- - - - - -], \tag{1}$$

where $T = T(x, y; u, v)$ and $T' = T'(x, y; u, v)$. To simplify our notation, we write the above equation as

$$f(x,y) \;:\; \operatorname*{Max}_{G} \operatorname*{Min}_{G} F(x,y;f;G,G')$$

$$= \operatorname*{Max}_{G} \operatorname*{Min}_{G} [- - - - - -],$$

where $F(x, y; u, v) = \int_u \int_v [R(x, y; u, v) + h(x, y; u, v) f(T, T')] dG(u) dG'(v)$ and G and G' are the distribution functions for A and B, respectively. To prove the existence of the solution of (1), we use the following preliminary lemmas.

*Received by the editors June 3, 1999; accepted for publication (in revised form) February 19, 2002; published electronically August 8, 2002.
http://www.siam.org/journals/sicon/41-3/35724.html
†K.S.U.B. College, Bhanganar (Ganjam), Orissa, India (kdsenapati@yahoo.com).
‡Mathematics Group (FD-III), Birla Institute of Technology and Science, Pilani, Rajasthan, India (Geetapa@yahoo.com).

2. Some preliminary lemmas.

LEMMA 2.1. *If, for* $i = 1, 2$,

$$\Psi_i(x,y) \;:\; \operatorname*{Max}_G \operatorname*{Min}_G \left[\int_u \int_v [R(x,y;u,v) + h(x,y;u,v) f(T,T')] dG(u) dG'(v)\right]$$
$$= \operatorname*{Max}_G \operatorname*{Min}_G [- - - - - -],$$

then $|\Psi_1(x,y) - \Psi_2(x,y)| \le \mathrm{Max}_u \mathrm{Max}_v [|(x,y;u,v)|\,|f_1(T,T') - f_2(T,T')|]$, *where* T *and* T' *are defined as in section* 1.

Proof. To simplify, we denote

$$\Psi_i(x,y) \;:\; \operatorname*{Max}_G \operatorname*{Min}_{G'} F(x,y;f_i;G,G')$$
$$= \operatorname*{Max}_{G'} \operatorname*{Min}_G [- - - - - -].$$

Let G_i, G_i' be a pair of functions yielding the optimal value of $\Psi_i(x,y)$ for $i = 1, 2$.

Then $\Psi_1(x,y) = F(x,y;f_1;G_1,G_2') \ge F(x,y;f_1;G_2,G_1')$ and $\Psi_1(x,y) \le F(x,y;f_1;G_1,G_2')$.

Again, $\Psi_2(x,y) = F(x,y;f_2;G_2,G_2') \ge F(x,y;f_2;G_1,G_2')$ and $\Psi_2(x,y) \le F(x,y;f_2;G_2,G_1')$.

Combining these inequalities, we get

$$\int_u \int_v [h(x,y;u,v)(f_1(T,T') - f_2(T,T'))] dG_2(u) dG_1'(v)$$
$$\le \Psi_1(x,y) - \Psi_2(x,y)$$
$$\le \int_u \int_v [h(x,y;u,v)(f_1(T,T') - f_2(T,T'))] dG_1(u) dG_2'(v).$$

This implies that

$$|\Psi_1(x,y) - \Psi_2(x,y)| \le \operatorname*{Max}_u \operatorname*{Max}_v [|h(x,y;u,v)|\,|f_1(T,T') - f_2(T,T')|]. \qquad \square$$

LEMMA 2.2. *Let* $\langle M, \rho\rangle$ *be a complete metric space.* A *is a mapping from* M *into itself. If the following conditions hold, then* A *has a unique fixed point.*

(i) *For any* x, $y \in M$, $\rho(Ax, Ay) \le \phi(\rho(x,y))$, *where* $\phi : [0,\infty) \to [0,\infty)$ *is nondecreasing continuous on the right and* $\phi(r) < r$ *for* $r > 0$.

(ii) *For every* $x \in M$, *there is a positive number* k *such that* $\rho(x, A^n x) \le k$ *for all* n.

LEMMA 2.3. *Let* $\langle M, \rho\rangle$ *be a complete metric space, and let* A *be a mapping of* M *into itself. Suppose that, for all* x, y *in* M, $\rho(Ax, Ay) \le \phi(\rho(x,y))$, *where* $\phi : [0,\infty) \to [0,\infty)$ *is nondecreasing, and, for every positive number* r, *the series* $\sum \phi^n(r)$ *is convergent. Then* A *has a unique fixed point.*

Lemma 2.2 is an extension of Brauer's fixed-point theorem [3], and the proof of Lemma 2.3 is in light of [4, Theorem 3.2, p. 12], which is easy and straightforward.

3. Solution of the functional equation.

THEOREM 3.1. *The functional equation* (1) *possesses unique bounded solution on* S *under the following conditions:*

(i) R *and* h *are bounded;*

(ii) $\mathrm{Max}_u\mathrm{Max}_v[|h(x,y;u,v)|\,|f_1(T,T') - f_2(T,T')|] \le \phi(\mathrm{Max}_u\mathrm{Max}_v[|f_1(T,T') - f_2(T,T')|])$,

where $\phi : [0,\infty) \to [0,\infty)$ *is nondecreasing and continuous on the right such that* $\phi(r) < r$ *for* $r > 0$.

Proof. Let $B(S)$ be the set of all real valued bounded functions on S. For Ψ_1, $\Psi_2 \in B(S)$, let

$$\rho(\Psi_1,\Psi_2) = \operatorname*{Max}_u \operatorname*{Max}_v \left[|\Psi_1(x,y) - \Psi_2(x,y)|\right].$$

Then ρ is a metric on $B(S)$ and $\langle B(S),\rho\rangle$ is a complete metric space. Let A be any function defined on $B(S)$ by $Ag = \Psi$ for any $g \in B(S)$,

$$\Psi(x,y) \;:\; \operatorname*{Max}_G \operatorname*{Min}_{G'} \left[\int_u\int_v [R(x,y;u,v) + h(x,y;u,v)g(T,T')]\,dG(u)dG'(v)\right]$$
$$= \operatorname*{Min}_{G'} \operatorname*{Max}_G[-----].$$

Since R, h, and g are bounded, Ψ is also bounded, and $\Psi \in B(S)$. Hence $A : B(S) \to B(S)$. For $\Psi_1\Psi_2 \in B(S)$, let

$$\Psi_i(x,y) : \operatorname*{Max}_G \operatorname*{Min}_{G'} \left[\int_u\int_v [R(x,y;u,v) + h(x,y;u,v)g_i(T,T')]dG(u)dG'(v)\right]$$
$$\operatorname*{Min}_{G'} \operatorname*{Max}_G[-----].$$

Let G_i, G_i' be the distribution functions yielding $\Psi_i(x,y)$ for $i = 1,2$. Then, using Lemma 2.1 and condition (ii), we have

$$|\Psi_1(x,y) - \Psi_2(x,y)| \le \phi\left(\operatorname*{Max}_u \operatorname*{Max}_v[|g_1(T,T') - g_2(T,T')|]\right) = \phi(\rho(g_1,g_2)),$$

i.e., $\rho(Ag_1,Ag_2) \le \phi(\rho(g_1,g_2))$. Let $A^ng = g_n$ for $g \in B(S)$, where

$$g_n(x,y) \;:\; \operatorname*{Max}_G \operatorname*{Min}_{G'} \left[\int_u\int_v [R(x,y;u,v) + h(x,y;u,v)g_{n-1}(T,T')]dG(u)dG'(v)\right]$$
$$= \operatorname*{Min}_{G'} \operatorname*{Max}_G[-----]$$

for $n = 2,3,\ldots$, and

$$g_1(x,y) \;:\; \operatorname*{Max}_G \operatorname*{Min}_{G'} \left[\int_u\int_v R(x,y;u,v)dG(u)dG'(v)\right]$$
$$= \operatorname*{Min}_{G'} \operatorname*{Max}_G[-----].$$

Since R, h, and g are all bounded functions, $|R(x,y;u,v)| \le \lambda_1$, $|h(x,y;u,v)g_n(T,T')| \le \lambda_2$, and $|g(x,y)| \le \lambda_3$ for all $(x,y) \in S$, $(u,v) \in D$, $g(T,T') \in R$, where λ_1, λ_2, λ_3 are constants. This yields that $|g_n(x,y)| \le \lambda_1 + \lambda_2\lambda_3$ for all $(x,y) \in S$. Now

$$|g(x,y) - g_n(x,y)| \le |g(x,y)| + |g_n(x,y)| \le \lambda_1 + \lambda_2\lambda_3 + \lambda_3 = \lambda \text{ for all } n.$$

So $\rho(g,g_n) = \rho(g,A^ng) \le \lambda$ for all n. Therefore, by Lemma 2.2, the mapping A has a unique fixed point; i.e., the functional equation (1) has a unique bounded solution on S. ∎

THEOREM 3.2. *Let $h: S \times D \to \mathbf{R}$ be such that $|h(x,y;u,v)| \le a < 1$ for all $(x,y) \in S$, $(u,v) \in D$ is bounded, and condition* (ii) *of Theorem* 3.1 *is satisfied. Then the functional equation* (1) *possesses a unique bounded solution on S.*

Proof. Let $\phi : [0,\infty) \to [0,\infty)$ be defined by $\phi(r) = ar$, $r > 0$. Then ϕ is nondecreasing and continuous on the right. Also $\sum \phi^n(r) = r \sum a^n$. Since $a < 1$, $\sum \phi^n(r)$ is convergent for a very positive number r. Proceeding as in Theorem 3.1 and using Lemma 2.1, we have

$$\begin{aligned}\rho(Ag_1, Ag_2) = \rho(\Psi_1, \Psi_2) &\le \operatorname*{Max}_u \operatorname*{Max}_v [|h(x,y;u,v)|\,|g_1(x,y) - g_2(x,y)|] \\ &\le \phi\left(\operatorname*{Max}_u \operatorname*{Max}_v [|g_1(x,y) - g_2(x,y)|]\right) \\ &= \phi(\rho(g_1, g_2)).\end{aligned}$$

Hence, by Lemma 2.3 the functional equation (1) possesses a unique bounded solution on S. □

THEOREM 3.3. *Suppose the following conditions hold:*

(i) $|f(T,T')| \le \|(T,T')\| \le \phi(\|(x,y)\|)$ *for all* $(x,y) \in S$, *where* ϕ: $[0,\infty] \to [0,\infty)$ *is nondecreasing, and, for every positive number* r, $\sum \phi^n(r)$ *is convergent.*

(ii) $0 \le R(x,y;u,v) \le \|(x,y)\|$ *for all* $(u,v) \in D$.

(iii) $0 \le h(x,y;u,v) f(T,T') \le 1 f(T,T')|$.

Then the functional equation (1) *possesses a unique solution on S.*

Proof. Let $\{f_n\}$ be a sequence of functions defined on S by

$$\begin{aligned}f_0(x,y) \;:\; &\operatorname*{Max}_G \operatorname*{Min}_{G'} \left[\int_u \int_v R(x,y;u,v)\,dG(u)\,dG'(v)\right] \\ &= \operatorname*{Min}_{G'} \operatorname*{Max}_G [------].\end{aligned}$$

and

$$\begin{aligned}f_{n+1}(x,y) \;:\; &\operatorname*{Max}_G \operatorname*{Min}_{G'} \left[\int_u \int_v R(x,y;u,v) + h(x,y;u,v) f_n(T,T')\,dG(u)\,dG'(v)\right] \\ &= \operatorname*{Min}_{G'} \operatorname*{Max}_G [------].\end{aligned}$$

By (ii), $f_0(x,y) \ge 0$ for all $(x,y) \in S$.

$$\begin{aligned}f_1(x,y) \;:\; &\operatorname*{Max}_G \operatorname*{Min}_{G'} \left[\int_u \int_v R(x,y;u,v) + h(x,y;u,v) f_0(T,T')\,dG(u)\,dG'(v)\right] \\ &= \operatorname*{Min}_{G'} \operatorname*{Max}_G [------] \\ &\ge f_0(x,y).\end{aligned}$$

Let $f_n(x,y) \ge f_{n-1}(x,y)$.

Then

$$\begin{aligned}f_{n+1}(x,y) \;:\; &\operatorname*{Max}_G \operatorname*{Min}_{G'} F(x,y;f_n;G,G') \\ &= \operatorname*{Min}_{G'} \operatorname*{Max}_G [------].\end{aligned}$$

As in Lemma 2.1, if G_1, G_1' and G_2, G_2' yield the value of $f_{n+1}(x,y)$ and $f_n(x,y)$, we get

$$f_{n+1}(x,y) - f_n(x,y) \geq \int_u \int_v h(x,y;u,v)[f_n(T,T') - f_{n-1}(T,T')]dG_2'(u)dG_1'(v) \geq 0;$$

i.e., $f_{n+1}(x,y) \geq f_n(x,y)$ for all n. Thus $\{f_n(x,y)\}$ is a monotonically increasing sequence. To show that $\{f_n(x,y)\}$ is bounded above, let $(x,y) \in S$ and r be a positive number such that $\|(x,y)\| = r$. Since $0 \leq R(x,y;u,v) \leq \|(x,y)\|$ for all $(u,v) \in D$, $0 \leq f_0(x,y) \leq \|(x,y)\|$. Again,

$$\begin{aligned} 0 \leq f_1(x,y) &\leq R(x,y;u,v) + h(x,y;u,v)f_0(T,T') \\ &\leq \|(x,y)\| + |f_0(T,T')| \\ &\leq \|(x,y)\| + \|(T,T')\|. \end{aligned}$$

Thus $0 \leq f_1(x,y) \leq \|(x,y)\| + \phi(\|(x,y)\|)$. Next

$$\begin{aligned} 0 \leq f_2(x,y) &\leq R(x,y;u,v) + h(x,y;u,v)f_1(T,T') \\ &\leq \|(x,y)\| + |f_1(T,T')| \\ &\leq \|(x,y)\| + \|(T,T')\| + \phi(\|(T,T')\|) \\ &\leq \|(x,y)\| + \phi\|(x,y)\|) + \phi^2(\|(x,y)\|); \end{aligned}$$

i.e., $0 \leq f_2(x,y) \leq r + \phi(r) + \phi^2(r)$. Proceeding in this way, we obtain $0 \leq f_n(x,y) \leq \sum \phi^n(r)$. Since $\sum \phi^n(r)$ is convergent, $\{f_n(x,y)\}$ is bounded hence convergent.

Let $\bar{f}(x,y) = \lim_{n\to\infty} f_n(x,y)$.

Next show that $\bar{f}(x,y)$ is the solution of the functional equation (1). Since f_n is increasing and $\bar{f}$ is its limit, $f_{n+1}(x,y) \leq \bar{f}(x,y)$; i.e.,

$$\operatorname*{Max}_{G} \operatorname*{Min}_{G'} \left[\int_u \int_v R(x,y;u,v) + h(x,y;u,v)f_n(T,T')dG(u)dG'(v)\right] \leq \bar{f}(x,y).$$

Letting $n \to \infty$,

$$\operatorname*{Max}_{G} \operatorname*{Min}_{G'} \left[\int_u \int_v R(x,y;u,v) + h(x,y;u,v)\bar{f}(T,T')dG(u)dG'(v)\right] \leq \bar{f}(x,y). \tag{2}$$

For any $(x,y) \in S$, $(u,v) \in D$, and any positive integer n, we have

$$f_n(x,y) \leq \operatorname*{Max}_{G} \operatorname*{Min}_{G'} \left[\int_u \int_v R(x,y;u,v) + h(x,y;u,v)\bar{f}(T,T')dG(u)dG'(v)\right].$$

Letting $n \to \infty$,

$$\bar{f}(x,y) \leq \operatorname*{Max}_{G} \operatorname*{Min}_{G'} \left[\int_u \int_v R(x,y;u,v) + h(x,y;u,v)\bar{f}(T,T')dG(u)dG'(v)\right]. \tag{3}$$

From (2) and (3),

$$\begin{aligned} \bar{f}(x,y) &\leq \operatorname*{Max}_{G} \operatorname*{Min}_{G'} \left[\int_u \int_v R(x,y;u,v) + h(x,y;u,v)\bar{f}(T,T')dG(u)dG'(v)\right] \\ &= \operatorname*{Min}_{G'} \operatorname*{Max}_{G}[------]. \end{aligned}$$

If we let $Ax = T(x,y;u,v)$ and $Ay = T'(x,y;u,v)$, then $\|(Ax,Ay)\| = \|(T,T')\| \leq \phi(\|(x,y)\|)$, and the above solution is unique by Lemma 2.3. □

4. Conclusion (effective of the solution). In a multistage allocation process, the functional equation discussed above has some importance in game theory. We have established the existence of its solution under several assumptions that this infinite process admits a value to both of the players. The solution is effective if it gives sufficient information to the players to obtain this value. Suppose that G, G' yield the value of the game. Then the corresponding solution becomes effective if A uses a distribution function $G(u)$; whatever B may use, A can guarantee himself a return of at least $f(x, y)$. At this fixed strategy, A's return will be at worst determined by the functional equation

$$F(x,y) \leq \operatorname*{Min}_{G'} \left[\int_u \int_v R(x,y;u,v) + h(x,y;u,v) F(T,T') dG(u) dG'(v) \right].$$

Under the assumptions of our theorems, the above equation has a unique solution, and the solution is obtained by the successive approximation of the functions

$$F_0(x,y) = \operatorname*{Min}_{G'} \int_u \int_v R(x,y;u,v) dG(u) dG'(v),$$

$$F_{n+1}(x,y) = \operatorname*{Min}_{G'} \left[\int_u \int_v R(x,y;u,v) + h(x,y;u,v) f_n(T,T') dG(u) dG'(v) \right].$$

It is clear that $F = f_1$ and $F_{n+1} = f_{n+1}$. Thus $F(x, y) = \lim_{n\to\infty} F_n = \lim_{n\to\infty} f_n = f(x, y)$.

Hence the solution is effective. A similar result holds if B uses a distribution function $G'(v)$, whatever A may use.

Acknowledgment. The authors are thankful to the referees for their valuable suggestions in preparing the final version of this paper.

REFERENCES

[1] R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.
[2] P. C. Bhakta and S. Mitra, *Some existence theorems for functional equations arising in dynamic programming*, J. Math. Anal. Appl., 98 (1984), pp. 340–361.
[3] F. Brauer, *A note on uniqueness and convergence of successive approximations*, Canad. Math. Bull., 2 (1959), pp. 5–8.
[4] J. Dugundji and A. Granas, *Fixed Point Theory*, PWN Polish Scientific Publications, Warsaw, 1984.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 3, pp. 826–850

LIMIT HAMILTON–JACOBI–ISAACS EQUATIONS FOR SINGULARLY PERTURBED ZERO-SUM DYNAMIC (DISCRETE TIME) GAMES*

PENG SHI†

Abstract. In this paper we study a singularly perturbed zero-sum dynamic game with full information. We introduce the upper (lower) value function of the dynamic game, in which the minimizer (maximizer) can be guaranteed if at the beginning of each interval his move (the choice of decision) precedes the move of the maximizer (minimizer). We show that when the singular perturbations parameter tends to zero, the upper (lower) value function of the dynamic game has a limit which coincides with a viscosity solution of a Hamilton–Jacobi–Isaacs-type equation. Two examples are given to demonstrate the potential of the proposed technique.

Key words. discrete time games, Hamilton–Jacobi–Isaacs equations, singularly perturbed systems, value functions, viscosity solutions

AMS subject classifications. 91A25, 91A50, 49L25

PII. S036301290037908X

1. Introduction.

Singularly perturbed control systems (SPCS) evolving in a discrete time scale arise in many applications as well as in the construction of the difference approximations of SPCS evolving in continuous time. Despite this, discrete time SPCS were studied in the literature much less intensively than their continuous time counterparts. A reason for such a discrepancy might be the fact that a common approach to SPCS was one based on the boundary layer method [25, 30, 31], which is not easy to adapt to deal with the discrete time SPCS.

Recently, a number of averaging-type methods allowing one to treat general SPCS in continuous time were developed (see, e.g., [1, 2, 8, 9, 10, 11, 16, 18, 32]). These methods are much more adaptable to the discrete time scale. For example, a full analogy between the averaging procedures in problems of optimal control of SPCS evolving in continuous and discrete times was established in [17].

In this paper, we extend a similar analogy by showing that the results about averaging in singularly perturbed (SP) zero-sum differential games obtained in [12] have their counterparts in a discrete time setting. More specifically, we will show that the upper and lower values of an SP dynamic (discrete time) game with full information converge (as the SP parameter tends to zero) to the viscosity solutions of some Hamilton–Jacobi-type equations, with Hamiltonians being the limit averages of the upper and (respectively) lower value functions of a certain associated "fast" discrete time game. This result leads, in particular, to a very important conclusion that states that if the limit averages of the upper and lower values of the associated fast game coincide, then the limits of the upper and lower values of the original discrete time SP game coincide too and, thus, the latter has value in the limit as the singular perturbations parameter tends to zero.

*Received by the editors October 5, 2000; accepted for publication (in revised form) March 2, 2002; published electronically September 12, 2002.

http://www.siam.org/journals/sicon/41-3/37908.html

†Land Operations Division, Defence Science and Technology Organisation, Department of Defence, P.O. Box 1500, Edinburgh 5111 SA, Australia (peng.shi@dsto.defence.gov.au). This work was carried out while the author was with the Centre for Industrial and Applicable Mathematics, School of Mathematics, The University of South Australia.

An important issue in the theory of SPCS is a justification of a so-called reduction technique approach (RTA). According to this approach the fast variables are replaced by their steady states obtained with "frozen" slow variables and controls, and the slow dynamics is approximated by the corresponding reduced order system. Although the RTA may fail to provide a proper approximation for the SPCS in a general case [9, 10, 11], its application was very successful in many important special cases (see [15, 24, 22, 23, 28, 29] and the references therein). In the differential game context the efficiency of the RTA was established for SP linear quadratic games in [14, 21] and for the SP H^∞ problem with linear dynamics in [26, 27].

The general averaging approach that we develop in this paper allows us to verify applicability of the RTA in SP discrete time games. This point is illustrated by a special example with linear fast variables and controls dynamics.

It should be mentioned that the motivation for the study in this paper stems from the work of [7] and [12]. In [7], the differential game and its properties were studied, but by discretizing the continuous time systems and using a discrete approach. By looking at the behavior of small time intervals (tending to zero), the solution to the differential game was thus provided. While stimulated by the work of [7], the continuous time counterpart of this paper was investigated in [12]. Both the results obtained in [12] and in this paper were encouraged by the pioneering work of [7], but emphasis is laid on the impact of singular perturbations and its asymptotic behavior. By the results in [7] and [12], the PDE (2.24) would be helpful to solve discrete time games by looking at the discrete instant behavior. As we mention in Remark 4.3, Theorem 4.1 is asymptotic in nature. Our emphasis is to establish the convergence of the upper and lower values of the singularly perturbed dynamic game to the viscosity solutions of corresponding limited Hamilton–Jacobi–Isaacs (LHJI) equations.

This paper is organized as follows. In section 2, we describe our main results without going into technical details. In section 3, these results are illustrated by a consideration of a special class of SP discrete time games and two examples, one of which illustrates an applicability of RTA. In section 4, we introduce our main assumptions and establish one theorem about the convergence of the upper and lower value functions of an SP discrete time zero-sum game to the viscosity solutions of the corresponding LHJI equations (defined in section 2). Assumptions 4.4, 4.5, and 4.6 are verified in section 5.

Notation. Throughout this paper R^n and $R^{n\times m}$ denote, respectively, the n-dimensional Euclidean space and the set of all $n \times m$ real matrices. $\|\cdot\|$ will refer to the L_∞ norm in the finite-dimensional space. That is, for $q \in R^k$ and $A \in R^{n\times k}$, $\|q\| = \max_{i=1,2,\ldots,k} |q_i|$ and $\|A\| = \max_{\|q\|=1} \|Aq\|$. $\lfloor x \rfloor$ stands for the greatest integer which is smaller than or equal to x.

2. Problem formulation and preliminaries.

2.1. Singularly perturbed discrete time game. Consider a discrete time system

$$z(k+1) = z(k) + \epsilon f_1\left(z(k), y(k), u_k, v_k\right), \tag{2.1}$$

$$y(k+1) = y(k) + f_2\left(z(k), y(k), u_k, v_k\right) \tag{2.2}$$

for $k = l, l+1, \ldots, N_\epsilon - 1$ with $N_\epsilon \stackrel{\text{def}}{=} \lfloor \frac{T}{\epsilon} \rfloor$ and

$$z(l) = z, \quad y(l) = y, \tag{2.3}$$

where $l = 0, 1, \ldots, N_\epsilon - 1$. Here ϵ is a small positive parameter, $T > 0$ defines the final time, and $f_1(\cdot)$ and $f_2(\cdot)$ are maps from $R^m \times R^n \times R^p \times R^q$ to R^m and R^n, respectively.

The appearance of the small parameter in the right-hand side of (2.1) implies that the rate of change of the z-components of the phase vector is of the order ϵ and, thus, this component can be considered to be slow with respect to the y-components changing with the rate $O(1)$. The number of discrete moments of time in which the phase vector changes its values according to (2.1)–(2.2) has the order $\frac{1}{\epsilon}$. Hence, the slow components z may have significant (not tending with ϵ to zero) deviations from their initial values. Systems allowing similar properties on a continuous time scale are commonly called singularly perturbed (SP). Adopting this terminology, we shall call (2.1)–(2.2) a *singularly perturbed discrete time (SPDT) system*.

Assume that there are two controllers in our system, one responsible for the choice of u_k (maximizer) and the other for the choice of v_k (minimizer). The controls are chosen to satisfy the inclusions

$$u_k \in U, \quad v_k \in V, \quad k = l, l+1, \ldots, N_\epsilon - 1, \tag{2.4}$$

where U and V are given compact subsets of R^p and R^q, respectively. Motivated by the results in [7], we consider a sequence, $l = 0, 1, \ldots, N_\epsilon - 1$, of the dynamic games each starting at the moment l and consisting of $N_\epsilon - l$ steps. The strategies of the players are defined in these games as follows:

Let Γ_l^l and Δ_l^l be any vectors in U and V. Let Γ_{l+k}^l and Δ_{l+k}^l be any maps:

$$\Gamma_{l+k}^l : \quad \underbrace{(U \times V) \times (U \times V) \times \cdots \times (U \times V)}_{k} \to U,$$

$$\Delta_{l+k}^l : \quad \underbrace{(U \times V) \times (U \times V) \times \cdots \times (U \times V)}_{k} \to V$$

with $k = 1, \ldots, N_\epsilon - l - 1$.

The vector

$$\Gamma^l = \left(\Gamma_l^l, \ldots, \Gamma_{N_\epsilon - 1}^l\right)$$

is called a *strategy for the maximizer* in the game starting at the moment l, while the vector

$$\Delta^l = \left(\Delta_l^l, \ \ldots, \ \Delta_{N_\epsilon - 1}^l\right)$$

is called a *strategy for the minimizer* in the game starting at the moment l.

Given any pair (Γ^l, Δ^l) one can uniquely construct the control sequences

$$u^l = (u_k, \ k = l, \ldots, N_\epsilon - 1), \quad v^l = (v_k, \ k = l, \ldots, N_\epsilon - 1)$$

with $u_l \stackrel{\text{def}}{=} \Gamma_l^l$, $v_l \stackrel{\text{def}}{=} \Delta_l^l$ and

$$u_k = \Gamma_k^l \left(u_l, v_l, \ldots, u_{k-1}, v_{k-1}\right), \tag{2.5}$$

$$v_k = \Delta_k^l \left(u_l, v_l, \ldots, u_{k-1}, v_{k-1}\right) \tag{2.6}$$

for $k = l + 1, \ldots, N_\epsilon - 1$. This pair of control sequences is called the outcome of (Γ^l, Δ^l).

Let (Γ^l, Δ^l) be a pair of strategies of the players in the game starting at the moment l, let $(u^l, v^l) = \{(u_k, v_k),\ k = l, l+1, \dots, N_\epsilon - 1\}$ be its outcome, and let $(z(k), y(k))$ be the corresponding solution of (2.1)–(2.2). Define the payoff of the game by the equation

$$P_\epsilon(l, z, y, \Gamma^l, \Delta^l) = P_\epsilon(l, z, y, u^l, v^l) \stackrel{\text{def}}{=} G(z(N_\epsilon)) + \epsilon \sum_{k=l}^{N_\epsilon - 1} \Phi\left(z(k), y(k), u_k, v_k\right),$$

$$l = 0, 1, \dots, N_\epsilon - 1, \tag{2.7}$$

where $G : R^m \to R^1, \Phi : R^m \times R^n \times R^p \times R^q \to R^1$.

Define the *upper* and *lower value functions* of the SPDT game as follows:

$$B_\epsilon^{up}(l, z, y) = \inf_{\Delta_l^l} \sup_{\Gamma_l^l} \dots \inf_{\Delta_{N_\epsilon - 1}^l} \sup_{\Gamma_{N_\epsilon - 1}^l} P_\epsilon\ (l, z, y, \Gamma^l, \Delta^l), \quad l = 0, 1, \dots, N_\epsilon - 1 \tag{2.8}$$

$$B_\epsilon^{lo}(l, z, y) = \sup_{\Gamma_l^l} \inf_{\Delta_l^l} \dots \sup_{\Gamma_{N_\epsilon - 1}^l} \inf_{\Delta_{N_\epsilon - 1}^l} P_\epsilon(l, z, y, \Gamma^l, \Delta^l), \quad l = 0, 1, \dots, N_\epsilon - 1, \tag{2.9}$$

and also take

$$B_\epsilon^{up}(N_\epsilon, z, y) = B_\epsilon^{lo}(N_\epsilon, z, y) = G(z). \tag{2.10}$$

Given any l, z, and y, $B_\epsilon^{up}(l, z, y)$ is the value of the payoff functional (2.7) which the v-player (the minimizer) can be guaranteed if at the beginning of each interval k his move (the choice of Δ_k) precedes the move of the u-player (the choice of Γ_k), whereas $B_\epsilon^{lo}(l, z, y)$ is the value which the u-player (the maximizer) can be guaranteed if at the beginning of each interval his move precedes the move of the minimizer.

Notice that using an argument similar to that in [7], one can show that $B_\epsilon^{up(lo)}(l, z, y)$ also allows the following representations:

$$B_\epsilon^{up}(l, z, y) = \inf_{v_l} \sup_{u_l} \dots \inf_{v_{N_\epsilon - 1}} \sup_{u_{N_\epsilon - 1}} P_\epsilon(l, z, y, u^l, v^l), \tag{2.11}$$

$$B_\epsilon^{lo}\ (l, z, y) = \sup_{u_l} \inf_{v_l} \dots \sup_{u_{N_\epsilon - 1}} \inf_{v_{N_\epsilon - 1}} P_\epsilon(l, z, y, u^l, v^l), \tag{2.12}$$

where the sups and infs are sought over $u_l \in U$ and $v_l \in V$ and $(u^l, v^l) = (u_k, v_k),\ k = l, l+1, \dots, N_\epsilon - 1$.

Definition 2.1. *We shall say that the* SPDT game has value in the limit *if corresponding to any compact set $D_1 \times P_1 \subset R^m \times R^n$ there exists a function $\mu(\epsilon)$ tending to zero as ϵ tends to zero such that*

$$\left|B_\epsilon^{up}\ (l, z, y) - B_\epsilon^{lo}\ (l, z, y)\right| \le \mu(\epsilon) \quad \forall l = 0, 1, \dots, N_\epsilon - 1,\ (z, y) \in D_1 \times P_1. \tag{2.13}$$

2.2. Associated fast discrete time game. Consider a system

$$\tilde{y}(k+1) = \tilde{y}(k) + f_2(z, \tilde{y}(k), \tilde{u}_k, \tilde{v}_k), \qquad \tilde{y}(0) = y, \quad z = constant,$$
$$k = 0, 1, \dots, N - 1, \tag{2.14}$$

where N is a positive integer number. In contrast to (2.2), z here is a vector of constant parameters. System (2.14) is called an *associated system*.

Let $\tilde{\Gamma}_0$ and $\tilde{\Delta}_0$ be any functions in U and V and let $\tilde{\Gamma}_j$ and $\tilde{\Delta}_j$ be any map from

$$\underbrace{(U \times V) \times \dots \times (U \times V)}_{j}$$

into U and V, respectively, for $j = 1, \ldots, N-1$. The vectors

$$\tilde{\Gamma} = (\tilde{\Gamma}_0, \ldots, \tilde{\Gamma}_{N-1}), \quad \tilde{\Delta} = (\tilde{\Delta}_0, \ldots, \tilde{\Delta}_{N-1})$$

are called *strategies* for $\tilde{u}$- and $\tilde{v}$-players in the associated fast discrete time (AFDT) game, respectively. Similar to (2.5)–(2.6), given a pair $(\tilde{\Gamma}, \tilde{\Delta})$, one can construct the corresponding outcome, that is, the sequence of control pairs

$$(\tilde{u}, \tilde{v}) = (\tilde{u}_k, \tilde{v}_k), \qquad k = 0, 1, \ldots, N-1,$$

with

$$\tilde{u}_k \in U, \quad \tilde{v}_k \in V, \qquad k = 0, 1, \ldots, N-1, \tag{2.15}$$

and thus determine the solution $y_z = \tilde{y}_z(k)$ of (2.14).

Define the *payoff* of the AFDT game corresponding to the pair $(\tilde{\Gamma}, \tilde{\Delta})$ by the equation

$$\begin{aligned} Q(z, \lambda, N, y, \tilde{\Gamma}, \tilde{\Delta}) \; &= \; Q(z, \lambda, N, y, \tilde{u}, \tilde{v}) \\ &\stackrel{\text{def}}{=} N^{-1} \sum_{k=0}^{N-1} \Big[\Phi\left(z, \tilde{y}_z(k), \tilde{u}_k, \tilde{v}_k\right) + \lambda^T f_1\left(z, \tilde{y}_z(k), \tilde{u}_k, \tilde{v}_k\right) \Big], \end{aligned} \tag{2.16}$$

where Φ is the same as in the payoff functional (2.7), $f_1(\cdot)$ is the function defining the slow subsystem (2.1), and $z \in R^m$ and $\lambda \in R^m$ are vectors of constant parameters. The *upper* and *lower value functions* of the AFDT game are defined by the following equations:

$$R^{up}(z, \lambda, N, y) = \inf_{\tilde{\Delta}_0} \sup_{\tilde{\Gamma}_0} \ldots \inf_{\tilde{\Delta}_{N-1}} \sup_{\tilde{\Gamma}_{N-1}} \; Q(z, \lambda, N, y, \tilde{\Gamma}, \tilde{\Delta}), \tag{2.17}$$

$$R^{lo}(z, \lambda, N, y) = \sup_{\tilde{\Gamma}_0} \inf_{\tilde{\Delta}_0} \ldots \sup_{\tilde{\Gamma}_{N-1}} \inf_{\tilde{\Delta}_{N-1}} \; Q(z, \lambda, N, y, \tilde{\Gamma}, \tilde{\Delta}). \tag{2.18}$$

Similar to (2.11) and (2.12), $R^{up(lo)}(z, \lambda, N, y)$ allows the representations [7]

$$R^{up}(z, \lambda, N, y) = \inf_{\tilde{v}_0} \sup_{\tilde{u}_0} \ldots \inf_{\tilde{v}_{N-1}} \sup_{\tilde{u}_{N-1}} \; Q(z, \lambda, N, y, \tilde{u}, \tilde{v}), \tag{2.19}$$

$$R^{lo}(z, \lambda, N, y) = \sup_{\tilde{u}_0} \inf_{\tilde{v}_0} \ldots \sup_{\tilde{u}_{N-1}} \inf_{\tilde{v}_{N-1}} \; Q(z, \lambda, N, y, \tilde{u}, \tilde{v}), \tag{2.20}$$

where the sups and infs are sought over $\tilde{u}_j \in U$ and $\tilde{v}_j \in V$.

In what follows we shall use the following assumption about the upper and lower value functions of the AFDT game.

Assumption 2.1. There exist the limits

$$\lim_{N \to \infty} \; R^{up}(z, \lambda, N, y) \stackrel{\text{def}}{=} \; R^{up}(z, \lambda), \tag{2.21}$$

$$\lim_{N \to \infty} \; R^{lo}(z, \lambda, N, y) \stackrel{\text{def}}{=} \; R^{lo}(z, \lambda) \tag{2.22}$$

which do not depend on the initial values y in (2.14).

Remark 2.1. Assumption 2.1 is a discrete-time counterpart of the one introduced in [12]. Note that from a control theory point of view, if a system is controllable, then it is possible to drive the system from any initial state to any specified state within a certain time. Consequently, the limit behavior of the upper and lower value functions of the AFDT game should not depend upon the initial state y.

We shall say that the *AFDT game has value in the limit* if

$$R^{up}(z, \lambda) = R^{lo}(z, \lambda) \stackrel{\text{def}}{=} R(z, \lambda) \quad \forall (z, \lambda) \in R^m \times R^m. \tag{2.23}$$

2.3. LHJI equations for the SPDT game. Let us consider the Hamilton–Jacobi-type equations

$$-\frac{\partial B(t,z)}{\partial t} + H\left(z, \frac{\partial B(t,z)}{\partial z}\right) = 0, \qquad (t,z) \in [0,T) \times R^m, \tag{2.24}$$

with Hamiltonians $H(z,\lambda)$ being equal to $-R^{up}(z,\lambda)$, $-R^{lo}(z,\lambda)$, or $-R(z,\lambda)$ in the case in which (2.23) is true. These equations will be referred to as LHJI equations for the SPDT game. Let us denote by $B^{up}(t,z)$, $B^{lo}(t,z)$, $B(t,z)$ the viscosity solutions of these equations which satisfy the boundary condition

$$B(T,z) = G(z) \quad \forall z \in R^m. \tag{2.25}$$

In the following sections it will be shown that the upper and lower value functions of the SPDT game, $B^{up}_\epsilon(l,z)$ and $B^{lo}_\epsilon(l,z)$, converge to $B^{up}(t,z)$ and $B^{lo}(t,z)$, respectively, as ϵ tends to zero. This will imply in particular that if the AFDT game has the value in the limit, that is, (2.23) is true, then both $B^{up}_\epsilon(t,z)$ and $B^{lo}_\epsilon(t,z)$ converge to $B(t,z)$ as ϵ tends to zero, and thus the SPDT game has the value in the limit as well.

As in SP differential games, the above results can be considered to be a justification of a decomposition of the SPDT game into the AFDT game, allowing one to describe an asymptotically optimal behavior of the players if the slow parameters are fixed and the LHJI equations are responsible for a "near-optimality" of the slow dynamics.

Notice that for some classes of dynamic games (see the examples in section 3) the SPDT and AFDT games can be interpreted as in the Isaacs equations [19] for some specially constructed "slow" differential games.

3. Special case. To illustrate the results mentioned above let us consider a special case. Let

$$y^T = (y_1^T, y_2^T), \quad f_2^T(z,y,u,v) = \left(f_{21}^T(z,y_1,u), f_{22}^T(z,y_2,v)\right), \tag{3.1}$$

$$y^T(0) = (y_{10}^T, y_{20}^T), \qquad y_i \in R^{n_i},\ f_{2i} \in R^{n_i},\ n_1 + n_2 = n, \tag{3.2}$$

$$f_1(z,y,u,v) = f_{10}(z) + f_{11}(z,y_1,u) + f_{12}(z,y_2,v), \tag{3.3}$$

$$\Phi(z,y,u,v) = \Phi_1(z,y_1,u) + \Phi_2(z,y_2,v). \tag{3.4}$$

It can be easily shown that under the above assumptions about the structure of the SPDT game, the AFDT game is equivalent to two optimal control problems:

$$\sup_{u_k \in U} \Bigg\{ N^{-1} \sum_{k=0}^{N-1} \left[\Phi_1\left(z, y_1(k), u_k\right) + \lambda^T f_{11}\left(z, y_1(k), u_k\right)\right] \Bigg| \ y_1(k+1) = y_1(k) + f_{21}(z, y_1(k), u_k),\ y_1(0) = y_{10} \Bigg\} \stackrel{\text{def}}{=} r_1(z, \lambda, N, y_{10}) \tag{3.5}$$

and

$$\inf_{v_k \in V} \Bigg\{ N^{-1} \sum_{k=0}^{N-1} \left[\Phi_2\left(z, y_2(k), v_k\right) + \lambda^T f_{12}\left(z, y_2(k), v_k\right)\right] \Bigg| \ y_2(k+1) = y_2(k) + f_{22}(z, y_2(k), v_k),\ y_2(0) = y_{20} \Bigg\} \stackrel{\text{def}}{=} r_2(z, \lambda, N, y_{20})\ . \tag{3.6}$$

It can be shown, in particular, that

$$\begin{aligned} R^{up}(z,\lambda,N,y) &= R^{lo}(z,\lambda,N,y) \\ &= \lambda^T f_{10}(z) + r_1(z,\lambda,N,y_{10}) + r_2(z,\lambda,N,y_{20}), \end{aligned} \tag{3.7}$$

which implies that

$$R^{up}(z,\lambda) = R^{lo}(z,\lambda) = R(z,\lambda) = \lambda^T f_{10}(z) + r_1(z,\lambda) + r_2(z,\lambda), \tag{3.8}$$

where $r_i(z,\lambda)$ is the limit of $r_i(z,\lambda,N,y_{i0})$ as N tends to infinity, $i = 1,2$.

Example 3.1. Assume in addition to (3.1)–(3.4) that

$$\begin{aligned} f_{11}(z,y_1,u) &= L_1(z)y_1 + C_1(z)u, \\ f_{12}(z,y_2,v) &= L_2(z)y_2 + C_2(z)v, \\ f_{21}(z,y_1,u) &= I_1(z) + J_1(z)y_1 + K_1(z)u, \\ f_{22}(z,y_2,v) &= I_2(z) + J_2(z)y_2 + K_2(z)v, \end{aligned}$$

where C_i, I_i, J_i, K_i, and L_i are matrix functions of appropriate dimensions, with $J_i(z)$, $i = 1,2$, is nonsingular and the eigenvalues of $J_i(z) + I, i = 1,2$, are inside the unit circle for any $z \in R^m$. The latter condition ensures, in particular, the fulfillment of Assumption 2.1 (see Remark 4.4 below). Assume that $\Phi_1(z,y_1,u)$ is concave in (y_1,u) and $\Phi_2(z,y_2,v)$ is convex in (y_2,v) and also that U and V are convex compact sets. Similar to the example in [11, p. 892], it can be shown that

$$r_1(z,\lambda) = \max_u \left\{ \Phi_1\left(z,\psi_1(z,u),u\right) + \lambda^T f_{11}(z,\psi_1(z,u),u) \mid u \in U \right\}, \tag{3.9}$$

$$r_2(z,\lambda) = \min_v \left\{ \Phi_2\left(z,\psi_2(z,v),v\right) + \lambda^T f_{12}(z,\psi_2(z,v),v) \mid v \in V \right\}, \tag{3.10}$$

where $y_1 = \psi_1(z,u)$ and $y_2 = \psi_2(z,v)$ are the roots of the equations

$$f_{21}(z,y_1,u) = 0, \quad f_{22}(z,y_2,v) = 0.$$

That is,

$$\psi_1(z,u) \stackrel{\text{def}}{=} -\left(J_1(z)\right)^{-1} \left(I_1(z) + K_1(z)u\right), \tag{3.11}$$

$$\psi_2(z,v) \stackrel{\text{def}}{=} -\left(J_2(z)\right)^{-1} \left(I_2(z) + K_2(z)v\right). \tag{3.12}$$

It is easy to see that the LHJI equation (2.24) with

$$H(z,\lambda) = -\left(\lambda^T f_{10}(z) + r_1(z,\lambda) + r_2(z,\lambda)\right),$$

where $r_i(z,\lambda)$, $i = 1,2$, are defined by (3.9)–(3.12), coincides with the Isaacs equation for the "slow" differential game with the dynamics described by the equation

$$\dot{z} = f_{10}(z) + f_{11}(z,\psi_1(z,u),u) + f_{12}(z,\psi_2(z,v),v),$$

with the payoff function being

$$\int_0^T \Big[\Phi_1\left(z,\psi_1(z,u),u\right) + \Phi_2\left(z,\psi_2(z,v),v\right) \Big] dt.$$

Notice that this game is similar to a game which would be obtained if one uses an RTA to an SP control system

$$\begin{aligned}\dot{z}(t) &= f_1(z(t), y(t), u(t), v(t)),\\ \epsilon\dot{y}(t) &= f_2(z(t), y(t), u(t), v(t)),\\ &\qquad u(t) \in U,\ v(t) \in V,\ \forall t \in [0, T],\end{aligned}$$

evolving on a continuous time scale (see [12, 14, 21] for more details).

Example 3.2. Assume that $f_{1i}, f_{2i}, \Phi_i, i = 1, 2$, in (3.1)–(3.4) do not depend on z and $f_{10}(z) = Az,\ G(z) = C^T z$, where $A \in R^{m\times m}$ is a constant matrix and $C \in R^m$ is a constant vector.

The optimal values of (3.5), (3.6) do not depend on z in this case. Let us denote by $r_1(\lambda),\ r_2(\lambda)$ the limits of these values as N tends to infinity. By (3.8),

$$R^{up}(z, \lambda) = R^{lo}(z, \lambda) = R(z, \lambda) = \lambda^T Az + r_1(\lambda) + r_2(\lambda).$$

The LHJI equation (2.24) and the boundary conditions (2.25) in this case take the form

$$\begin{aligned}&\frac{\partial B(t,z)}{\partial t} + \left(\frac{\partial B(t,z)}{\partial z}\right)^T Az + r_1\left(\frac{\partial B(t,z)}{\partial z}\right) + r_2\left(\frac{\partial B(t,z)}{\partial z}\right) = 0,\\ &\qquad B(T, z) = C^T z.\end{aligned} \tag{3.13}$$

It can be verified via a direct substitution that the solution of (3.13) allows an explicit representation

$$B(t, z) = \lambda^T(t)z + \int_t^T \left[r_1\left(\lambda(\theta)\right) + r_2\left(\lambda(\theta)\right)\right] d\theta,$$

where $\lambda(t)$ is the solution of the system

$$\dot{\lambda}(\theta) = -A^T\lambda(\theta),\ \lambda(T) = C.$$

Notice that (3.13) is the Isaacs equation for the dynamic game with the payoff functional $C^T z(T)$ and the dynamics described by the equation

$$z(k + 1) = Az(k) + w_1(k) + w_2(k), \quad z(l) = z,$$

where the maximizer chooses $w(\theta) \in W_1$ and the minimizer chooses $w_2(\theta) \in W_2$. W_1 and W_2 are convex compact sets defined by their supporting functions,

$$\max_{w\in W_1}\{\lambda^T w\} = r_1(\lambda), \quad \min_{w\in W_2}\{\lambda^T w\} = r_2(\lambda).$$

4. Convergence of the upper and lower value functions of the SPDT game. Before presenting our main results in this paper, we introduce the following assumptions.

Assumption 4.1. All the functions used in the definitions of the SPDT and AFDT games in section 2 are continuous on $R^m \times R^n \times U \times V \times [0, T]$, and they also satisfy the local Lipschitz conditions in (z, y).

Assumption 4.2. Corresponding to any compact set $D \times P \subset R^m \times R^n$, there exists a compact set $D_1 \times P_1 \in R^m \times R^n$ such that if the initial values $(z, y) \in D \times P$,

then all the solutions of (2.1)–(2.2) obtained with different control functions satisfying (2.4) do not leave $D_1 \times P_1$.

Assumption 4.3. Corresponding to any compact set $D \times P \in R^m \times R^n$, there exists a compact set $P_1 \in R^n$ such that if $z \in D$ and the vector of initial values $y \in P$, then the solutions of (2.14) obtained with different control functions satisfying (2.15) do not leave P_1.

Remark 4.1. It is worthwhile to mention that Assumptions 4.2 and 4.3 might be replaced by some equivalent conditions with the help of the viability theory [3], which may provide another way to verify the correctness of these two assumptions.

Assumption 4.4. Assumption 2.1 is true and the convergence in (2.21) and (2.22) is uniform with respect to (z, λ, y) from any compact subset of $R^m \times R^m \times R^n$. That is, corresponding to any compact set $D \times \Lambda \times P \in R^m \times R^m \times R^n$, there exists a function $\gamma(N)$ tending to zero as N tends to infinity such that for any $(z, \lambda, y) \in D \times \Lambda \times P$,

$$\left| R^{up(lo)}(z, \lambda, N, y) - R^{up(lo)}(z, \lambda) \right| \leq \gamma(N). \tag{4.1}$$

Assumption 4.5. Corresponding to any compact set $D \times \Lambda \times P \subset R^m \times R^m \times R^n$, there exists a monotone function $\nu(\alpha)$ satisfying

$$\lim_{\alpha \to 0} \nu(\alpha) = 0 \tag{4.2}$$

such that

$$\left| R^{up(lo)}(z^1, \lambda) - R^{up(lo)}(z^2, \lambda) \right| \leq \nu(\|z^1 - z^2\|) \quad \forall z^i \in D,\ i = 1, 2. \tag{4.3}$$

Remark 4.2. Notice that we used (originally to establish our main result; see Theorem 4.1 below) a much stronger assumption (see Assumption 4.5 in [13]) which is similar to the corresponding assumption in [12] for continuous time differential games. It was noted by Artstein that this assumption can be replaced by a weaker one (similar to Assumption 4.5 above) in optimal control setting. The replacement of the assumption by a weaker one leads to changes in the proof (with respect to the one shown in [13]). We introduce these changes using some ideas suggested in [1].

Assumption 4.6. Corresponding to any compact set $[0, T] \times D_1 \times P_1 \subset [0, T] \times R^m \times R^n$, there exist a constant $L > 0$ and continuous functions $\omega_1(\alpha)$ and $\omega_2(\alpha)$ tending to zero as α tends to zero such that for any (l_1, z_1, y_1) and $(l_2, z_2, y_2) \in [0, T] \times D_1 \times P_1$,

$$\left| B_\epsilon^{up(lo)}(l_1, z_1, y_1) - B_\epsilon^{up(lo)}(l_2, z_2, y_2) \right| \leq \epsilon L |l_1 - l_2| + \omega_1(|z_1 - z_2|) + \omega_2(\epsilon)$$

with

$$B_\epsilon^{up(lo)}(N_\epsilon, z, y) = G(z) \quad \forall (z, y) \in R^m \times R^n. \tag{4.4}$$

Notice that Assumptions 4.4, 4.5, and, particularly, 4.6 imposed on the upper and lower value functions of SPDT games are hard to verify in a general case. In Remark 4.4 below we shall provide some sufficient conditions for these assumptions to be satisfied.

Now, we are ready to formulate the main result of this paper.

THEOREM 4.1. *Let Assumptions* 4.1–4.3 *and* 4.4–4.6 *be satisfied. Then there exists a function* $\mu(\epsilon)$ *tending to zero as* ϵ *tends to zero such that*

$$\left| B_\epsilon^{up(lo)}(l, z, y) - B^{up(lo)}(l\epsilon, z) \right| \leq \mu(\epsilon), \qquad l = 0, 1, \ldots, N_\epsilon - 1, \tag{4.5}$$

with the convergence being uniform on any compact subset of $R^m \times R^n$ *with* $l = 0, 1, \ldots, N_\epsilon - 1$.

From Theorem 4.1, the following corollary immediately follows.

COROLLARY 4.2. *If the AFDT game has value in the limit, that is,* (2.23) *is true and the LHJI equation* (2.24) *with* $H(z, \lambda) = -R(z, \lambda)$ *has the unique viscosity solution* $B(t, z)$*, then the SPDT game has value in the limit as well, and*

$$\left| B_\epsilon^{up}(l, z, y) - B(l\epsilon, z) \right| \le \mu(\epsilon), \qquad l = 0, 1, 2, \ldots, N_\epsilon - 1,$$

$$\left| B_\epsilon^{lo}(l, z, y) - B(l\epsilon, z) \right| \le \mu(\epsilon), \qquad l = 0, 1, 2, \ldots, N_\epsilon - 1.$$

Remark 4.3. An upper semicontinuous function $X(t, z)$ is called a *viscosity subsolution* of (2.24) if

$$-\frac{\partial x(\bar{t}, \bar{z})}{\partial t} + H\left(\bar{z}, \frac{\partial x(\bar{t}, \bar{z})}{\partial z}\right) \le 0$$

for any $(\bar{t}, \bar{z}) \in [0, T) \times R^m$ and for each function $x(t, z)$ which has continuous partial derivatives on $[0, T) \times R^m$ and satisfies the conditions $x(\bar{t}, \bar{z}) = X(\bar{t}, \bar{z})$ and $x(t, z) \ge X(t, z)$ in some neighborhood of $(\bar{t}, \bar{z})$.

A lower semicontinuous function $X(t, z)$ is called a *viscosity supersolution* of (2.24) if

$$-\frac{\partial x(\bar{t}, \bar{z})}{\partial t} + H\left(\bar{z}, \frac{\partial x(\bar{t}, \bar{z})}{\partial z}\right) \ge 0$$

for any $(\bar{t}, \bar{z}) \in [0, T) \times R^m$ and for each function $x(t, z)$ which has continuous partial derivatives on $[0, T) \times R^m$ and which satisfies the conditions $x(\bar{t}, \bar{z}) = X(\bar{t}, \bar{z})$ and $x(t, z) \le X(t, z)$ in some neighborhood of $(\bar{t}, \bar{z})$. A function $X(t, z)$ which is *both a viscosity sub- and a supersolution* is called a *viscosity solution* of (2.24).

As shown in a number of works (see [5] and the references therein), upper and lower values of a differential game coincide with the viscosity solutions of the corresponding Hamilton–Jacobi–Isaacs equations. In contrast to these results, Theorem 4.1 is of an asymptotic nature. We establish the convergence of the upper and lower values of the SP dynamic game to the viscosity solutions of LHJI equations which do not explicitly include a part of phase variables representing the fast motions.

Remark 4.4. Similar to Theorems 6.1 and 7.1 in [12], it can be established that Assumptions 4.4, 4.5, and 4.6 will be satisfied if the following assumption is true.

Assumption 4.7. Corresponding to any compact set $D \subset R^m$ there exists a sequence $\xi(k)$ tending to zero as k tends to infinity such that for any $z \in D$ the solutions of (2.14), $\tilde{y}_z^i(k)$, $i = 1, 2$, obtained with arbitrary initial conditions $\tilde{y}_z^i(0) = y^i$, $i = 1, 2$, and any control pairs $(\tilde{u}_k, \tilde{v}_k) \in U \times V$ satisfy the inequality

$$\|\tilde{y}_z^1(k) - \tilde{y}_z^2(k)\| \le \xi(k)\|y^1 - y^2\|.$$

It should be remarked that Assumption 4.7 is of stability type, which was used in [9]. The difference between the solutions of (2.14) induced by a difference in the initial values is decreasing to zero as time tends to infinity. Also, it is easy to see that this assumption is satisfied if the function $f_2(\cdot)$ in the right-hand side of (2.2) is presented in the form

$$f_2(z, y, u, v) = q(z, y) + p(u, v)$$

and $g(z,y) \stackrel{\text{def}}{=} y+q(z,y)$ defines a contractive operator. More precisely, if corresponding to any compact set $D \subset R^m$ there exists a constant $\alpha_D \in (0,1)$ such that for any y^1 and y^2 in R^n

$$\|g(z,y^2)-g(z,y^1)\| \le \alpha_D \|y^2-y^1\| \quad \forall z \in D,$$

then Assumption 4.7 is satisfied with $\xi(k)=\alpha_D^k$.

Proof of Theorem 4.1. Let $t_l = l\epsilon,\ l=0,1,\ldots,N_\epsilon = \lfloor \frac{T}{\epsilon} \rfloor$. For any $t \in [0,T]$, we define

$$\begin{aligned}
&\bar{B}_\epsilon^{up(lo)}(t,z,y) \\
&= \begin{cases} (1-\frac{t-t_l}{t_{l+1}-t_l})B_\epsilon^{up(lo)}(l,z,y) + \frac{(t-t_l)}{t_{l+1}-t_l}B_\epsilon^{up(lo)}(l+1,z,y), \\ \qquad\qquad\qquad\qquad t \in [t_l,t_{l+1}],\ l=0,1,\ldots,N_\epsilon-1; \\ B_\epsilon^{up(lo)}(N_\epsilon,z,y)=G(z), \qquad t \in [t_{N_\epsilon},T]. \end{cases}
\end{aligned} \tag{4.6}$$

Notice that $\bar{B}_\epsilon^{up(lo)}(t,z,y)$ is defined in $[0,T]\times R^m \times R^n$ and is continuous in t for $t \in [0,T]$.

Also note that if $l_1 \le l_2 < N_\epsilon - 1$, then for any (t_1,z_1,y_1) and $(t_2,z_2,y_2) \in [0,T]\times R^m \times R^n$ and $t_1 \in [t_{l_1},t_{l_1+1}]$, $t_2 \in [t_{l_2},t_{l_2+1}]$ one has

$$\begin{aligned}
&\left|\bar{B}_\epsilon^{up(lo)}(t_1,z_1,y_1)-\bar{B}_\epsilon^{up(lo)}(t_2,z_2,y_2)\right| \\
&\le \left|\bar{B}_\epsilon^{up(lo)}(t_1,z_1,y_1)-\bar{B}_\epsilon^{up(lo)}(t_2,z_1,y_1)\right| + \left|\bar{B}_\epsilon^{up(lo)}(t_2,z_1,y_1)-\bar{B}_\epsilon^{up(lo)}(t_2,z_2,y_2)\right|.
\end{aligned} \tag{4.7}$$

Using the expressions (4.6) and Assumption 4.6, one can evaluate the first and second terms in the right-hand side of (4.7) as follows:

$$\begin{aligned}
&\left|\bar{B}_\epsilon^{up(lo)}(t_1,z_1,y_1)-\bar{B}_\epsilon^{up(lo)}(t_2,z_1,y_1)\right| \\
&\le \left|\bar{B}_\epsilon^{up(lo)}(t_1,z_1,y_1)-\bar{B}_\epsilon^{up(lo)}(t_{l_1},z_1,y_1)\right| \\
&\qquad + \left|\bar{B}_\epsilon^{up(lo)}(t_{l_1},z_1,y_1)-\bar{B}_\epsilon^{up(lo)}(t_{l_2},z_1,y_1)\right| \\
&\qquad + \left|\bar{B}_\epsilon^{up(lo)}(t_{l_2},z_1,y_1)-\bar{B}_\epsilon^{up(lo)}(t_2,z_1,y_1)\right| \\
&\le \frac{t_1-t_{l_1}}{t_{l_1+1}-t_{l_1}}\left|B_\epsilon^{up(lo)}(l_1+1,z_1,y_1)-B_\epsilon^{up(lo)}(l_1,z_1,y_1)\right| \\
&\qquad + \left|B_\epsilon^{up(lo)}(l_1,z_1,y_1)-B_\epsilon^{up(lo)}(l_2,z_1,y_1)\right| \\
&\qquad + \frac{t_2-t_{l_2}}{t_{l_2+1}-t_{l_2}}\left|B_\epsilon^{up(lo)}(l_2+1,z_1,y_1)-B_\epsilon^{up(lo)}(l_2,z_1,y_1)\right| \\
&\le 4\epsilon L + 3\omega_2(\epsilon) + 2\epsilon L|t_1-t_2|
\end{aligned} \tag{4.8}$$

and

$$\begin{aligned}
&\left|\bar{B}_\epsilon^{up(lo)}(t_2,z_1,y_1)-\bar{B}_\epsilon^{up(lo)}(t_2,z_2,y_2)\right| \\
&\le \left|\bar{B}_\epsilon^{up(lo)}(t_2,z_1,y_1)-\bar{B}_\epsilon^{up(lo)}(t_{l_2},z_1,y_1)\right|
\end{aligned}$$

$$
\begin{aligned}
&\quad + \left|\bar{B}_\epsilon^{up(lo)}(t_{l_2}, z_1, y_1) - \bar{B}_\epsilon^{up(lo)}(t_2, z_2, y_2)\right| \\
&\le \frac{t_2 - t_{l_2}}{t_{l_2+1} - t_{l_2}} \left|B_\epsilon^{up(lo)}(l_2+1, z_1, y_1) - B_\epsilon^{up(lo)}(l_2, z_1, y_1)\right| \\
&\quad + \left|\bar{B}_\epsilon^{up(lo)}(t_{l_2}, z_1, y_1) - \bar{B}_\epsilon^{up(lo)}(t_{l_2}, z_2, y_2)\right| \\
&\quad + \left|\bar{B}_\epsilon^{up(lo)}(t_{l_2}, z_2, y_2) - \bar{B}_\epsilon^{up(lo)}(t_2, z_2, y_2)\right| \\
&\le \frac{t_2 - t_{l_2}}{t_{l_2+1} - t_{l_2}} \left|B_\epsilon^{up(lo)}(l_2+1, z_1, y_1) - B_\epsilon^{up(lo)}(l_2, z_1, y_1)\right| \\
&\quad + \left|B_\epsilon^{up(lo)}(l_2, z_1, y_1) - B_\epsilon^{up(lo)}(l_2, z_2, y_2)\right| \\
&\quad + \frac{t_2 - t_{l_2}}{t_{l_2+1} - t_{l_2}} \left|B_\epsilon^{up(lo)}(l_2+1, z_2, y_2) - B_\epsilon^{up(lo)}(l_2, z_2, y_2)\right| \\
&\le 2\epsilon L + \omega_1(|z_1 - z_2|) + 3\omega_2(\epsilon). \qquad (4.9)
\end{aligned}
$$

Substituting (4.8) and (4.9) back into (4.7), one obtains

$$
\begin{aligned}
&\left|\bar{B}_\epsilon^{up(lo)}(t_1, z_1, y_1) - \bar{B}_\epsilon^{up(lo)}(t_2, z_2, y_2)\right| \\
&\qquad \le 6\epsilon L + \omega_1(|z_1 - z_2|) + 6\omega_2(\epsilon) + \epsilon L|t_1 - t_2|. \qquad (4.10)
\end{aligned}
$$

Let us introduce the notation

$$
\bar{B}_\epsilon^{up}(t, z, 0) \stackrel{\text{def}}{=} X_\epsilon(t, z).
$$

By Assumption 4.6, if (t, z, y) belongs to a compact set $[0, T] \times D_1 \times P_1$, then

$$
\begin{aligned}
&\left|\bar{B}_\epsilon^{up}(t, z, y) - X_\epsilon(t, z)\right| \le \omega_2(\epsilon) \qquad (4.11) \\
&\quad \Longrightarrow |B_\epsilon^{up}(l, z, y) - X_\epsilon(l\epsilon, z)| \le \omega_2(\epsilon) \quad \forall l = 0, 1, \ldots, N_\epsilon - 1.
\end{aligned}
$$

Hence, to prove (4.5) it is sufficient to show that

$$
\lim_{\epsilon \to 0} X_\epsilon(t, z) = B^{up}(t, z), \qquad (4.12)
$$

where the convergence is uniform with respect to (t, z) from any compact subset of $[0, T] \times R^m$. For the sake of brevity we shall refer to this sort of convergence as *U-convergence* and the corresponding limits will be called *U-limits*.

From (4.10) and (4.11) it follows that for any $(t_i, z_i) \in D_1 \times P_1, i = 1, 2,$

$$
|X_\epsilon(t_1, z_1) - X_\epsilon(t_2, z_2)| \le 6\epsilon L + \omega_1(|z_1 - z_2|) + 6\omega_2(\epsilon) + \epsilon L|t_1 - t_2|. \qquad (4.13)
$$

Before completing the proof of Theorem 4.1, let us recall the following lemma needed in our derivation.

LEMMA 4.3 (see [12]). *Given any sequence ϵ_i tending to zero, one can find a subsequence $\epsilon_{i_j} = \epsilon_j$ of this sequence such that there exists the U-limit*

$$
\lim_{\epsilon_j \to 0} X_{\epsilon_j}(t, z) \stackrel{\text{def}}{=} X(t, z). \qquad (4.14)
$$

Let us show that any function obtained as U-limit (4.14) coincides with $B^{up}(t, z)$. Notice that, by (4.13), any such function $X(t, z)$ is continuous on $[0, T] \times R^m$ and, by (4.4) and (4.11), it satisfies the condition

$$
X(T, z) = G(z) \quad \forall z \in R^m.
$$

Thus, to show that it coincides with $B^{up}(t,z)$, it is enough to show that it is a viscosity solution of (2.24) with $H(z,\lambda) = -R^{up}(t,z)$.

Let $n(\epsilon)$ be a function of ϵ taking positive integer values such that

$$n(\epsilon) \longrightarrow \infty \qquad \text{and} \qquad \epsilon n(\epsilon) \longrightarrow 0 \qquad \text{as } \epsilon \to 0. \tag{4.15}$$

For any $\bar{t} \in [0,T)$ and any $\epsilon > 0$, there exists $l(\epsilon) > 0$ such that $\bar{t} \in [l(\epsilon)\epsilon, (l(\epsilon)+1)\epsilon]$, where $l(\epsilon) = \lfloor \frac{\bar{t}}{\epsilon} \rfloor$.

Let $(z(k), y(k))$ be the solution of (2.1), (2.2) on the interval $[\epsilon l(\epsilon), \epsilon l(\epsilon) + \epsilon n(\epsilon)]$ obtained with the initial conditions $(z(l(\epsilon)), y(l(\epsilon))) = (\bar{z}, \bar{y})$ and with the use of the control $(u_k, v_k) \in U \times V$, $k = l(\epsilon), l(\epsilon)+1, \ldots, l(\epsilon)+n(\epsilon)-1$.

By (4.11),

$$\bar{B}^{up}_{\epsilon}(\bar{t}, \bar{z}, \bar{y}) = X_{\epsilon}(\bar{t}, \bar{z}) + O(\omega_2(\epsilon)). \tag{4.16}$$

Using representation (2.11) and the estimate (4.10), one can obtain

$$\begin{aligned}
\bar{B}^{up}_{\epsilon}(\bar{t}, \bar{z}, \bar{y}) &= B^{up}_{\epsilon}(l(\epsilon), \bar{z}, \bar{y}) + O(\epsilon) \\
&= \inf_{v_{l(\epsilon)}} \sup_{u_{l(\epsilon)}} \cdots \inf_{v_{l(\epsilon)+n(\epsilon)-1}} \sup_{u_{l(\epsilon)+n(\epsilon)-1}} \left\{ \epsilon \sum_{k=l(\epsilon)}^{l(\epsilon)+n(\epsilon)-1} \Phi(z(k), y(k), u_k, v_k) \right. \\
&\qquad \left. + B^{up}_{\epsilon}(l(\epsilon)+n(\epsilon), z(l(\epsilon)+n(\epsilon)), y(l(\epsilon)+n(\epsilon))) \right\} + O(\epsilon) \\
&= \inf_{v_{l(\epsilon)}} \sup_{u_{l(\epsilon)}} \cdots \inf_{v_{l(\epsilon)+n(\epsilon)-1}} \sup_{u_{l(\epsilon)+n(\epsilon)-1}} \left\{ \epsilon \sum_{k=l(\epsilon)}^{l(\epsilon)+n(\epsilon)-1} \Phi(z(k), y(k), u_k, v_k) \right. \\
&\qquad \left. + \bar{B}^{up}_{\epsilon}(t_{l(\epsilon)+n(\epsilon)}, z(l(\epsilon)+n(\epsilon)), y(l(\epsilon)+n(\epsilon))) \right\} + O(\epsilon) \\
&= \inf_{v_{l(\epsilon)}} \sup_{u_{l(\epsilon)}} \cdots \inf_{v_{l(\epsilon)+n(\epsilon)-1}} \sup_{u_{l(\epsilon)+n(\epsilon)-1}} \left\{ \epsilon \sum_{k=l(\epsilon)}^{l(\epsilon)+n(\epsilon)-1} \Phi(z(k), y(k), u_k, v_k) \right. \\
&\qquad \left. + X_{\epsilon}(t_{l(\epsilon)+n(\epsilon)}, z(l(\epsilon)+n(\epsilon))) \right\} + O(\epsilon) + O(\omega_2(\epsilon)).
\end{aligned} \tag{4.17}$$

Let D and P be compact sets such that $(\bar{t}, \bar{z}, \bar{y}) \in [0,T) \times D \times P$. Then, by Assumption 4.2, $(z(l(\epsilon)+n(\epsilon)), y(l(\epsilon)+n(\epsilon))) \in D_1 \times P_1$, where D_1 and P_1 are compact sets in R^m and R^n. Since the convergence in (4.14) is uniform with respect to (t,z) from any compact subset of $[0,T] \times R^m$, there exists a function $\tilde{\nu}(\epsilon_j)$,

$$\lim_{\epsilon_j \to 0} \tilde{\nu}(\epsilon_j) = 0, \tag{4.18}$$

such that

$$\left| X_{\epsilon_j}(t,z) - X(t,z) \right| \le \tilde{\nu}(\epsilon_j) \quad \forall (t,z) \in [0,T] \times D_1.$$

Using this and (4.11), one obtains from (4.6) with (4.17)

$$X(\bar{t},\bar{z}) = \inf_{v_{l(\epsilon_j)}} \sup_{u_{l(\epsilon_j)}} \dots \inf_{v_{l(\epsilon_j)+n(\epsilon_j)-1}} \sup_{u_{l(\epsilon_j)+n(\epsilon_j)-1}} \left\{ \epsilon_j \sum_{k=l(\epsilon_j)}^{l(\epsilon_j)+n(\epsilon_j)-1} \Phi(z(k),y(k),u_k,v_k) \right.$$

$$\left. + X(t_{l(\epsilon_j)+n(\epsilon_j)},\ z(l(\epsilon_j)+n(\epsilon_j))) \right\} + O(\tilde{\mu}(\epsilon_j)), \tag{4.19}$$

where

$$\tilde{\mu}(\epsilon_j) = \max\{\omega_2(\epsilon_j),\ \tilde{\nu}(\epsilon_j),\epsilon_j\}. \tag{4.20}$$

Now let $x(t,z)$ have continuous partial derivatives and satisfy the conditions $x(\bar{t},\bar{z}) = X(\bar{t},\bar{z})$ and $x(t,z) \geq X(t,z)$ for (t,z) in some neighborhood of $(\bar{t},\bar{z})$. From (4.19) it then follows that

$$x(\bar{t},\bar{z}) \leq \inf_{v_{l(\epsilon_j)}} \sup_{u_{l(\epsilon_j)}} \dots \inf_{v_{l(\epsilon_j)+n(\epsilon_j)-1}} \sup_{u_{l(\epsilon_j)+n(\epsilon_j)-1}} \left\{ \epsilon_j \sum_{k=l(\epsilon_j)}^{l(\epsilon_j)+n(\epsilon_j)-1} \Phi(z(k),y(k),u_k,v_k) \right.$$

$$\left. + x(t_{l(\epsilon_j)+n(\epsilon_j)}, z(l(\epsilon_j)+n(\epsilon_j))) \right\} + O(\tilde{\mu}(\epsilon_j)). \tag{4.21}$$

By definition

$$z\big(l(\epsilon_j)+n(\epsilon_j)\big) = \bar{z} + \epsilon_j \sum_{k=l(\epsilon_j)}^{l(\epsilon_j)+n(\epsilon_j)-1} f_1(z(k),y(k),u_k,v_k) \tag{4.22}$$

and

$$t_{l(\epsilon_j)+n(\epsilon_j)} = \big(l(\epsilon_j)+n(\epsilon_j)\big)\epsilon_j = \bar{t} + O(\epsilon_j) + \epsilon_j n(\epsilon_j).$$

Since the function $f_1(\cdot)$ is continuous and its arguments belong to compact sets, the second term in the right-hand side of (4.22) is of the order $O(\epsilon_j n(\epsilon_j))$. Thus, substituting (4.22) into (4.21) and using Taylor's expansion of $x(t_{l(\epsilon_j)+n(\epsilon_j)}, z(l(\epsilon_j)+n(\epsilon_j)))$ at $(\bar{t},\bar{z})$, one obtains

$$\frac{\partial x(\bar{t},\bar{z})}{\partial t} + \frac{1}{\epsilon_j n(\epsilon_j)} \inf_{v_{l(\epsilon_j)}} \sup_{u_{l(\epsilon_j)}}$$

$$\dots \inf_{v_{l(\epsilon_j)+n(\epsilon_j)-1}} \sup_{u_{l(\epsilon_j)+n(\epsilon_j)-1}} \left\{ \epsilon_j \sum_{k=l(\epsilon_j)}^{l(\epsilon_j)+n(\epsilon_j)-1} \left[\Phi(z(k),y(k),u_k,v_k) \right.\right.$$

$$\left.\left. + \left(\frac{\partial x(\bar{t},\bar{z})}{\partial z}\right)^T f_1(z(k),y(k),u_k,v_k) \right] \right\} + \frac{O(\tilde{\mu}(\epsilon_j))}{\epsilon_j n(\epsilon_j)} + \frac{o(\epsilon_j n(\epsilon_j))}{\epsilon_j n(\epsilon_j)} \geq 0. \tag{4.23}$$

To proceed with the proof, one needs to establish the following.

LEMMA 4.4. *Let $K(\epsilon)$ be a function of ϵ having integer values and satisfying the conditions*

$$K(\epsilon) \to \infty \qquad as\ \epsilon \to 0, \quad K(\epsilon) \le \frac{1}{2\ln(1+C)} \ln \frac{1}{\epsilon}, \tag{4.24}$$

where C is a Lipschitz constant of the function f_2 on the compact set containing the trajectories of (2.1), (2.2), *and* (2.14) *(see Assumptions* 4.1–4.3*). Let $(z(k), y(k))$, $k = l, l+1, \dots, l+K(\epsilon)$, be the solutions of* (2.1), (2.2) *obtained with some controls (u_k, v_k), $k = l, l+1, \dots, l+K(\epsilon)-1$, and with the initial values* (2.3). *Let $\tilde{y}_{\bar{z}}(k)$, $k = l, l+1, \dots, l+K(\epsilon)$, be the solution of the system*

$$\tilde{y}_{\bar{z}}(k+1) = \tilde{y}_{\bar{z}}(k) + f_2(\bar{z}, \tilde{y}_{\bar{z}}, u_k, v_k) \tag{4.25}$$

obtained with the same sequence of controls as above, (u_k, v_k), $k = l, l+1, \dots, l+K(\epsilon)-1$, and with the initial conditions

$$\tilde{y}_{\bar{z}}(l) = y, \tag{4.26}$$

the same as in (2.3). *Assume that $\bar{z}$ in* (4.25) *satisfies the inequality*

$$\|\bar{z} - z(k)\| \le M\epsilon K(\epsilon) \quad \forall k = l, l+1, \dots, l+K(\epsilon)-1, \tag{4.27}$$

with some constant M. Then

$$\max_{k=l,l+1,\dots,l+K(\epsilon)} \|y(k) - \tilde{y}_{\bar{z}}(k)\| \le \gamma_1(\epsilon), \tag{4.28}$$

where

$$\gamma_1(\epsilon) \stackrel{\text{def}}{=} M\epsilon K(\epsilon)(1+C)^{K(\epsilon)+1} \to 0 \qquad as\ \epsilon \to 0. \tag{4.29}$$

The fact that $\gamma_1(\epsilon)$ tends to zero follows from the second inequality in (4.24).

Proof. Subtracting (4.25) from (2.2) and using Lipschitz conditions, one obtains

$$\begin{aligned}
&\|y_z(k+1) - \tilde{y}_{\bar{z}}(k+1)\| \\
&\le \|y_z(k) - \tilde{y}_{\bar{z}}(k)\| + C\left(\|z(k) - \bar{z}\| + \|y_z(k) - \tilde{y}_{\bar{z}}(k)\|\right) \\
&\le CM\epsilon K(\epsilon) + (1+C)\|y_z(k) - \tilde{y}_{\bar{z}}(k)\|,
\end{aligned}$$

which implies

$$\begin{aligned}
&\|y(k) - \tilde{y}_{\bar{z}}(k)\| \\
&\le M\epsilon K(\epsilon)(1+C)^{k+1-l} \\
&\le M\epsilon K(\epsilon)(1+C)^{K(\epsilon)+1} \quad \forall k = l, l+1, \dots, l+K(\epsilon).
\end{aligned}$$

This justifies (4.28). ∎

COROLLARY 4.5. *Let $\phi(z, y, u, v)$ satisfy Lipschitz conditions in z and y. Then*

$$\begin{aligned}
&\left| \frac{1}{K(\epsilon)} \sum_{k=l}^{l+K(\epsilon)-1} \phi(z(k), y(k), u_k, v_k) - \frac{1}{K(\epsilon)} \sum_{k=l}^{l+K(\epsilon)-1} \phi(\bar{z}, \tilde{y}_{\bar{z}}(k), u_k, v_k) \right| \\
&\le C_1 K(\epsilon)\left(M\epsilon K(\epsilon) + \gamma_1(\epsilon)\right) \stackrel{\text{def}}{=} \gamma_2(\epsilon),
\end{aligned} \tag{4.30}$$

where C_1 is a Lipschitz constant of ϕ.

Notice that, by (4.29), (4.24),

$$\gamma_2(\epsilon) \to 0 \qquad as\ \epsilon \to 0. \tag{4.31}$$

Returning to the proof of the main theorem let us define

$$n(\epsilon_j) = r(\epsilon_j)K(\epsilon_j) + K(\epsilon_j) \tag{4.32}$$

with

$$r(\epsilon_j) = \left\lfloor \frac{(\tilde{\mu}(\epsilon_j))^{1/2}}{\epsilon_j K(\epsilon_j)} \right\rfloor. \tag{4.33}$$

Notice that

$$\epsilon_j n(\epsilon_j) \to 0 \qquad \text{as}\ \ \epsilon_j \to 0. \tag{4.34}$$

Denoting for the brevity

$$\Phi(z,y,u,v) - \left(\frac{\partial x(\bar{t},\bar{z})}{\partial z}\right)^T f_1(z,y,u,v) \stackrel{\text{def}}{=} \phi(z,y,u,v), \tag{4.35}$$

let us rewrite the second term in (4.23) as follows:

$$\begin{aligned}
&\frac{1}{n(\epsilon_j)} \underbrace{\inf\sup\cdots\inf\sup}_{n(\epsilon_j)} \sum_{k=l(\epsilon_j)}^{l(\epsilon_j)+n(\epsilon_j)-1} \phi(z(k),y(k),u_k,v_k) \\
&= \frac{1}{r(\epsilon_j)+1} \Bigg[\underbrace{\inf\sup\cdots\inf\sup}_{K(\epsilon_j)} \frac{1}{K(\epsilon_j)} \left\{ \sum_{k=l(\epsilon_j)}^{l(\epsilon_j)+K(\epsilon_j)-1} \phi(z(k),y(k),u_k,v_k) \right. \\
&+ \underbrace{\inf\sup\cdots\inf\sup}_{K(\epsilon_j)} \left\{ \frac{1}{K(\epsilon_j)} \sum_{k=l(\epsilon_j)+K(\epsilon_j)}^{l(\epsilon_j)+2K(\epsilon_j)-1} \phi(z(k),y(k),u_k,v_k) + \cdots \right. \\
&+ \underbrace{\inf\sup\cdots\inf\sup}_{K(\epsilon_j)} \left\{ \frac{1}{K(\epsilon_j)} \sum_{k=l(\epsilon_j)+(r(\epsilon_j)-1)K(\epsilon_j)}^{l(\epsilon_j)+r(\epsilon_j)K(\epsilon_j)-1} \phi(z(k),y(k),u_k,v_k) \right. \\
&\left. \left. \left. + \underbrace{\inf\sup\cdots\inf\sup}_{K(\epsilon_j)} \left\{ \frac{1}{K(\epsilon_j)} \sum_{k=l(\epsilon_j)+r(\epsilon_j)K(\epsilon_j)}^{l(\epsilon_j)+(r(\epsilon_j)+1)K(\epsilon_j)-1} \phi(z(k),y(k),u_k,v_k) \underbrace{\Big\}\Big\}\Big\}\cdots\Big\}}_{r(\epsilon_j)+1} \right] \right.
\end{aligned} \tag{4.36}$$

where the inf sups in the right-hand side of (4.36) are taken over the controls included in the corresponding sums. Let us consider the last sum in the right-hand side of (4.36). Notice that for any controls

$$\|z(k) - \bar{z}_r\| \le M\epsilon_j K(\epsilon_j) \quad \forall k = l(\epsilon_j) + r(\epsilon_j)K(\epsilon_j) + 1, \ldots, l(\epsilon_j) + n(\epsilon_j), \tag{4.37}$$

where M is some constant and

$$\bar{z}_r \stackrel{\text{def}}{=} z(l(\epsilon_j) + r(\epsilon_j)K(\epsilon_j)).$$

By Lemma 4.4,

$$\|y(k) - \tilde{y}_{\bar{z}_r}(k)\| \le \gamma_1(\epsilon) \quad \forall k = l(\epsilon_j) + r(\epsilon_j)K(\epsilon_j) + 1, \dots, l(\epsilon_j) + (r(\epsilon_j) + 1)K(\epsilon_j), \tag{4.38}$$

where $\tilde{y}_{\bar{z}_r}(k)$ is the solution of system (4.25) with $\bar{z} = \bar{z}_r$ and with the initial conditions

$$\tilde{y}_{\bar{z}_r}(l(\epsilon_j) + r(\epsilon_j)K(\epsilon_j)) = y(l(\epsilon_j) + r(\epsilon_j)K(\epsilon_j)) \stackrel{\text{def}}{=} \bar{y}_r. \tag{4.39}$$

Estimates (4.37), (4.38) imply the inequality (see Corollary 4.5)

$$\begin{aligned} &\left| \frac{1}{K(\epsilon_j)} \sum_{k=l(\epsilon_j)+r(\epsilon_j)K(\epsilon_j)}^{l(\epsilon_j)+(r(\epsilon_j)+1)K(\epsilon_j)-1} \phi(z(k), y(k), u_k, v_k) \right. \\ &\left. - \frac{1}{K(\epsilon_j)} \sum_{k=l(\epsilon_j)+r(\epsilon_j)K(\epsilon_j)}^{l(\epsilon_j)+(r(\epsilon_j)+1)K(\epsilon_j)-1} \phi(\bar{z}_r, \tilde{y}_{\bar{z}_r}(k), u_k, v_k) \right| \\ &\le \gamma_2(\epsilon). \end{aligned} \tag{4.40}$$

This inequality is satisfied uniformly with respect to (u_k, v_k). Hence, the application of the operations inf sups do not violate the inequality. Thus, the last $r(\epsilon_j) + 1$ term in the right-hand side of (4.38) can be presented in the form

$$R^{up}\left(\bar{z}_r, \frac{\partial x(\bar{t}, \bar{z})}{\partial z}, K(\epsilon_j), \bar{y}_r\right) + \theta_{r(\epsilon_j)+1}, \tag{4.41}$$

with

$$\left|\theta_{r(\epsilon_j)+1}\right| \le \gamma_2(\epsilon_j). \tag{4.42}$$

By Assumption 4.4,

$$R^{up}\left(\bar{z}_r, \frac{\partial x(\bar{t}, \bar{z})}{\partial z}, K(\epsilon_j), \bar{y}_r\right) = R^{up}\left(\bar{z}_r, \frac{\partial x(\bar{t}, \bar{z})}{\partial z}\right) + \Gamma_{r(\epsilon_j)}, \tag{4.43}$$

where

$$\left|\Gamma_{r(\epsilon_j)}\right| \le \gamma(K(\epsilon_j)). \tag{4.44}$$

Also, by Assumption 4.5,

$$R^{up}\left(\bar{z}_r, \frac{\partial x(\bar{t}, \bar{z})}{\partial z}\right) = R^{up}\left(\bar{z}, \frac{\partial x(\bar{t}, \bar{z})}{\partial z}\right) + \Delta_{r(\epsilon_j)}, \tag{4.45}$$

$$\left|\Delta_{r(\epsilon_j)}\right| \le \nu\left(M\epsilon_j n(\epsilon_j)\right), \tag{4.46}$$

where the last estimate is obtained on the basis of the fact that

$$|\bar{z}_r - \bar{z}| \le M\epsilon_j r(\epsilon_j)K(\epsilon_j) \le M\epsilon_j n(\epsilon_j). \tag{4.47}$$

Summing up, one can conclude that the $r(\epsilon_j)+1$ term in (4.36) is equal to

$$R^{up}\left(\bar{z}, \frac{\partial x(\bar{t},\bar{z})}{\partial z}, K(\epsilon_j)\right) + \eta_{r(\epsilon_j)+1}, \tag{4.48}$$

with

$$\left|\eta_{r(\epsilon_j)+1}\right| \le \gamma_2(\epsilon_j) + \gamma(K(\epsilon_j)) + \nu\left(M\epsilon_j n(\epsilon_j)\right). \tag{4.49}$$

As follows from (4.48), the $r(\epsilon_j)+1$ term in (4.36) does not depend (within the proximity given in (4.49)) on the values of controls (u_k, v_k) for $k < l(\epsilon_j)+r(\epsilon_j)K(\epsilon_j)$. This allows us to obtain a similar estimate for the term $r(\epsilon_j)$ (next to the last) in the right-hand side of (4.36). Proceeding in the same way $r(\epsilon_j)+1$ times, one can obtain that the second term in (4.23) is equal to

$$R^{up}\left(\bar{z}, \frac{\partial x(\bar{t},\bar{z})}{\partial z}\right) + \eta(\epsilon_j), \tag{4.50}$$

$$|\eta(\epsilon_j)| \le \gamma_2(\epsilon_j) + \gamma(K(\epsilon_j)) + \nu\left(M\epsilon_j n(\epsilon_j)\right), \tag{4.51}$$

which after the substitution to (4.23) gives

$$\begin{aligned}\frac{\partial x(\bar{t},\bar{z})}{\partial t} + R^{up}\left(\bar{z}, \frac{\partial x(\bar{t},\bar{z})}{\partial z}, K(\epsilon_j)\right) + O\left(\gamma_2(\epsilon_j)\right) + O\left(\gamma(K(\epsilon_j))\right) \\ + O\left(\nu\left(M\epsilon_j n(\epsilon_j)\right)\right) + \frac{O(\tilde{\mu}(\epsilon_j))}{\epsilon_j n(\epsilon_j)} + \frac{o(\epsilon_j n(\epsilon_j))}{\epsilon_j n(\epsilon_j)} \ge 0.\end{aligned} \tag{4.52}$$

Notice that, by (4.33),

$$\frac{\tilde{\mu}(\epsilon_j)}{\epsilon_j n(\epsilon_j)} \le \frac{\tilde{\mu}(\epsilon_j)}{\epsilon_j} \frac{1}{\frac{(\tilde{\mu}(\epsilon_j))^{1/2}}{\epsilon_j}} = \left(\tilde{\mu}(\epsilon_j)\right)^{1/2}.$$

Taking into account this as well as (4.23) and passing to the limit as ϵ_j tends to zero, one obtains

$$\frac{\partial x(\bar{t},\bar{z})}{\partial t} + R^{up}\left(\bar{z}, \frac{\partial x(\bar{t},\bar{z})}{\partial z}\right) \ge 0 \Rightarrow -\frac{\partial x(\bar{t},\bar{z})}{\partial t} + H\left(\bar{z}, \frac{\partial x(\bar{t},\bar{z})}{\partial z}\right) \le 0.$$

This establishes that $X(t,z)$ is a viscosity subsolution of (2.24) on $[0,T)\times R^m$. Similarly, taking $x(t,z)$ having continuous partial derivatives and satisfying the conditions $x(\bar{t},\bar{z}) = X(\bar{t},\bar{z})$ and $x(t,z) \le X(t,z)$ in some neighborhood of $(\bar{t},\bar{z}) \in [0,T)\times R^m$, one can obtain that

$$-\frac{\partial x(\bar{t},\bar{z})}{\partial t} + H\left(\bar{z}, \frac{\partial x(\bar{t},\bar{z})}{\partial z}\right) \ge 0,$$

which means that $X(t,z)$ is a viscosity supersolution of (2.24) on $[0,T)\times R^m$. Thus, $X(t,z)$ is a viscosity solution (2.24) on $[0,T)\times R^m$ and, consequently, it coincides with $B^{up}(t,z)$.

This proves that $\bar{B}^{up}_{\epsilon}(t,z,y)$ U-converges (as ϵ tends to zero) to $B^{up}(t,z)$ (since, otherwise, by Lemma 4.3, one would be able to choose a subsequence ϵ_j tending to zero such that the U-limit (4.14) does not coincide with $B^{up}(t,z)$).

Similarly, it is established that $\bar{B}^{lo}_{\epsilon}(t,z,y)$ U-converges to $\bar{B}^{lo}(t,z)$ as ϵ tends to zero. □

Remark 4.5. It should be noted in Theorem 4.1 that, from [6, 20], the continuous viscosity solution $B^{up}(t,z)$, $(B^{lo}(t,z))$ of (2.24) with $H(z,\lambda) = -R^{up}(z,\lambda)(-R^{lo}(z,\lambda))$ satisfying boundary conditions (2.25) is unique. Notice that the Hamiltonians of the LHJI equations mentioned in the theorems are continuous under the assumptions made.

Remark 4.6. The conditions that U and V are compact sets introduced in the definition of the SP game can be replaced by the coercivity-type condition. Namely, one can assume that corresponding to any compact set $D \times \Lambda \times P \in R^m \times R^m \times R^n$, there exists a constant M such that for any $(z,\lambda,y) \subset D \times \Lambda \times P$, $l = 0,1,\dots,N_\epsilon$ and $N > 0$, the upper and lower values (2.8), (2.9) and (2.17), (2.18) are not changed with the replacement of U and V by

$$U_M = \{u | u \in U, \| u \| \leq M\}, \quad V_M = \{v | v \in V, \| v \| \leq M\}.$$

Remark 4.7. Note that since the closed-loop strategies are considered in this paper, in our future studies the connections between SPDT games and positional differential games would be worth further investigation, in which feedback strategies would be used (see, for example, [4]). We conjecture that our results (and approach) will be useful for the readers working in the fields of (SP) positional differential games.

5. Asymptotics of the AFDT game: Verification of assumptions. We have the following result.

THEOREM 5.1. *Assumptions* 4.1*,* 4.3*, and* 4.7 *imply Assumption* 4.4.

The proof of Theorem 5.1 is based on the following.

LEMMA 5.2. *Corresponding to any compact set $D \times \Lambda \times P \subset R^m \times R^m \times R^n$, there exists a constant L and a monotone function $\gamma_1(N)$ tending to zero as N tends to infinity such that for any $(z,\lambda,y) \in D \times \Lambda \times P$ and any $y^i \in P, i = 1,2,$*

$$\left| R^{up(lo)}(z,\lambda,N,y^1) - R^{up(lo)}(z,\lambda,N,y^2) \right| \leq \gamma_1(N), \qquad N = 0,1,2,\dots; \tag{5.1}$$

$$\left| R^{up(lo)}(z,\lambda,lK,y) - R^{up(lo)}(z,\lambda,K,y) \right| \leq \gamma_1(K) \quad \forall K = 0,1,2,\dots,\ l = 0,1,2,\dots; \tag{5.2}$$

$$\left| R^{up(lo)}(z,\lambda,N_1,y) - R^{up(lo)}(z,\lambda,N_2,y) \right| \leq L\left(\frac{1}{N_1} + \frac{N_2 - N_1}{N_2}\right), \qquad N_2 \geq N_1 \geq 1. \tag{5.3}$$

Proof of Theorem 5.1. We need to show that there exist the limits

$$\lim_{K \to \infty} R^{up(lo)}(z,\lambda,K,y) \stackrel{\text{def}}{=} R^{up(lo)}(z,\lambda). \tag{5.4}$$

By (5.2) and (5.3), for any integers $K_1 \geq K_0$, $K_2 \geq K_0$, and $K_0 > 0$,

$$\begin{aligned}
&\left| R^{up(lo)}(z,\lambda,K_2,y) - R^{up(lo)}(z,\lambda,K_1,y) \right| \\
&\leq \left| R^{up(lo)}(z,\lambda,K_2,y) - R^{up(lo)}\left(z,\lambda,\left\lfloor \frac{K_2}{K_0} \right\rfloor K_0, y\right) \right| \\
&\quad + \left| R^{up(lo)}\left(z,\lambda,\left\lfloor \frac{K_2}{K_0} \right\rfloor K_0, y\right) - R^{up(lo)}(z,\lambda,K_0,y) \right|
\end{aligned}$$

$$
\begin{aligned}
&+\left|R^{up(lo)}(z,\lambda,K_0,y)-R^{up(lo)}\left(z,\lambda,\left\lfloor\frac{K_1}{K_0}\right\rfloor K_0,y\right)\right|\\
&+\left|R^{up(lo)}\left(z,\lambda,\left\lfloor\frac{K_1}{K_0}\right\rfloor K_0,y\right)-R^{up(lo)}(z,\lambda,K_1,y)\right|\\
&\le L\left(\frac{1}{\lfloor\frac{K_2}{K_0}\rfloor K_0}+\frac{(K_2-\lfloor\frac{K_2}{K_0}\rfloor K_0)}{K_2}\right)\\
&\quad+2\gamma_1(K_0)+L\left(\frac{1}{\lfloor\frac{K_1}{K_0}\rfloor K_0}+\frac{(K_1-\lfloor\frac{K_1}{K_0}\rfloor K_0)}{K_1}\right)\\
&\le 2\gamma_1(K_0)+\frac{2L}{K_0}+\frac{LK_0}{K_1}+\frac{LK_0}{K_2}.
\end{aligned}
$$

Choosing $K_i \ge K_0^2,\ i=1,2$, one obtains

$$
\left|R^{up(lo)}(z,\lambda,K_2,y)-R^{up(lo)}(z,\lambda,K_1,y)\right| \le 2\gamma_1(K_0)+\frac{4L}{K_0},
$$

which implies the existence of the limits (5.4). Passing to the limit as l tends to infinity in (5.2) then allows us to obtain

$$
\left|R^{up(lo)}(z,\lambda)-R^{up(lo)}(z,\lambda,K,y)\right| \le \gamma_1(K).
$$

Note that from (5.1), the above inequalities will hold for any $y \in P$, which establishes (4.1). □

Proof of Lemma 5.2. By Assumption 4.3, the trajectories of (2.14) which started at $y \in P$ do not leave a compact set P_1. By Assumption 4.1, the functions Φ and f_1 satisfy Lipschitz conditions in z and y on $D \times P_1$ with some constant L_1. Hence, for any admissible pair of controls $\tilde{u}_k, \tilde{v}_k$

$$
\begin{aligned}
&\left|Q(z,\lambda,N,y^1,\tilde{u},\tilde{v})-Q(z,\lambda,N,y^2,\tilde{u},\tilde{v})\right|\\
&\le (L_1+\parallel\lambda\parallel L_1)\,N^{-1}\sum_{k=0}^{N-1}\parallel y_z^1(k)-y_z^2(k)\parallel\\
&\le L_1\left(1+\max_{\lambda\in\Lambda}\parallel\lambda\parallel\right)\left(N^{-1}\sum_{k=0}^{N-1}\xi(k)\|y^1-y^2\|\right)\stackrel{\text{def}}{=}\gamma_1(N). \qquad (5.5)
\end{aligned}
$$

Notice that from the fact that $\xi(k)$ tends to zero as k tends to infinity, it follows that $\gamma_1(N)$ tends to zero as N tends to infinity. Also without loss of generality $\gamma_1(N)$ can be taken to be a monotone function. Inequality (5.5) and representations (2.19), (2.20) imply (5.1).

Let us verify (5.2). For $l=1$ it is obvious. Let us use induction to show that (5.2) is true for R^{up}. Assume that for $l=r\ge 1$,

$$
|R^{up}(z,\lambda,rK,y)-R^{up}(z,\lambda,K,y)| \le \gamma_1(K), \qquad (5.6)
$$

with $\gamma_1(N)$ being a monotone function which tends to zero as N tends to infinity and which allows (5.1) with $y^i \in P_1, i=1,2$, where P_1 is a compact set containing all the trajectories of (2.14) started in P. Using (2.19) one can write

$$R^{up}(z,\lambda,(r+1)K,y)=\inf_{\tilde v_0}\sup_{\tilde u_0}\dots\inf_{\tilde v_{rK-1}}\sup_{\tilde u_{rK-1}}\left\{\frac{1}{(r+1)K}\sum_{k=0}^{rK-1}q(z,\tilde y_z(k),\tilde u_k,\tilde v_k)\right.$$

$$\left.+\frac{1}{r+1}R^{up}(z,\lambda,K,\tilde y_z(rK))\right\},\tag{5.7}$$

where $q(z,y,u,v)\overset{\text{def}}{=}\Phi(z,y,u,v)+\lambda^T f_1(z,y,u,v)$ and $\tilde y_z(k)$ is the solution of (2.14) obtained with a control pair $(\tilde u_k,\tilde v_k)$ and with the initial conditions $\tilde y_z(0)=y$. By (5.1),

$$\left|\frac{1}{r+1}R^{up}(z,\lambda,K,\tilde y_z(rK))-\frac{1}{r+1}R^{up}(z,\lambda,K,y)\right|\le\frac{\gamma_1(K)}{r+1}.$$

Hence,

$$\begin{aligned}R^{up}(z,\lambda,(r+1)K,y)&\le\inf_{\tilde v_0}\sup_{\tilde u_0}\dots\inf_{\tilde v_{rK-1}}\sup_{\tilde u_{rK-1}}\left\{\frac{r}{(r+1)}\cdot\frac{1}{rK}\sum_{k=0}^{rK-1}q(z,\tilde y_z(k),\tilde u_k,\tilde v_k)\right\}\\&\quad+\frac{1}{r+1}R^{up}(z,\lambda,K,y)+\frac{\gamma_1(K)}{r+1}\\&=\frac{r}{r+1}R^{up}(z,\lambda,rK,y)+\frac{1}{r+1}R^{up}(z,\lambda,K,y)+\frac{\gamma_1(K)}{r+1}\end{aligned}\tag{5.8}$$

and, similarly,

$$R^{up}(z,\lambda,(r+1)K,y)\ge\frac{r}{r+1}R^{up}(z,\lambda,rK,y)+\frac{1}{r+1}R^{up}(z,\lambda,K,y)-\frac{\gamma_1(K)}{r+1}.\tag{5.9}$$

From (5.8) and (5.6) it follows that

$$\begin{aligned}R^{up}(z,\lambda,(r+1)K,y)&\le\frac{r}{r+1}(R^{up}(z,\lambda,K,y)+\gamma_1(K))\\&\quad+\frac{1}{r+1}R^{up}(z,\lambda,K,y)+\frac{\gamma_1(K)}{r+1}\\&=R^{up}(z,\lambda,K,y)+\gamma_1(K).\end{aligned}$$

In the same way, from (5.9) and (5.6) it follows that

$$R^{up}(z,\lambda,(r+1)K,y)\ge R^{up}(z,\lambda,K,y)-\gamma_1(K).$$

This completes the proof of (5.2) for R^{up}. The proof for R^{lo} is similar.

To prove (5.3) let us notice that for any control pairs $(\tilde u_k,\tilde v_k)$ defined on $[0,N_2]$ $(N_2\ge N_1\ge 1)$,

$$|N_2Q(z,\lambda,N_2,\tilde u,\tilde v)-N_1Q(z,\lambda,N_1,\tilde u,\tilde v)|\le M\,(N_2-N_1)\,,\tag{5.10}$$

where

$$M=\max\left\{|\Phi(z,y,u,v)|+\|\lambda\|\,\|f_1(z,y,u,v)\|\right.$$

$$\left.\Big|(z,\lambda,y,u,v)\in D\times\Lambda\times P_1\times U\times V\right\},\tag{5.11}$$

where P_1 is a compact subset of R^n containing all the trajectories of (2.14) which start in P. On the basis of (5.10), (2.19), and (2.20), it can be easily verified that

$$|N_2 R^{up(lo)}(z,\lambda,N_2,y) - N_1 R^{up(lo)}(z,\lambda,N_1,y)| \le M\,(N_2 - N_1)\,. \tag{5.12}$$

By (5.11), $|R^{up(lo)}(z,\lambda,N,y)| \le M\ \forall N \ge 0$, which along with (5.12) allows us to establish the inequalities

$$\begin{aligned}
&\left|R^{up(lo)}(z,\lambda,N_2,y) - R^{up(lo)}(z,\lambda,N_1,y)\right| \\
&\le \left|R^{up(lo)}(z,\lambda,N_2,y) - N_1 {N_2}^{-1} R^{up(lo)}(z,\lambda,N_1,y)\right| \\
&\quad + \left|N_1 {N_2}^{-1} - 1\right| \left|R^{up(lo)}(z,\lambda,N_1,y)\right| \\
&\le 2M\,(N_2 - N_1)\,{N_2}^{-1},
\end{aligned}$$

which implies (5.3), where L can be taken to be equal to $2M$.

THEOREM 5.3. *Assume that all assumptions of Theorem* 5.1 *are satisfied and the function* $\xi(k)$ *in Assumption* 4.7 *has the form* $\xi(k) = \xi^k C$, *where* $0 < \xi < 1$ *and* $C > 0$ *is a constant. Then Assumptions* 4.5 *and* 4.6 *are true.*

Proof. It can be worked out essentially following a similar line as in the proof of the results of Gaitsgory [12] and combining with Lemma 5.4 below. □

LEMMA 5.4. *Let* $\tilde{z}(k)$ *be an arbitrary function belonging to a compact set for* $k = 0, 1, \ldots, N$. *Let* $\tilde{u}_k$ *be an arbitrary admissible control and let* $\tilde{y}(k)$ *and* $\tilde{y}_z(k)$ *be the solutions of the system*

$$\tilde{y}(k+1) = \tilde{y}(k) + f_2(\tilde{z}(k), \tilde{y}(k), \tilde{u}_k), \quad \tilde{y}(0) = y, \tag{5.13}$$

and the system

$$\tilde{y}_z(k+1) = \tilde{y}_z(k) + f_2(z, \tilde{y}_z(k), \tilde{u}_k), \quad \tilde{y}_z(0) = y, \tag{5.14}$$

respectively (where, in contrast to (5.13), z *is constant in* (5.14)*). Then*

$$\max_{k=0,1,\ldots,N} \|\tilde{y}(k) - \tilde{y}_z(k)\| \le C \max_{k=0,1,\ldots,N} \|\tilde{z}(k) - z\|, \tag{5.15}$$

where $C_1 > 0$ *is a constant.*

Proof of Lemma 5.4. The proof can be carried out by similar techniques as those in [9].

Choose a natural number K in such a way that

$$C_1 \xi^K \stackrel{\text{def}}{=} \delta < 1 \tag{5.16}$$

and define $\tau_l = lK,\ l = 0, 1, \ldots, \lfloor \frac{N}{K} \rfloor,\ \tau_{l+1} \stackrel{\text{def}}{=} N$.

To continue the proof, we need the following simple proposition.

PROPOSITION 5.5. *Let* $\alpha_k \ge 0,\ k = \tau_l, \tau_l + 1, \ldots, \tau_{l+1}$, *satisfy the inequalities*

$$\alpha_{\tau_l} \le M_1 \Delta, \tag{5.17}$$

$$\alpha_{k+1} \le M_2 \Delta + M_3 \sum_{i=\tau_l}^{k} \alpha_i, \tag{5.18}$$

where M_i, $i = 1, 2, 3$, are constants and Δ is a positive parameter. Then there exists a constant θ such that

$$\alpha_k \le \theta\Delta, \qquad k = \tau_l, \tau_l + 1, \ldots, \tau_{l+1}. \tag{5.19}$$

Proof. It can be easily derived by induction. □

Returning to the proof of the main lemma, denote by $\tilde{y}_z^l(k)$, $k = \tau_l, \tau_l+1, \ldots, \tau_{l+1}$, the solution of (5.14) obtained with the controls $\tilde{u}_k$ and the initial condition

$$\tilde{y}_z^l(\tau_l) = \tilde{y}(\tau_l). \tag{5.20}$$

By definition $\tilde{y}_z^l(k)$ satisfies the equation

$$\tilde{y}_z^l(k+1) = \tilde{y}(\tau_l) + \sum_{\tau=\tau_l}^{k} f_2\left(z, \tilde{y}_z^l(\tau), \tilde{u}_\tau\right). \tag{5.21}$$

Also, $\tilde{y}(k)$ satisfies the equation

$$\tilde{y}(k+1) = \tilde{y}(\tau_l) + \sum_{\tau=\tau_l}^{k} f_2\left(\tilde{z}, \tilde{y}(\tau), \tilde{u}_\tau\right). \tag{5.22}$$

Subtracting (5.21) from (5.22) and using Lipschitz conditions ($f_2(\cdot)$ is assumed to satisfy these conditions in z and y with the constant L), one obtains

$$\|\tilde{y}_z^l(k+1) - \tilde{y}(k+1)\| \le LK\Delta + L\sum_{\tau=\tau_l}^{k} \|\tilde{y}_z^l(\tau) - \tilde{y}(\tau)\|, \tag{5.23}$$

where

$$\Delta \stackrel{\text{def}}{=} \max_{k=0,1,\ldots,N} \|z(k) - z\|. \tag{5.24}$$

Denoting

$$\|\tilde{y}_z^l(k) - \tilde{y}(k)\| \stackrel{\text{def}}{=} \alpha_k$$

and taking

$$M_1 = 0, \quad M_2 = LK, \quad M_3 = L,$$

one obtains from Proposition 5.5 that there exists a constant θ_1 such that

$$\|\tilde{y}_z^l(k) - \tilde{y}(k)\| \le \theta_1\Delta, \qquad k = \tau_l, \tau_l + 1, \ldots, \tau_{l+1}. \tag{5.25}$$

On the basis of the assumption on $\xi(k)$ in Theorem 5.3, and (5.16) and (5.25), one can write

$$\begin{aligned}
&\|\tilde{y}_z(\tau_{l+1}) - \tilde{y}(\tau_{l+1})\| \\
&\le \|\tilde{y}_z(\tau_{l+1}) - \tilde{y}_z^l(\tau_{l+1})\| + \|\tilde{y}_z^l(\tau_{l+1}) - \tilde{y}(\tau_{l+1})\| \\
&\le \delta\|\tilde{y}_z(\tau_l) - \tilde{y}_z^l(\tau_l)\| + \theta_1\Delta \\
&= \delta\|\tilde{y}_z(\tau_l) - \tilde{y}(\tau_l)\| + \theta_1\Delta \\
&\le \delta\left(\delta\|\tilde{y}_z(\tau_{l-1}) - \tilde{y}(\tau_{l-1})\| + \theta_1\Delta\right) + \theta_1\Delta \\
&\ldots\ldots \\
&\le (\delta^{l-1} + \delta^{l-2} + \cdots + 1)\theta_1\Delta \le \frac{\theta_1\Delta}{1-\delta}\,.
\end{aligned} \tag{5.26}$$

The solution $\tilde{y}_z(k)$ of (5.14) satisfies the equation

$$\tilde{y}_z(k+1) = \tilde{y}_z(k) + \sum_{\tau=\tau_l}^{k} f_2\left(z, \tilde{y}_z(\tau), \tilde{u}_\tau\right). \tag{5.27}$$

Subtracting this from (5.22), one obtains, by using (5.26),

$$\begin{aligned} &\|\tilde{y}(k+1) - \tilde{y}_z(k+1)\| \\ &\le \|\tilde{y}(\tau_l) - \tilde{y}_z(\tau_l)\| + LK\Delta + L\sum_{\tau=\tau_l}^{k} \|\tilde{y}(\tau_l) - \tilde{y}_z(\tau)\| \\ &\le \left[\frac{\theta_1}{1-\delta} + LK\right]\Delta + L\sum_{\tau=\tau_l}^{k} \|\tilde{y}(\tau_l) - \tilde{y}_z(\tau)\|. \end{aligned} \tag{5.28}$$

Denoting

$$\|\tilde{y}(k) - \tilde{y}_z(k)\| = \alpha_k$$

and taking

$$M_1 = \frac{\theta_1\Delta}{1-\delta}, \quad M_2 = \frac{\theta_1}{1-\delta} + LK, \quad M_3 = L,$$

one obtains from Proposition 5.5 that there exists a constant θ_2 such that

$$\|\tilde{y}(k) - \tilde{y}_z(k)\| \le \theta_2\Delta, \qquad k = \tau_l, \tau_l + 1, \ldots, \tau_{l+1}. \tag{5.29}$$

Note that the constant θ_2 does not depend on l and, hence, (5.29) proves the lemma with $C = \theta_2$. □

6. Conclusions. In this paper, the problem of singularly perturbed zero-sum dynamic games with full information has been investigated. With the aid of an associated fast game, it has been shown that the upper and lower value functions of this game have limits as the singular perturbations parameter tends to zero. It has also been established that these limits coincide with viscosity solutions of some Hamilton–Jacobi-type equations. Two examples were presented to illustrate the general results.

Acknowledgments. The author would like to express his sincere gratitude and appreciation to Professor Vladimir Gaitsgory from the University of South Australia for discussion and encouragement during this work. Also, the author wishes to thank the associate editor and referees for their valuable comments and suggestions which have improved the presentation.

REFERENCES

[1] Z. Artstein and V. Gaitsgory, *Tracking fast trajectories along a slow dynamics: A singular perturbations approach*, SIAM J. Control Optim., 35 (1997), pp. 1487–1507.

[2] Z. Artstein and A. Vigodner, *Singularly perturbed ordinary differential equations with dynamic limits*, Proc. Roy. Soc. Edinburgh Sect. A, 126 (1996), pp. 541–569.

[3] J. Aubin, *Viability Theory*, Birkhäuser, Boston, 1991.

[4] F. Clarke, Y. Ladyaev, E. Sontag, and A. Subbotin, *Asymptotic controllability implies feedback stabilization*, IEEE Trans. Automat. Control, 42 (1997), pp. 1394–1407.

[5] L. C. Evans and P. E. Souganidis, *Differential games and representations formulas for solutions of Hamilton-Jacobi-Isaacs equations*, Indiana Univ. Math. J., 33 (1984), pp. 773–797.

[6] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.
[7] A. Friedman, *Differential Games*, Wiley, New York, 1971.
[8] V. Gaitsgory, *Use of the averaging method in control problems*, Differential Equations, 22 (1986), pp. 1290–1299.
[9] V. Gaitsgory, *Suboptimization of singularly perturbed control systems*, SIAM J. Control Optim., 30 (1992), pp. 1228–1249.
[10] V. Gaitsgory, *Control of Systems with Slow and Fast Motions*, Nauka, Moscow, 1991 (in Russian).
[11] V. Gaitsgory, *Suboptimal control of singularly perturbed systems and periodic optimization*, IEEE Trans. Automat. Control, 38 (1993), pp. 888–903.
[12] V. Gaitsgory, *Limit Hamilton-Jacobi-Isaacs equations for singularly perturbed zero-sum differential games*, J. Math. Anal. Appl., 202 (1996), pp. 862–899.
[13] V. Gaitsgory and P. Shi, *Limit Hamilton-Jacobi-Isaacs equations for singularly perturbed zero-sum dynamic (discrete time) games*, in Proceedings of the 7th International Symposium on Dynamic Games and Applications, Kanagawa, Japan, International Society of Dynamic Games, 1996, pp. 168–174.
[14] B. F. Gardner and J. B. Cruz, *Well-posedness of singular perturbed Nash games*, J. Franklin Inst., 30 (1978), pp. 355–374.
[15] F. Garofalo and G. Leitmann, *Nonlinear composite control of a class of nominally linear singularly perturbed uncertain systems*, in Deterministic Control of Uncertain Systems, A.S.I. Zinober, ed., IEE Press, London, 1990, pp. 269–288.
[16] G. Grammel, *Controllability of differential inclusions*, J. Dynam. Control Systems, 1 (1995), pp. 581–595.
[17] G. Grammel, *Singularly perturbed control systems: Recent progress*, in Proceedings of the 35th IEEE Conference on Decision and Control, Kobe, Japan, IEEE Control Systems Society, 1996, pp. 505–510.
[18] G. Grammel, *Singularly perturbed differential inclusions: An averaging approach*, Set-Valued Anal., 4 (1996), pp. 361–374.
[19] R. Isaacs, *Differential Games*, Wiley, New York, 1965.
[20] H. Ishii, *Uniqueness of unbounded viscosity solutions of Hamilton-Jacobi equations*, Indiana Univ. Math. J., 26 (1984), pp. 721–748.
[21] H. K. Khalil and P. V. Kokotovic, *Feedback and well-posedness of singularly perturbed Nash games*, IEEE Trans. Automat. Control, 24 (1979), pp. 699–708.
[22] P. V. Kokotović, *Applications of singular perturbation techniques to control problems*, SIAM Rev., 26 (1984), pp. 501–550.
[23] P. V. Kokotovic, H. Khalil, and J. O'Reilly, *Singular Perturbations in Control: Analysis and Design*, Academic Press, New York, 1984.
[24] P. V. Kokotovic, R. E. O'Malley, and P. Sannuti, *Singular perturbations and order reduction in control theory*, Automatica J. IFAC, 12 (1976), pp. 123–132.
[25] R. E. O'Malley, *Introduction to Singular Perturbations*, Academic Press, New York, 1974.
[26] Z. Pan and T. Basar, *H^∞-optimal control for singularly perturbed systems. Part* I: *Perfect state measurements*, Automatica J. IFAC, 29 (1993), pp. 401–423.
[27] Z. Pan and T. Basar, *H^∞-optimal control for singularly perturbed systems. Part* II: *Imperfect state measurements*, IEEE Trans. Automat. Control, 39 (1994), pp. 280–299.
[28] A. A. Pervozvansky and V. Gaitsgory, *Theory of Suboptimal Decisions*, Kluwer Academic, Dordrecht, The Netherlands, 1988.
[29] M. Quincampox, *Contribution a L'etude des perturbations singulieres pour les systemes controles et les inclusions differentialles*, C. R. Acad. Sci. Paris Sér. I Math., 316 (1993), pp. 133–138.
[30] A. N. Tichonov, *Systems of differential equations containing small parameters near derivations*, Mat. Sb. (N.S.), 31 (1952), pp. 575–586 (in Russian).
[31] A. B. Vasil'eva and A. F. Butuzov, *Asymptotic Expansions of Solutions to Singularly Perturbed Equations*, Nauka, Moscow, 1973 (in Russian).
[32] A. Vigodner, *Limits of singularly perturbed control problems with statistical limits of fast motions*, SIAM J. Control Optim., 35 (1997), pp. 1–28.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 3, pp. 851–874

SIMPLE MECHANICAL CONTROL SYSTEMS WITH CONSTRAINTS AND SYMMETRY*

J. CORTÉS†, S. MARTÍNEZ†, J. P. OSTROWSKI‡, AND H. ZHANG§

Abstract. We develop tools for studying the control of underactuated mechanical systems that evolve on a configuration space with a principal fiber bundle structure. Taking the viewpoint of affine connection control systems, we derive reduced formulations of the Levi–Civita and the nonholonomic affine connections, along with the symmetric product, in the presence of symmetries and nonholonomic constraints. We note that there are naturally two kinds of connections to be considered here, affine and principal connections, leading to what we term a "connection within a connection." These results are then used to describe controllability tests that are specialized to simple, underactuated mechanical systems on principal fiber bundles, including the notion of fiber configuration controllability. We present examples of the use of these tools in studying the planar rigid body with a variable direction (vectored) thruster and the snakeboard robot.

Key words. nonlinear control, reduction, configuration controllability, symmetric product

AMS subject classifications. 53B05, 70Q05, 93B03, 93B05, 93B29

PII. S0363012900381741

1. Introduction.

In the area of control for mechanical systems, there is a newly emerging body of work that utilizes the special Lagrangian structure of such systems to help focus the control analysis [5, 6, 8, 12, 20, 21]. This perspective, in which the dynamics of simple mechanical systems is interpreted using an affine connection, has led to new insights into both control and motion planning for a number of underactuated mechanical systems. In this paper, we study the effect of symmetries and constraints on the tools that are used in studying affine connection control systems, namely, the *affine connection* and the *symmetric product*.

In studying the controllability of a mechanical system, classical tools from nonlinear control theory [27] suggest that one compute the closure by the Lie bracket of all the control inputs and the drift vector field. When the control inputs enter in as forcing terms for second-order ODEs, such as is the case with forces or torques, this procedure requires the system to be transformed into a first-order form. The drawback of this, however, is that the conversion requires that one treat the velocities as a part of the state and, more importantly, that the intrinsic structure of the mechanical system as a second-order Lagrangian system is covered up. However, work by Lewis and Murray [21] has shown that a proper geometric interpretation of simple mechanical systems can be achieved through the use of the affine connection formalism and the symmetric product that derives from it.

*Received by the editors November 28, 2000; accepted for publication (in revised form) January 17, 2002; published electronically September 12, 2002. The research of the first and second authors was partially supported by FPU and FPI grants from the Spanish Ministerio de Educación y Cultura and grant DGICYT PB97-1257. The research of the third and fourth authors was partially supported by NSF grants IRI-9711834, IIS-9876301, and ECS-0086931, and by ARO grant DAAH04-96-1-0007.
http://www.siam.org/journals/sicon/41-3/38174.html

†Laboratory of Dynamical Systems, Mechanics and Control, Instituto de Matemáticas y Física Fundamental, CSIC, Serrano 123, Madrid 28006, Spain (j.cortes@imaff.cfmac.csic.es, s.martinez@imaff.cfmac.csic.es).

‡General Robotics, Automation, Sensing and Perception Laboratory, University of Pennsylvania, 3401 Walnut Street, Philadelphia, PA 19104-6228 (jpo@grip.cis.upenn.edu).

§Mechanical Engineering Department, Rowan University, 133 Rowan Hall, 201 Mullica Hill Road, Glassboro, NJ 08028-1701 (zhang@rowan.edu).

Bullo, Leonard, and Lewis later applied these results to underactuated Lagrangian systems evolving on a Lie group [6, 7]. They took advantage of the special Lie group structure to derive algorithms for generating the control inputs that lead to motion along the directions generated through the operation of the symmetric product. This work serves as a starting point for this paper, in which we explore a generalization of these ideas to systems that evolve on principal fiber bundles, which are locally the product of a Lie group and a general smooth manifold.

We also note that there has been extensive work in the area of understanding the role of symmetries in mechanical systems (see, e.g., [2, 4, 9, 10, 16, 23, 24, 25] and references therein). We focus on one aspect of such systems, where internal shape variables play an important role in determining the motion of a system along a Lie group. Lagrangian reduction provides powerful tools for analyzing mechanical systems on fiber bundles. Generally (as is the case for our examples), the Lie group describes the position and orientation of the system, while the remaining variables constitute an internal shape space. Some examples of the shape variables that result are the thruster angle of a blimp [35, 36], the leg angle of a robot leg [21], and the wheel direction angles of a snakeboard [30, 31]. Through a local trivialization, we can use the internal symmetries of the system to decouple the dynamics into two parts, vertical and horizontal, and connect them with a mechanical connection (or constraints) [4, 31]. Likewise, we can apply the same technique to the computation of covariant derivatives by finding the vertical and horizontal parts and then use the Lie bracket and symmetric product to take advantage of the geometric structure of the system, leading to simplified tests of configuration accessibility and controllability. When nonholonomic constraints are present, the situation is further complicated, though it was shown by Lewis [20] that one can use a *nonholonomic affine connection* that directly extends the controllability results.

Our motivation for studying this class of systems comes from robotics, where it has been noted that *robotic locomotion systems* possess this structure—the dynamics evolve on a product bundle between a Lie group and a general "shape" manifold [15, 28, 31]. This leads us to consider mechanical systems on principal fiber bundles, in which the motion of the system is generated through a complex interaction of thrusts/forces and internal changes in the shape or configuration of the robot. There is an extensive literature studying such systems, including kinematic versions [15], dynamic systems that evolve purely on Lie groups [6], and dynamic systems with nonholonomic constraints [28, 30]. An important quantity for such robotic systems that is highlighted here is the notion of *fiber controllability*, introduced by Kelly and Murray [15] for driftless, kinematic systems. The notion of fiber controllability stems from the fact that, for many robotic systems of this form, one cares only that the robot be able to control its position and orientation, without regard to the configuration of its internal shape. Thus the emphasis is on understanding whether a system is controllable only along the fiber (position and orientation). We extend this notion to dynamic systems with symmetries living on trivial principal fiber bundles.

The paper is organized as follows. In section 2, we give some background on simple mechanical control systems and the role of symmetries. In section 3, we study the reduced version of the Levi–Civita affine connection and hence the symmetric product for principal fiber bundles. We present the computations in terms of local forms of the quantities that arise, including the *mechanical connection* and the *locked inertia tensor*, since these allow for a reduced and compact representation. We follow up this derivation in section 4 with a parallel formulation of the nonholonomic affine

connection that arises when constraints are present. In section 5, we describe how these results extend previous notions of configuration controllability to fiber bundles and introduce a new concept of fiber controllability. In section 6, we demonstrate the use of these tools in two motivating examples: the underactuated rigid body (or planar blimp) and the snakeboard. Finally, section 7 is devoted to some concluding remarks.

2. Background on simple mechanical control systems. In this section, we describe the geometric framework utilized in the study of mechanical control systems. We follow [20, 21] in the exposition of affine connection control systems. The reader is referred to [1, 17] for more details on notions such as principal bundles or affine connections.

2.1. Affine connection control systems. Let Q be an n-dimensional manifold. We denote by TQ the tangent bundle of Q, by $\mathfrak{X}(Q)$ the set of vector fields on Q, and by $C^\infty(Q)$ the set of smooth functions on Q. A *simple mechanical control system* is defined by a tuple $(Q, \mathcal{G}, V, \mathcal{F})$, where Q is the manifold of configurations of the system, $\mathcal{G}$ is a Riemannian metric on Q (the kinetic energy metric of the system), $V \in C^\infty(Q)$ is the potential function, and $\mathcal{F} = \{F^1, \dots, F^m\}$ is a set of m linearly independent 1-forms on Q, which physically correspond to forces or torques.

The dynamics of simple mechanical control systems is classically described by the *forced Euler–Lagrange equations*

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) - \frac{\partial L}{\partial q} = \sum_{i=1}^{m} u_i(t) F^i, \tag{2.1}$$

where $L(q, \dot{q}) = \frac{1}{2}\mathcal{G}(\dot{q}, \dot{q}) - V(q)$ is the *Lagrangian* of the system.

Alternatively, one can express the control equations (2.1) using the natural affine connection associated to the metric $\mathcal{G}$, the Levi–Civita connection. An *affine connection* [17] is defined as an assignment

$$\begin{array}{rccc} \nabla : & \mathfrak{X}(Q) \times \mathfrak{X}(Q) & \longrightarrow & \mathfrak{X}(Q) \\ & (X, Y) & \longmapsto & \nabla_X Y, \end{array}$$

which is $\mathbb{R}$-bilinear and satisfies $\nabla_{fX} Y = f \nabla_X Y$ and $\nabla_X (fY) = f \nabla_X Y + X(f) Y$ for any X, $Y \in \mathfrak{X}(Q)$, $f \in C^\infty(Q)$. In local coordinates,

$$\nabla_X Y = \left(\frac{\partial Y^a}{\partial q^b} X^b + \Gamma^a_{bc} X^b Y^c\right) \frac{\partial}{\partial q^a},$$

where $\Gamma^a_{bc}(q)$ are the *Christoffel symbols* of the affine connection defined by

$$\nabla_{\frac{\partial}{\partial q^b}} \frac{\partial}{\partial q^c} = \Gamma^a_{bc} \frac{\partial}{\partial q^a}. \tag{2.2}$$

For simple mechanical control systems, the *Levi–Civita connection* $\nabla^{\mathcal{G}}$ associated to the metric $\mathcal{G}$ is determined by the formula

$$\begin{aligned} 2\,\mathcal{G}(Z, \nabla_X Y) = {} & X(\mathcal{G}(Z, Y)) + Y(\mathcal{G}(Z, X)) - Z(\mathcal{G}(Y, X)) + \mathcal{G}(X, [Z, Y]) \\ & + \mathcal{G}(Y, [Z, X]) - \mathcal{G}(Z, [Y, X]), \quad X, Y, Z \in \mathfrak{X}(Q). \end{aligned} \tag{2.3}$$

One can compute the Christoffel symbols of $\nabla^{\mathcal{G}}$ to be

$$\Gamma^a_{bc} = \frac{1}{2}\mathcal{G}^{ad}\left(\frac{\partial \mathcal{G}_{db}}{\partial q^c} + \frac{\partial \mathcal{G}_{dc}}{\partial q^b} - \frac{\partial \mathcal{G}_{bc}}{\partial q^d}\right),$$

where $(\mathcal{G}^{ad})$ denotes the inverse matrix of $(\mathcal{G}_{da} = \mathcal{G}(\frac{\partial}{\partial q^d}, \frac{\partial}{\partial q^a}))$. Instead of the input forces $F^1, \ldots, F^m$, we shall make use of the vector fields $Y_1, \ldots, Y_m$, defined as $Y_i = \sharp_{\mathcal{G}}(F^i)$, where $\sharp_{\mathcal{G}} = \flat_{\mathcal{G}}^{-1}$ and $\flat_{\mathcal{G}} : TQ \longrightarrow T^*Q$ is the musical isomorphism given by $\flat_{\mathcal{G}}(X)(Y) = \mathcal{G}(X, Y)$. In local coordinates, we have that $Y_i^a = \mathcal{G}^{ab}F_b^i$ for each $1 \leq i \leq m$. Roughly speaking, this corresponds to considering the effect of the controls on "accelerations" rather than on forces. The control equations (2.1) for the mechanical system may then be recast as

$$\nabla^{\mathcal{G}}_{\dot{c}(t)}\dot{c}(t) = -\operatorname{grad} V + \sum_{i=1}^{m} u^i(t) Y_i(c(t)), \tag{2.4}$$

where $\operatorname{grad} V = \sharp_{\mathcal{G}}(dV)$. Observe that we can use a general affine connection in (2.4) instead of the Levi–Civita connection without changing the structure of the equation. This is particularly interesting, since nonholonomic mechanical control systems also give rise to equations of the form of (2.4), as we review in the following [20]. This observation is actually very powerful, since controllability analyses based on a general affine connection (cf. section 5) are valid for both unconstrained and constrained control systems.

A *constrained mechanical control system* $(Q, \mathcal{G}, V, \mathcal{F}, \mathcal{D})$ is a simple mechanical control system $(Q, \mathcal{G}, V, \mathcal{F})$ subject to the constraints given by the $(n-l)$-dimensional (nonholonomic) distribution $\mathcal{D}$ on Q. In a local description, $\mathcal{D}$ can be defined by the vanishing of l independent constraint functions $\omega_j(q)\dot{q}$, $1 \leq j \leq l$. The application of Lagrange–d'Alembert's principle leads to the constrained equations of motion

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) - \frac{\partial L}{\partial q} = \sum_{i=1}^{m} u_i(t) F^i + \sum_{j=1}^{l} \lambda^j \omega_j, \tag{2.5}$$

which, together with the constraint equations $\omega_j(q)\dot{q} = 0$, describe the dynamics of the nonholonomic system. Here, the λ^j are the Lagrange multipliers. The term $\sum_{j=1}^{l} \lambda^j \omega_j$ represents the "reaction force" due to the constraints.

The second-order equation (2.5) can alternatively be written as

$$\begin{cases} \nabla^{\mathcal{G}}_{\dot{c}(t)}\dot{c}(t) = \lambda(t) - \operatorname{grad} V + \sum_{i=1}^{m} u_i(t) Y_i(c(t)), \\ \dot{c}(t) \in \mathcal{D}_{c(t)}, \end{cases} \tag{2.6}$$

where now λ is seen as a section of $\mathcal{D}^{\perp}$, the $\mathcal{G}$-orthogonal complement to $\mathcal{D}$, along the curve c. Letting $\mathcal{P} : TQ \longrightarrow \mathcal{D}$, $\mathcal{Q} : TQ \longrightarrow \mathcal{D}^{\perp}$ denote the complementary $\mathcal{G}$-orthogonal projectors, we can define an affine connection

$$\overline{\nabla}_X Y = \nabla^{\mathcal{G}}_X Y + (\nabla^{\mathcal{G}}_X \mathcal{Q})(Y) = \mathcal{P}(\nabla^{\mathcal{G}}_X Y) + \nabla^{\mathcal{G}}_X(\mathcal{Q}(Y)),$$

such that the nonholonomic control equations (2.6) can be rewritten as

$$\overline{\nabla}_{\dot{c}(t)}\dot{c}(t) = -\mathcal{P}(\operatorname{grad} V) + \sum_{i=1}^{m} u_i(t) \mathcal{P}(Y_i(c(t))), \tag{2.7}$$

and where we select the initial velocity in $\mathcal{D}$ (cf. [20] for details). Observe that the inputs Y_i act on the system only through their $\mathcal{D}$-component. Indeed, the Lagrange multiplier $\lambda \in \mathcal{D}^{\perp}$ absorbs their $\mathcal{D}^{\perp}$-components. The connection $\overline{\nabla}$ is called the

nonholonomic affine connection [3, 19, 20, 33]. Note that (2.4) and (2.7) have the same structure.

It can be easily deduced from its definition that $\overline{\nabla}$ restricts to $\mathcal{D}$; that is,

$$\overline{\nabla}_X Y = \mathcal{P}(\nabla^{\mathcal{G}}_X Y) \in \mathcal{D} \text{ for all } Y \in \mathcal{D},\ X \in \mathfrak{X}(Q).$$

This kind of affine connection, which restricts to a given distribution, has been studied in [19]. In particular, such a behavior implies that the distribution $\mathcal{D}$ is *geodesically invariant*; that is, for every geodesic $c(t)$ of $\overline{\nabla}$ starting from a point in $\mathcal{D}$, $\dot{c}(0) \in \mathcal{D}_{c(0)}$, we have that $\dot{c}(t) \in \mathcal{D}_{c(t)}$.

As we shall see later, a key tool in the controllability analysis and description of mechanical control systems is the *symmetric product* $\langle \cdot : \cdot \rangle$ associated to an affine connection ∇ (see [13, 21, 34]). Given X, $Y \in \mathfrak{X}(Q)$, define

$$\langle X : Y \rangle = \nabla_X Y + \nabla_Y X.$$

The symmetric product characterizes geodesically invariant distributions. Indeed, one can prove that $\mathcal{D}$ is geodesically invariant for the nonholonomic connection if and only if $\langle X : Y \rangle \in \mathcal{D}$ for all X, $Y \in \mathcal{D}$ (see [19]). Recently, Bullo [5] has shown that the evolution of mechanical control systems when starting from rest can be described by a series involving repeated symmetric products of the input vector fields, extending the possibilities of use of the symmetric product to the design of motion control algorithms.

2.2. Principal fiber bundles. The notion of principal fiber bundles is present in many locomotion and robotic systems, since they commonly exhibit translational and rotational symmetries. Examining the configuration space Q, one can observe that there exists a splitting $Q = G \times M$ between variables describing the position and orientation of the robot, i.e., the *pose* coordinates $g \in G$, and variables describing the internal shape of the system, the *shape* coordinates $r \in M$. This exactly corresponds to the case of a trivial principal fiber bundle, decomposed into *fiber* space, G, and *base* space, M, respectively.

Geometrically, this situation is described as follows. Assume there is a Lie group G acting on Q

$$\begin{array}{rccl} \Phi: & G \times Q & \longrightarrow & Q \\ & (g, q) & \longmapsto & \Phi(g, q) = \Phi_g(q) = gq. \end{array}$$

The orbit through a point q is $\mathrm{Orb}_G(q) = \{gq \mid g \in G\}$. We denote by $\mathfrak{g}$ the Lie algebra of G. For any element $\xi \in \mathfrak{g}$, let ξ_Q denote the corresponding infinitesimal generator of the group action on Q. Then

$$T_q(\mathrm{Orb}_G(q)) = \{\xi_Q(q) \mid \xi \in \mathfrak{g}\}.$$

If the action Φ is free and proper, we can endow the quotient space $Q/G \cong M$ with a manifold structure such that the canonical projection $\pi : Q \longrightarrow M$ is a surjective submersion. Then we have that $Q(M, G, \pi)$ is a principal bundle with bundle space Q, base space M, structure group G, and projection π. Note that the kernel of $\pi_*(= T\pi)$ consists of the *vertical* tangent vectors, i.e., the vectors tangent to the orbits of G in Q. We denote the bundle of vertical vectors by $\mathcal{V}$, with $\mathcal{V}_q = T_q(\mathrm{Orb}_G(q))$, $q \in Q$.

Throughout the paper, we will usually deal with general principal fiber bundles, unless otherwise stated. Locally, one can always trivialize Q and work with $Q \supset$

$\pi^{-1}(U) \equiv G \times U$, where $U \subset M$ is an open subset of M. In the bundle coordinates (g, r), the projection reads $\pi(g, r) = r$, and the Lagrangian L can be written as

$$L(q, \dot{q}) = \frac{1}{2}(\dot{g}^T \, \dot{r}^T) \, \mathcal{G} \begin{pmatrix} \dot{g} \\ \dot{r} \end{pmatrix} - V(g, r),$$

where we note the abuse of notation resulting from changing between $\dot{g}$ as an argument in TG and as a vector (the same stands for $\mathcal{G}$ seen as a bilinear form or as a matrix). In the remainder of the paper, we will often make use of the same notation for coordinate-free and matrix formulas. The precise meaning should be clear from the context.

A *principal connection* on $Q(M, G, \pi)$ can be defined as a G-invariant distribution $\mathcal{H}$ on Q satisfying $T_qQ = \mathcal{H}_q \oplus \mathcal{V}_q$ for all $q \in Q$. The subspace $\mathcal{H}_q$ of T_qQ is called the *horizontal subspace* at q determined by the connection.

Alternatively, a principal connection can be characterized by a $\mathfrak{g}$-valued 1-form $\mathcal{A}$ on Q satisfying the following conditions:

(i) $\mathcal{A}(\xi_Q(q)) = \xi$ for all $\xi \in \mathfrak{g}$,

(ii) $\mathcal{A}((\Phi_g)_* X) = \mathrm{Ad}_g(\mathcal{A}(X))$ for all $X \in TQ$.

The horizontal subspace at q is then given by $\mathcal{H}_q = \{v_q \in T_qQ \mid \mathcal{A}(v_q) = 0\}$. In coordinates, using (i) and (ii), we can write

$$\begin{aligned} \mathcal{A}(g, r, \dot{g}, \dot{r}) &= \mathcal{A}(g(e, r, \xi, \dot{r})) = Ad_g \mathcal{A}(e, r, \xi, \dot{r}) \\ &= Ad_g(\mathcal{A}(e, r, \xi, 0) + \mathcal{A}(e, r, 0, \dot{r})) = Ad_g(\xi + A(r)\dot{r}). \end{aligned}$$

Note that A depends only on the shape variables. It is called the *local form of the connection* $\mathcal{A}$.

Given a principal connection, we have that every vector $v \in T_qQ$ can be uniquely written as $v = v^{hor} + v^{ver}$, with $v^{hor} \in \mathcal{H}_q$ and $v^{ver} = \mathcal{A}(v)_Q(q) \in \mathcal{V}_q$. The curvature $\mathcal{B}$ of the principal connection $\mathcal{A}$ is a $\mathfrak{g}$-valued 2-form on Q defined as follows: for each $q \in Q$ and $u, v \in T_qQ$,

$$\mathcal{B}(u, v) = d\mathcal{A}(u^{hor}, v^{hor}) = -\mathcal{A}([u^{hor}, v^{hor}]).$$

The curvature measures the lack of integrability of the horizontal distribution and plays a fundamental role in the theory of geometric phases (see [17] for a comprehensive treatment). In a local representation, the curvature can be written as

$$\mathcal{B}((g\xi, v), (g\eta, w)) = (B(r)(v, w)) = B^a_{\alpha\beta} v^\alpha w^\beta Ad_g e_a,$$

where $\{e_a\}_{a=1}^k$ is a basis of the Lie algebra $\mathfrak{g}$ and

$$B^a_{\alpha\beta} = \frac{\partial A^a_\alpha}{\partial r^\beta} - \frac{\partial A^a_\beta}{\partial r^\alpha} + c^a_{bc} A^b_\alpha A^c_\beta.$$

The c^a_{bc} are the *structure constants* of the Lie algebra defined by $[e_b, e_c] = c^a_{bc} e_a$.

An additional derivative operator related to a principal connection will appear in the derivations below. Let κ be a $\bigotimes_\nu \mathfrak{g}^*$-valued function on Q, $\kappa : Q \longrightarrow \bigotimes_\nu \mathfrak{g}^*$. Define then the *derivative of* κ *along* $\mathcal{A}$, $D\kappa : TQ \longrightarrow \bigotimes_\nu \mathfrak{g}^*$, by

$$D\kappa(\dot{q})(\xi_1, \ldots, \xi_\nu) = d\kappa(\dot{q})(\xi_1, \ldots, \xi_\nu) + \sum_{k=1}^{\nu} \kappa(q)(\xi_1, \ldots, ad_{\mathcal{A}\dot{q}} \xi_k, \ldots, \xi_\nu).$$

If the mapping κ is G-equivariant, $\kappa(g,r) = Ad^*_{g^{-1}}\kappa_{\rm loc}(r)$, where $\kappa_{\rm loc}(r) = \kappa(e,r)$, meaning

$$\kappa(g,r)(\xi_1,\ldots,\xi_\nu) = \kappa_{\rm loc}(r)(Ad_{g^{-1}}\xi_1,\ldots,Ad_{g^{-1}}\xi_\nu),$$

then one can see that

$$D\kappa(\dot g,\dot r) = Ad^*_{g^{-1}}D\kappa_{\rm loc}(\dot r).$$

In bundle coordinates, $D\kappa_{\rm loc}(\dot r)(\xi_1,\ldots,\xi_\nu) = (D\kappa_{\rm loc})_{\alpha a_1\ldots a_\nu}\dot r^\alpha \xi_1^{a_1}\ldots\xi_\nu^{a_\nu}$, where

$$(D\kappa_{\rm loc})_{\alpha a_1\ldots a_\nu} = \frac{\partial(\kappa_{\rm loc})_{a_1\ldots a_\nu}}{\partial r^\alpha} + \sum_{k=1}^{\nu}(\kappa_{\rm loc})_{a_1\ldots d_k\ldots a_\nu}A^e_\alpha c^{d_k}_{ea_k}.$$

2.3. Systems with symmetry. In the reduction of unconstrained mechanical systems with symmetry, there naturally arises a principal connection called the *mechanical* connection A^{mech}. Assume that the control system $(Q,\mathcal{G},V,\mathcal{F})$ is *invariant* under the action of a Lie group G, that is, $\Phi_g^*\mathcal{G} = \mathcal{G}$, $\Phi_g^*V = V$, and $\Phi_g^*F^i = F^i$ for $1 \le i \le m$ and all $g \in G$. (Note that it may happen that a particular element of the control system is invariant under the action of a larger Lie group H, $G \subseteq H$, but we are considering only Lie groups which leave invariant *all* the components of the problem.) The horizontal subspace of the mechanical connection is then given by the orthogonal complement of the vertical bundle $\mathcal{V}$ with respect to the kinetic energy metric $\mathcal{G}$, $\mathcal{H} = \mathcal{V}^\perp$. An explicit formula for its associated 1-form is the following. Define the *locked inertia tensor* at configuration $q \in Q$, $\mathcal{I}(q) : \mathfrak{g} \longrightarrow \mathfrak{g}^*$ by

$$\langle \mathcal{I}(q)\xi,\eta\rangle = \mathcal{G}(\xi_Q(q),\eta_Q(q)).$$

In local coordinates, this can be expressed as $\mathcal{I}(r,g) = Ad^*_{g^{-1}}I(r)Ad_{g^{-1}}$. $I(r)$, the *local form of* $\mathcal{I}$, has the interpretation of the inertia of the system when frozen at shape r. If we further defined the *momentum map* $J : TQ \longrightarrow \mathfrak{g}^*$ by $\langle J(\dot q),\xi\rangle = \langle \frac{\partial L}{\partial \dot q}(\dot q),\xi_Q(q)\rangle$, then the mechanical connection is just $A^{mech}(\dot q) = I(q)^{-1}J(\dot q)$.

The invariance of the metric and the potential function implies also that $L(g,r,\dot g,\dot r) = L(e,r,g^{-1}\dot g,\dot r) = \ell(r,\dot r,\xi)$, where $\xi = g^{-1}\dot g$. The function $\ell : TQ/G \longrightarrow \mathbb{R}$ is given by

$$\ell(r,\dot r,\xi) = \frac{1}{2}(\xi^T\ \dot r^T)\,\hat{\mathcal{G}}\begin{pmatrix}\xi\\ \dot r\end{pmatrix} - V(r),$$

where $\hat{\mathcal{G}}$ stands for the reduced metric [28]

$$\hat{\mathcal{G}} = \begin{pmatrix} I(r) & I(r)A(r)\\ A(r)^TI(r) & m(r)\end{pmatrix}. \tag{2.8}$$

Here A denotes the local form of the mechanical connection. This reduced metric is block diagonalized if we write it in terms of the shape variables $(r,\dot r)$ and the *locked body angular velocity*, $\Omega = \xi + A(r)\dot r$. Indeed, one can see that $\hat{\mathcal{G}}$ takes the form

$$\tilde{\mathcal{G}} = \begin{pmatrix} I(r) & 0\\ 0 & m(r) - A^T(r)I(r)A(r)\end{pmatrix} = \begin{pmatrix} I(r) & 0\\ 0 & \Delta(r)\end{pmatrix}.$$

We will see below that the terms I and Δ play a central role in deriving a local expression for the Levi–Civita affine connection.

The study of nonholonomic systems with symmetry has by now many contributions, starting from the work by Koiller on the kinematic case [16] and going through the use of the Hamiltonian formalism [2], Lagrangian reduction [4], the geometry of the tangent bundle [9, 10], and Poisson methods [23], among others. We review here some of the results found in [4, 28] for such systems which will be relevant for establishing later the decomposition for the nonholonomic affine connection.

Assume that the constrained mechanical control system is *invariant* under the action of a Lie group G, meaning that both $(Q, \mathcal{G}, V, \mathcal{F})$ and the constraint distribution $\mathcal{D}$ are invariant. Assume further that $\mathcal{D} + \mathcal{V} = TQ$ (the so-called *dimension assumption* [4]). We are interested in knowing which symmetry directions (i.e., tangent to the action of the Lie group) are compatible with the constraints. Consequently, we consider the intersection $S_q = \mathcal{V}_q \cap \mathcal{D}_q$ at each $q \in Q$. Since $S \subset \mathcal{V}$, we can consider a bundle $\mathfrak{g}^{\mathcal{D}} \longrightarrow Q$ whose fiber is given by $\mathfrak{g}^q = \{\xi \in \mathfrak{g} : \xi_Q(q) \in S_q\}$. The nonholonomic momentum map is then defined as

$$\begin{array}{rccccccc} J^{nh}: & TQ & \longrightarrow & \mathfrak{g}^{\mathcal{D}^*}, & & & & \\ & (q,\dot{q}) & \longmapsto & J^{nh}(q,\dot{q}): & \mathfrak{g}^q & \to & \mathbb{R}, \\ & & & & \xi^q & \longmapsto & \langle \frac{\partial L}{\partial \dot{q}}(\dot{q}), \xi_Q^q(q) \rangle. \end{array}$$

This momentum map can be used to "augment" the constraints and provide a principal connection on $Q \longrightarrow Q/G$, the so-called nonholonomic principal connection [4]. The horizontal subspace at $q \in Q$ of this connection is given by the orthogonal complement of S in the constraint distribution, $\mathcal{H}_q = S_q^{\perp} \cap \mathcal{D}_q$.

Alternatively, let $\{e_1(r), \ldots, e_s(r), e_{s+1}(r), \ldots, e_k(r)\} \in \mathfrak{g}$ be a basis of $\mathfrak{g}$ such that the first s elements span $\mathfrak{g}^{(r,e)}$ and both sets of generators are orthogonal in the kinetic energy metric restricted to $\mathcal{V}$. Denote by $\frac{\partial e_i}{\partial r^\alpha} = \sum_{a=1}^{k} \gamma_{i\alpha}^a e_a$ a notation which will be useful later. Define the momentum

$$p_i = \left\langle \frac{\partial \ell}{\partial \xi}, e_i(r) \right\rangle, \quad 1 \leq i \leq s.$$

Now consider the map

$$\begin{array}{rccc} A^{sym}: & T_qQ & \longrightarrow & S_q \\ & (q,\dot{q}) & \longmapsto & (\tilde{\mathcal{I}}^{-1}(q) J^{nh}(q,\dot{q}))_Q, \end{array}$$

where $\tilde{\mathcal{I}}(q) : \mathfrak{g}^{\mathcal{D}} \longrightarrow \mathfrak{g}^{\mathcal{D}^*}$ is the locked inertia tensor relative to $\mathfrak{g}^{\mathcal{D}}$. Notice that A^{sym} maps S onto itself. Additionally, let $A^{kin} : T_qQ \longrightarrow S_q^{\perp}$ be the orthogonal projection relative to the kinetic energy metric. The constraints plus the momentum can be written as

$$A^{kin}(q)\dot{q} = 0, \quad A^{sym}(q)\dot{q} = (\tilde{\mathcal{I}}^{-1}(q)p)_Q.$$

The nonholonomic connection 1-form is then given by

$$A^{nh} = A^{kin} + A^{sym}.$$

It is an instructive exercise to verify that A^{nh} indeed satisfies conditions (i) and (ii) defining a principal connection (cf. section 2.2). This principal connection plays a fundamental role in the reduction of nonholonomic systems with symmetry [4].

3. Decomposition of the Levi–Civita connection under symmetry. Given a mechanical control system with symmetry, it seems reasonable that the controllability tests can be simplified by taking into account the symmetry properties of the problem. In order to do that, we will obtain decompositions of the Levi–Civita connection and the nonholonomic affine connection according to the principal fiber bundle structure of the configuration space Q. This will be the subject of the following two sections.

Let $(Q, \mathcal{G}, V, \mathcal{F})$ be a simple mechanical control system invariant under the action of a Lie group G. The following simple lemma [14] will be helpful.

LEMMA 3.1. *The Levi–Civita connection associated to a left-invariant metric H on the Lie group G is given by*

$$\nabla^H_{g\xi} g\eta = \frac{1}{2} g \left([\xi, \eta] - \sharp_H \left(ad^*_\xi \flat_H \eta + ad^*_\eta \flat_H \xi \right) \right),$$

where $g\xi$ stands for $(L_g)_* \xi$ and so on. Consequently, the symmetric product associated to ∇^H takes the form

$$\langle g\xi : g\eta \rangle_H = -g \sharp_H \left(ad^*_\xi \flat_H \eta + ad^*_\eta \flat_H \xi \right).$$

Now we come to the main result of this section, where we derive the properties of the "connection within a connection." Emphasis is placed on the role of I, A, and Δ in determining $\nabla^{\mathcal{G}}$.

PROPOSITION 3.2. *Given G-invariant vector fields on Q, $X = (g\xi, v)$ and $Y = (g\eta, w)$, with $\xi(r)$, $\eta(r) \in \mathfrak{g}$ and v, $w \in TM$, the covariant derivative of Y along X can be expressed as*

$$\nabla^{\mathcal{G}}_X Y = g \left\{ \begin{pmatrix} \nabla^I_\Omega \Psi \\ \nabla^\Delta_v w \end{pmatrix} - \frac{1}{2} \begin{pmatrix} I^{-1}\mathbb{L} \\ \Delta^{-1}\mathbb{S} \end{pmatrix} \right\}, \tag{3.1}$$

where

$$\begin{aligned} \mathbb{L} &= -D(I\Omega)(\cdot, w) - D(I\Psi)(\cdot, v) + I([\Omega, \Psi] - [\xi, \eta] + \xi_r w - \eta_r v - A[v, w]) \\ &\quad + 2\, I(A(\nabla^{\mathcal{G}}_X Y)_M) \in \mathfrak{g}^*, \\ \mathbb{S} &= I(\Omega, B(w, \cdot)) + I(\Psi, B(v, \cdot)) + DI(\cdot)(\Omega, \Psi) \in T^*M, \end{aligned}$$

and $\Omega = \xi + Av$, $\Psi = \eta + Aw$, $\xi_r \equiv \frac{\partial \xi}{\partial r}$, $\eta_r \equiv \frac{\partial \eta}{\partial r}$.

Proof. As we have recalled above, the Levi–Civita connection can be characterized as the unique affine connection verifying (2.3). Let Z be a G-invariant vector field, $Z = (g\mu, u)$. The invariance of the metric implies

$$\mathcal{G}(Z, Y) = \hat{\mathcal{G}} \left((\mu, u), (\eta, w) \right) = \tilde{\mathcal{G}} \left((\Theta, u), (\Psi, w) \right),$$

where $\Theta = \mu + Aw$. The first three terms in (2.3) can be expanded in a similar way:

$$X(\mathcal{G}(Z, Y)) = X(\Theta^T I \Psi + u^T \Delta w) = v(\Theta^T I \Psi) + v(u^T \Delta w).$$

For the remaining ones, we have that

$$\begin{aligned} \mathcal{G}(X, [Z, Y]) &= \tilde{\mathcal{G}} \left((\Omega, v), ([\mu, \eta] + \eta_r u - w\mu_r + A[u, w], [u, w]) \right) \\ &= \Omega^T I[\mu, \eta] + \Omega^T I(\eta_r u - w\mu_r + A[u, w]) + v^T \Delta [u, w]. \end{aligned}$$

As a result, (2.3) can be written as $2\,\mathcal{G}(Z, \nabla^{\mathcal{G}}_X Y) = \langle(\delta, \gamma), (\mu, u)\rangle$, where $\delta = \delta_1 + \delta_2$, $\gamma = \gamma_1 + \gamma_2$, and

$$
\begin{aligned}
\delta_1 &= \xi^T I[\cdot, \eta] + \eta^T I[\cdot, \xi] - \cdot^T I[\eta, \xi],\\
\delta_2 &= \cdot^T v(I\Psi) + \cdot^T w(I\Omega) + (Av)^T I[\cdot, \eta] + (Aw)^T I[\cdot, \xi] - \cdot^T I(\xi_r w - \eta_r v + A[w, v]),\\
\gamma_1 &= v(\cdot^T \Delta w) + w(\cdot^T \Delta v) - \cdot(w^T \Delta v) + v^T \Delta[\cdot, w] + w^T \Delta[\cdot, v] - \cdot^T \Delta[w, v],\\
\gamma_2 &= v((A\cdot)^T I\Psi) + w((A\cdot)^T I\Omega) - \cdot(\Psi^T I\Omega) + \Omega^T I(\eta_r \cdot + A[\cdot, w])\\
&\quad + \Psi^T I(\xi_r \cdot + A[\cdot, v]) - (A\cdot)^T I([\eta, \xi] + \xi_r w - \eta_r v + A[w, v]).
\end{aligned}
$$

On the other hand, we have that

$$
2\,\mathcal{G}(Z, \nabla^{\mathcal{G}}_X Y) = 2\,(\mu^T, u^T) \begin{pmatrix} I & IA \\ A^T I & m \end{pmatrix} \begin{pmatrix} (\nabla^{\mathcal{G}}_X Y)_{\mathfrak{g}} \\ (\nabla^{\mathcal{G}}_X Y)_M \end{pmatrix}.
$$

As both expansions for $\mathcal{G}(Z, \nabla^{\mathcal{G}}_X Y)$ are valid for any Z, we can conclude that

$$
\begin{aligned}
2 \begin{pmatrix} (\nabla^{\mathcal{G}}_X Y)_{\mathfrak{g}} \\ (\nabla^{\mathcal{G}}_X Y)_M \end{pmatrix} &= \begin{pmatrix} I & IA \\ A^T I & m \end{pmatrix}^{-1} \begin{pmatrix} \delta \\ \gamma \end{pmatrix}\\
&= \begin{pmatrix} I^{-1} + A\Delta^{-1}A^T & -A\Delta^{-1} \\ -\Delta^{-1}A^T & \Delta^{-1} \end{pmatrix} \begin{pmatrix} \delta \\ \gamma \end{pmatrix}.
\end{aligned} \tag{3.2}
$$

Noting that $\delta_1 = 2I\nabla^I_\xi \eta$ (see Lemma 3.1) and $\gamma_1 = 2\Delta\nabla^{\Delta}_v w$, we can further develop the right-hand side of (3.2) as

$$
\begin{aligned}
&\begin{pmatrix} I^{-1} & 0 \\ 0 & \Delta^{-1} \end{pmatrix} \begin{pmatrix} \delta \\ \gamma \end{pmatrix} + \begin{pmatrix} A\Delta^{-1}A^T & -A\Delta^{-1} \\ -\Delta^{-1}A^T & 0 \end{pmatrix} \begin{pmatrix} \delta \\ \gamma \end{pmatrix}\\
&= 2 \begin{pmatrix} \nabla^I_\xi \eta \\ \nabla^{\Delta}_v w \end{pmatrix} + \begin{pmatrix} I^{-1} & 0 \\ 0 & \Delta^{-1} \end{pmatrix} \left\{ \begin{pmatrix} \delta_2 \\ \gamma_2 \end{pmatrix} + \begin{pmatrix} IA\Delta^{-1}A^T & -IA\Delta^{-1} \\ -A^T & 0 \end{pmatrix} \begin{pmatrix} \delta \\ \gamma \end{pmatrix} \right\}.
\end{aligned}
$$

In this way, we get

$$
\begin{pmatrix} (\nabla^{\mathcal{G}}_X Y)_{\mathfrak{g}} \\ (\nabla^{\mathcal{G}}_X Y)_M \end{pmatrix} = \begin{pmatrix} \nabla^I_\xi \eta \\ \nabla^{\Delta}_v w \end{pmatrix} - \frac{1}{2} \begin{pmatrix} I^{-1} & 0 \\ 0 & \Delta^{-1} \end{pmatrix} \begin{pmatrix} \mathbb{L}' \\ \mathbb{S} \end{pmatrix},
$$

where $\mathbb{L}' = -\delta_2 - IA\Delta^{-1}\mathbb{S} + IA\Delta^{-1}\gamma_1$ and $\mathbb{S} = A^T\delta - \gamma_2$. To complete the proof, we have only to identify these terms in a more geometrical manner, which we do in the following.

We begin with $\mathbb{S}$. Noting that

$$
A^b_\beta \frac{\partial v^\beta}{\partial r^\alpha} + \frac{\partial \xi^b}{\partial r^\alpha} - \frac{\partial \Omega^b}{\partial r^\alpha} = -\frac{\partial A^b_\beta}{\partial r^\alpha} v^\beta,
$$

we can rewrite γ_2 as

$$
\begin{aligned}
\gamma_2 = {} & v^\beta A^b_\alpha \frac{\partial (I\Psi)_b}{\partial r^\beta} + w^\beta A^b_\alpha \frac{\partial (I\Omega)_b}{\partial r^\beta} - \Psi^b \frac{\partial I_{ba}}{\partial r^\alpha} \Omega^a\\
& + (I\Psi)_b \left\{ \frac{\partial A^b_\alpha}{\partial r^\beta} - \frac{\partial A^b_\beta}{\partial r^\alpha} \right\} v^\beta + (I\Omega)_b \left\{ \frac{\partial A^b_\alpha}{\partial r^\beta} - \frac{\partial A^b_\beta}{\partial r^\alpha} \right\} w^\beta\\
& - A^b_\alpha I_{ba} c^a_{de} \eta^d \xi^e - A^b_\alpha I_{ba} \left\{ \frac{\partial \xi^a}{\partial r^\beta} w^\beta - \frac{\partial \eta^a}{\partial r^\beta} v^\beta + A^a_\beta [w, v]^\beta \right\}.
\end{aligned}
$$

Substituting the latter into the expression for $\mathbb{S}$, one obtains after some computations

$$\begin{aligned}-\mathbb{S} &= (I\Psi)_b \left\{ \frac{\partial A^b_\alpha}{\partial r^\beta} - \frac{\partial A^b_\beta}{\partial r^\alpha} + A^c_\beta E^b_{\alpha c} \right\} v^\beta + (I\Omega)_b \left\{ \frac{\partial A^b_\alpha}{\partial r^\beta} - \frac{\partial A^b_\beta}{\partial r^\alpha} + A^c_\beta E^b_{\alpha c} \right\} w^\beta \\ &\quad -\Psi^b \frac{\partial I_{ba}}{\partial r^\alpha}\Omega^a - \Omega^d I_{db} E^b_{\alpha e}\Psi^e - \Psi^d I_{db} E^b_{\alpha e}\Omega^e \\ &= -I(\Psi, B(v,\cdot)) - I(\Omega, B(w,\cdot)) - DI(\cdot)(\Omega,\Psi),\end{aligned}$$

where $E^b_{\alpha c} = c^b_{dc} A^d_\alpha$. Now we turn our attention to $\mathbb{L}'$. Note that

$$\begin{aligned}\mathbb{L}' &= IA\Delta^{-1}(\gamma_1 - \mathbb{S}) - \delta_2 \\ &= 2IA\left(\nabla^\Delta_v w - \frac{1}{2}\Delta^{-1}\mathbb{S}\right) - \delta_2 = 2IA(\nabla^{\mathcal{G}}_X Y)_M - \delta_2.\end{aligned}$$

Moreover, we have

$$\begin{aligned}\delta_2 = v^\alpha \frac{\partial (I\Psi)_a}{\partial r^\alpha} + w^\alpha \frac{\partial (I\Omega)_a}{\partial r^\alpha} + A^d_\alpha v^\alpha I_{db} c^b_{ae}\eta^e + A^d_\alpha w^\alpha I_{db} c^b_{ae}\xi^e \\ -I_{ba}\left\{ \frac{\partial \xi^b}{\partial r^\beta} w^\beta - \frac{\partial \eta^b}{\partial r^\beta} v^\beta + A^b_\beta [w,v]^\beta \right\}.\end{aligned}$$

Adding and subtracting $(I\Psi)_b E^b_{\alpha a} v^\alpha$ and $(I\Omega)_b E^b_{\alpha a} w^\alpha$ and regrouping, we obtain

$$\begin{aligned}\delta_2 = D(I\Psi)(\cdot, v) + D(I\Omega)(\cdot, w) + 2I\nabla^I_\Omega \Psi - 2I\nabla^I_\xi \eta \\ -I(\cdot, [\Omega,\Psi]) + I(\cdot, [\xi,\eta]) - I(\cdot, \xi_r w - \eta_r v + A[w,v]).\end{aligned}$$

Finally, we can write

$$(\nabla^{\mathcal{G}}_X Y)_{\mathfrak{g}} = \nabla^I_\xi \eta - \frac{1}{2} I^{-1}\mathbb{L}' = \nabla^I_\Omega \Psi - \frac{1}{2} I^{-1}\mathbb{L},$$

where $\mathbb{L}$ is as above. □

As a consequence of this proposition, we have the following interesting result.

COROLLARY 3.3. *The symmetric product associated to the Levi–Civita connection $\nabla^{\mathcal{G}}$ of two G-invariant vector fields, $X = (g\xi, v)$ and $Y = (g\eta, w)$, is given by*

$$\langle X : Y\rangle_{\mathcal{G}} = g\left\{ \begin{pmatrix} \langle \Omega : \Psi\rangle_I \\ \langle v : w\rangle_\Delta \end{pmatrix} - \begin{pmatrix} I^{-1}\mathbb{L}^s \\ \Delta^{-1}\mathbb{S} \end{pmatrix} \right\}, \tag{3.3}$$

where

$$\begin{aligned}\mathbb{L}^s &= -D(I\Omega)(\cdot, w) - D(I\Psi)(\cdot, v) + IA\left(\langle v : w\rangle_\Delta - \Delta^{-1}\mathbb{S}\right) \in \mathfrak{g}^*, \\ \mathbb{S} &= I(\Omega, B(w,\cdot)) + I(\Psi, B(v,\cdot)) + DI(\cdot)(\Omega,\Psi) \in T^*M,\end{aligned}$$

and $\langle\cdot:\cdot\rangle_I$, $\langle\cdot:\cdot\rangle_\Delta$ denote the symmetric products defined by the Levi–Civita connections ∇^I and ∇^Δ, respectively.

We shall return to these results in section 6 in computing the symmetric product in specific examples.

4. Decomposition of the nonholonomic affine connection under symmetry. Let $(Q, \mathcal{G}, V, \mathcal{F}, \mathcal{D})$ be a constrained mechanical control system invariant under the action of a Lie group G. As expected, the invariance of the Levi–Civita connection and the nonholonomic distribution $\mathcal{D}$ can be combined to find a decomposition of the nonholonomic affine connection similar to that of Proposition 3.2.

First, notice that if $\mathcal{D}$ is generated by a basis of G-invariant vector fields X_i, $1 \leq i \leq n-l$, the projector $\mathcal{P} : TQ \longrightarrow \mathcal{D}$ with respect to the orthogonal decomposition $TQ = \mathcal{D} \oplus \mathcal{D}^\perp$ is given by

$$\mathcal{P}(Z) = \sum_{i,j} C^{ij} \mathcal{G}(X_i, Z) X_j, \quad Z \in \mathfrak{X}(Q),$$

where (C^{ij}) is the inverse matrix of $(C_{ij} = \mathcal{G}(X_i, X_j))$. A geometrically revealing choice of generators of $\mathcal{D}$ making use of the exposition in section 2.3 is the following. Recall that the nonholonomic principal connection A^{nh} induces a decomposition of the tangent bundle, $TQ = \mathcal{H} \oplus \mathcal{V}$. This, in particular, implies that

$$\mathcal{D} = \mathcal{H} \oplus S.$$

On the one hand, we know that $S_{(r,e)} = \operatorname{span}\{e_1(r)_Q, \ldots, e_s(r)_Q\}$. Furthermore, the generators of $\mathcal{H}_{(r,e)}$ are of the form $(-\mathbb{A}\dot{r}, \dot{r})$, where $\mathbb{A}$ denotes the local form of A^{nh}. Hence we have that

$$\mathcal{D}_{(r,g)} = g\mathcal{D}_{(r,e)} = g \operatorname{span}\left\{(-\mathbb{A}\dot{r}, \dot{r}), (e_i, 0)\right\}.$$

For these vector fields, we compute

$$\begin{aligned}
\mathcal{G}\left(g\left(e_i, 0\right), g\left(e_j, 0\right)\right) &= e_i^T I e_j = e_i^T \tilde{I} e_j, \\
\mathcal{G}\left(g\left(-\mathbb{A}(\dot{r}), \dot{r}\right), g\left(e_j, 0\right)\right) &= -(\mathbb{A}\dot{r})^T I e_j + (A\dot{r})^T I e_j = (\tilde{A}\dot{r})^T I e_j = 0, \\
\mathcal{G}\left(g\left(-\mathbb{A}(\dot{r}), \dot{r}\right), g\left(-\mathbb{A}(\dot{r}), \dot{r}\right)\right) &= (\mathbb{A}\dot{r})^T I \mathbb{A}\dot{r} - (\mathbb{A}\dot{r})^T I A\dot{r} - (A\dot{r})^T I \mathbb{A}\dot{r} + \dot{r}^T m \dot{r} \\
&= \dot{r}^T (m + \mathbb{A}^T I \mathbb{A} - \mathbb{A}^T I A - A^T I \mathbb{A}) \dot{r} = \dot{r}^T \tilde{\Delta} \dot{r},
\end{aligned}$$

where $\tilde{A} = A - \mathbb{A}$, $\tilde{\Delta} = m - A^T I A + \tilde{A}^T I \tilde{A}$, and we have used the fact that $\tilde{A}\dot{r} \in S^\perp$. Hence we can write the matrix C as

$$C = \begin{pmatrix} \tilde{I} & 0 \\ 0 & \tilde{\Delta} \end{pmatrix}.$$

Now we are in a position to prove the following result.

PROPOSITION 4.1. *Given G-invariant vector fields, $X = (g\xi, v) \in TQ$, $Y = (g\eta, w) \in \mathcal{D}$ on Q, with $\xi(r)$, $\eta(r) \in \mathfrak{g}$ and v, $w \in TM$, the nonholonomic affine connection $\overline{\nabla}$ can be expressed as*

$$\overline{\nabla}_X Y = g\left\{ \begin{pmatrix} A^{sym}(\nabla^I_{\bar{\Omega}} \bar{\Psi}) \\ \nabla^{\tilde{\Delta}}_v w \end{pmatrix} - \frac{1}{2} \begin{pmatrix} \tilde{I}^{-1} \tilde{\mathbb{L}} + 2\mathbb{A}(\overline{\nabla}_X Y)_M \\ \tilde{\Delta}^{-1} \tilde{\mathbb{S}} \end{pmatrix} \right\}, \tag{4.1}$$

where

$$\begin{aligned}
\tilde{\mathbb{L}} &= -\mathbb{D}(I\bar{\Omega})(\cdot, w) - \mathbb{D}(I\bar{\Psi})(\cdot, v) + I(\tilde{A}v, \gamma.w - [\cdot, \eta]) + I(\tilde{A}w, \gamma.v - [\cdot, \xi]) \\
&\quad + I([\bar{\Omega}, \bar{\Psi}] - [\xi, \eta] + \xi_r w - \eta_r v - \mathbb{A}[v, w]) \in \mathfrak{g}^{\mathcal{D}^*}, \\
\tilde{\mathbb{S}} &= I(\bar{\Psi}, B(v, \cdot)) + I(\bar{\Omega}, B(w, \cdot)) + I(\tilde{A}w, \mathbb{B}(v, \cdot)) + I(\tilde{A}v, \mathbb{B}(w, \cdot)) \\
&\quad - D(I\bar{\Psi})(\tilde{A}\cdot, v) - D(I\bar{\Omega})(\tilde{A}\cdot, w) + \mathbb{D}I(\cdot)(\bar{\Omega} + \tilde{A}v, \bar{\Psi} + \tilde{A}w) - \mathbb{D}I(\cdot)(\tilde{A}v, \tilde{A}w) \\
&\quad - I([\xi, \eta], \tilde{A}\cdot) - I(\eta_r v - \xi_r v, \tilde{A}\cdot) - I(\mathbb{A}[v, w], \tilde{A}\cdot) \in T^*M,
\end{aligned}$$

and $\mathbb{D}$, $\mathbb{B}$ *denote, respectively, the local forms of the derivative along and the curvature of the nonholonomic connection* A^{nh} *and* $\bar{\Omega} = \xi + \mathbb{A}v$, $\bar{\Psi} = \eta + \mathbb{A}w$.

Proof. Since $Y \in \mathcal{D}$, $\overline{\nabla}_X Y = \mathcal{P}(\nabla^{\mathcal{G}}_X Y) = \sum C^{ij}\mathcal{G}(X_i, \nabla^{\mathcal{G}}_X Y)X_j$. We first compute

$$\mathcal{G}\left(g\left(e_i, 0\right), \nabla^{\mathcal{G}}_X Y\right) = e_i^T I\left\{(\nabla^{\mathcal{G}}_X Y)_{\mathfrak{g}} + A(\nabla^{\mathcal{G}}_X Y)_M\right\}, \tag{4.2}$$

$$\begin{aligned}\mathcal{G}\left(g\left(-\mathbb{A}\dot{r}, \dot{r}\right), \nabla^{\mathcal{G}}_X Y\right) &= (\tilde{A}\dot{r})^T I(\nabla^{\mathcal{G}}_X Y)_{\mathfrak{g}} + \dot{r}^T(m - A^T I\mathbb{A})(\nabla^{\mathcal{G}}_X Y)_M \\ &= (\tilde{A}\dot{r})^T I\left\{(\nabla^{\mathcal{G}}_X Y)_{\mathfrak{g}} + A(\nabla^{\mathcal{G}}_X Y)_M\right\} + \dot{r}^T\Delta(\nabla^{\mathcal{G}}_X Y)_M.\end{aligned} \tag{4.3}$$

Let us denote $(\nabla^{\mathcal{G}}_X Y)_{\mathfrak{g}} + A(\nabla^{\mathcal{G}}_X Y)_M = \widetilde{\nabla^{\mathcal{G}}_X Y}$ for brevity. In terms of $\bar{\Omega}$, $\bar{\Psi}$ and using Proposition 3.2, it can be expanded as

$$\begin{aligned}\widetilde{\nabla^{\mathcal{G}}_X Y} = \nabla^I_{\bar{\Omega}}\bar{\Psi} - \frac{1}{2}I^{-1}\Big\{ &- \mathbb{D}(I\bar{\Omega})(\cdot, w) - \mathbb{D}(I\tilde{A}v)(\cdot, w) - \mathbb{D}(I\bar{\Psi})(\cdot, v) \\ &- \mathbb{D}(I\tilde{A}w)(\cdot, v) - I(\tilde{A}w, [\cdot, \bar{\Omega}]) - I(\tilde{A}v, [\cdot, \bar{\Psi}]) \\ &+ I([\bar{\Omega}, \bar{\Psi}] - [\xi, \eta] + \xi_r w - \eta_r v - A[v, w], \cdot)\Big\}.\end{aligned}$$

Before plugging this expression into (4.2), notice that

$$-\mathbb{D}(I\tilde{A}v)(e_i, w) - I(\tilde{A}v, [e_i, \bar{\Psi}]) = I(\tilde{A}v, \gamma_i w - [e_i, \eta]),$$

where we have used the facts that $e_i \in S$ and $\tilde{A}v \in S^{\perp}$. After substituting, we find that (4.2) can be expressed as $\mathcal{G}\left(g\left(e_i, 0\right), \nabla^{\mathcal{G}}_X Y\right) = \langle I(\nabla^I_{\bar{\Omega}}\bar{\Psi}, \cdot) - \frac{1}{2}\tilde{\mathbb{L}}, e_i\rangle$, where

$$\begin{aligned}\tilde{\mathbb{L}} = &-\mathbb{D}(I\bar{\Omega})(\cdot, w) - \mathbb{D}(I\bar{\Psi})(\cdot, v) + I(\tilde{A}v, \gamma.w - [\cdot, \eta]) + I(\tilde{A}w, \gamma.v - [\cdot, \xi]) \\ &+ I([\bar{\Omega}, \bar{\Psi}] - [\xi, \eta], \cdot) - I(\eta_r v - \xi_r w + \mathbb{A}[v, w], \cdot).\end{aligned}$$

On the other hand, it is easy to see that

$$\begin{aligned}\Delta(\nabla^{\mathcal{G}}_X Y)_M &= \Delta\nabla^{\Delta}_v w - \frac{1}{2}\mathbb{S} \\ &= \tilde{\Delta}\nabla^{\tilde{\Delta}}_v w - D\nabla^D_v w - \frac{1}{2}\mathbb{S} = \tilde{\Delta}\nabla^{\tilde{\Delta}}_v w - \left(D\nabla^D_v w + \frac{1}{2}\mathbb{S}\right),\end{aligned}$$

where $D = \tilde{A}^T I\tilde{A}$ and $D\nabla^D_v w$ is a shorthand notation to denote the expression (2.3) for the symmetric tensor D. Then we can rewrite (4.3) as

$$\dot{r}^T\left(\tilde{\Delta}\nabla^{\tilde{\Delta}}_v w - \left(D\nabla^D_v w + \frac{1}{2}\mathbb{S} - \tilde{A}^T I\widetilde{\nabla^{\mathcal{G}}_X Y}\right)\right).$$

Therefore, $\overline{\nabla}_X Y$ becomes

$$\overline{\nabla}_X Y = \mathcal{P}(\nabla^{\mathcal{G}}_X Y) = \begin{pmatrix} g(\overline{\nabla}_X Y)_{\mathfrak{g}} \\ (\overline{\nabla}_X Y)_M \end{pmatrix},$$

with

$$\begin{aligned}(\overline{\nabla}_X Y)_{\mathfrak{g}} &= \tilde{I}^{-1}\left\{I(\nabla^I_{\bar{\Omega}}\bar{\Psi}, \cdot) - \frac{1}{2}\tilde{\mathbb{L}}\right\} - \mathbb{A}(\overline{\nabla}_X Y)_M \\ &= A^{sym}(\nabla^I_{\bar{\Omega}}\bar{\Psi}) - \frac{1}{2}\tilde{I}^{-1}\tilde{\mathbb{L}} - \mathbb{A}(\overline{\nabla}_X Y)_M, \\ (\overline{\nabla}_X Y)_M &= \nabla^{\tilde{\Delta}}_v w - \tilde{\Delta}^{-1}\left(D\nabla^D_v w + \frac{1}{2}\mathbb{S} - \tilde{A}^T I\widetilde{\nabla^{\mathcal{G}}_X Y}\right),\end{aligned}$$

where we have used the fact that $A^{sym}(\zeta) \equiv A^{sym}(\zeta_Q(e,r)) = \tilde{I}^{-1} I(\zeta)$ for $\zeta \in \mathfrak{g}$. To end the proof, let us write explicitly the terms in $(\overline{\nabla}_X Y)_M$. Adding and subtracting terms in the expression for $2\, D\nabla^D_v w$, we can find that

$$\begin{aligned}
2\, D\nabla^D_v w &= v^\beta \frac{\partial \tilde{A}^a_\alpha}{\partial r^\beta} I_{ab} \tilde{A}^b_\gamma w^\gamma + D(I\tilde{A}w)(\tilde{A}\cdot, v) - I(\tilde{A}v, [Aw, \tilde{A}\cdot]) \\
&\quad + w^\beta \frac{\partial \tilde{A}^a_\alpha}{\partial r^\beta} I_{ab} \tilde{A}^b_\gamma v^\gamma + D(I\tilde{A}v)(\tilde{A}\cdot, w) - I(\tilde{A}w, [Av, \tilde{A}\cdot]) \\
&\quad - w^\beta v^\gamma \frac{\partial \tilde{A}^a_\beta I_{ab} \tilde{A}^b_\gamma}{\partial r^\alpha} + \tilde{A}^T I \tilde{A}[v, w] \\
&= D(I\tilde{A}w)(\tilde{A}\cdot, v) + D(I\tilde{A}v)(\tilde{A}\cdot, w) + I(\tilde{A}w)B(\cdot, v) + I(\tilde{A}w)\mathbb{B}(v, \cdot) \\
&\quad + I(\tilde{A}v)B(\cdot, w) + I(\tilde{A}v)\mathbb{B}(w, \cdot) - \mathbb{D}I(\tilde{A}v, \tilde{A}w) + \tilde{A}^T I \tilde{A}[v, w].
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\mathbb{S} &= I(\Omega, B(w, \cdot)) + I(\Psi, B(v, \cdot)) + DI(\cdot)(\Omega, \Psi) \\
&= I(\Omega, B(w, \cdot)) + I(\Psi, B(v, \cdot)) + \mathbb{D}I(\cdot)(\Omega, \Psi) + I(\Omega, [\tilde{A}\cdot, \Psi]) + I(\Psi, [\tilde{A}\cdot, \Omega]),
\end{aligned}$$

and the term $\tilde{A}^T I \widetilde{\nabla^{\mathcal{G}}_X Y}$ can be written as

$$\begin{aligned}
-\tilde{A}^T I \widetilde{\nabla^{\mathcal{G}}_X Y} = &- \tilde{A}^T I \nabla^I_\Omega \Psi + \frac{1}{2}(-D(I\Psi)(\tilde{A}\cdot, v) - D(I\Omega)(\tilde{A}\cdot, w) \\
&+ I([\Omega, \Psi] - [\xi, \eta], \tilde{A}\cdot) - I(\eta_r v - \xi_r v, \tilde{A}\cdot) - \tilde{A}^T I A[v, w]).
\end{aligned}$$

Summing up these terms, we get the expression for $\tilde{\mathbb{S}}$ stated in the proposition. □

COROLLARY 4.2. *The symmetric product associated to $\overline{\nabla}$ of two G-invariant vector fields, $X = (g\xi, v) \in \mathcal{D}$ and $Y = (g\eta, w) \in \mathcal{D}$, is given by*

(4.4)
$$\langle X : Y \rangle = g \left\{ \begin{pmatrix} A^{sym}(\langle \bar{\Omega} : \bar{\Psi} \rangle_I) \\ \langle v : w \rangle_{\tilde{\Delta}} \end{pmatrix} - \begin{pmatrix} \tilde{I}^{-1} \tilde{\mathbb{L}}^s + \mathbb{A}\left(\langle v : w \rangle_{\tilde{\Delta}} - \tilde{\Delta}^{-1} \tilde{\mathbb{S}}^s \right) \\ \tilde{\Delta}^{-1} \tilde{\mathbb{S}}^s \end{pmatrix} \right\},$$

where

$$\begin{aligned}
\tilde{\mathbb{L}}^s &= -\mathbb{D}(I\bar{\Omega})(\cdot, w) - \mathbb{D}(I\bar{\Psi})(\cdot, v) + I(\tilde{A}v, \gamma.w - [\cdot, \eta]) + I(\tilde{A}w, \gamma.v - [\cdot, \xi]) \in \mathfrak{g}^{\mathcal{D}^*}, \\
\tilde{\mathbb{S}}^s &= I(\bar{\Psi}, B(v, \cdot)) + I(\bar{\Omega}, B(w, \cdot)) + I(\tilde{A}w, \mathbb{B}(v, \cdot)) + I(\tilde{A}v, \mathbb{B}(w, \cdot)) - D(I\bar{\Psi})(\tilde{A}\cdot, v) \\
&\quad -D(I\bar{\Omega})(\tilde{A}\cdot, w) + \mathbb{D}I(\cdot)(\bar{\Omega} + \tilde{A}v, \bar{\Psi} + \tilde{A}w) - \mathbb{D}I(\cdot)(\tilde{A}v, \tilde{A}w) \in T^*M,
\end{aligned}$$

and $\langle \cdot : \cdot \rangle_{\tilde{\Delta}}$ denotes the symmetric product defined by the Levi–Civita connection $\nabla^{\tilde{\Delta}}$.

5. Controllability analysis. The point in the approach of Lewis and Murray to simple mechanical control systems is precisely to know what is happening to configurations, rather than to states, since, in many of these systems, configurations may be controlled but not configurations and velocities at the same time. The basic question they pose is "what is the set of configurations that are attainable from a given configuration starting from rest?"

Consider the control equation

$$\nabla_{\dot{c}(t)} \dot{c}(t) = \sum_{i=1}^m u_i(t) Y_i(c(t)), \tag{5.1}$$

where the affine connection ∇ can be either the Levi–Civita affine connection associated to a kinetic energy metric or the nonholonomic affine connection for a constrained system. (Recall that, in the latter case, we select $\dot{c}(0) \in \mathcal{D}$ and Y_i denotes the projection by $\mathcal{P}$ to $\mathcal{D}$ of the ith input vector field.) Notice that we are considering now that $V \equiv 0$. The absence of the potential makes the picture considerably more clear while capturing the essential aspects of the analysis. On the other hand, a potential function could be incorporated to the controllability tests along the lines of [21].

Take $q_0 \in Q$, and let $U \subset Q$ be a neighborhood of q_0. Define

$$\mathcal{R}_Q^U(q_0, T) = \{q \in Q \,|\, \text{there exists a solution } (c, u) \text{ of (5.1) such that } \dot{c}(0) = 0_{q_0}, c(t) \in U \text{ for } t \in [0, T], \text{ and } \dot{c}(T) \in T_q Q\},$$

and denote $\mathcal{R}_Q^U(q_0, \leq T) = \cup_{0 \leq t \leq T} \mathcal{R}_Q^U(q_0, t)$.

We shall focus our attention on the following notions of accessibility and controllability [21].

DEFINITION 5.1. *The system* (5.1) *is* locally configuration accessible (LCA) at $q_0 \in Q$ *if there exists* $T > 0$ *such that* $\mathcal{R}_Q^U(q_0, \leq t)$ *contains a nonempty open set of* Q *for all neighborhoods* U *of* q_0 *and all* $0 \leq t \leq T$. *If this holds for any* $q_0 \in Q$, *then the system is called* LCA.

DEFINITION 5.2. *The system* (5.1) *is* small-time locally configuration controllable (STLCC) at $q_0 \in Q$ *if there exists* $T > 0$ *such that* $\mathcal{R}_Q^U(q_0, \leq t)$ *contains a nonempty open set of* Q *to which* q_0 *belongs for all neighborhoods* U *of* q_0 *and all* $0 \leq t \leq T$. *If this holds for any* $q_0 \in Q$, *then the system is called* STLCC.

Given the input vector fields $\mathcal{Y} = \{Y_1, \ldots, Y_m\}$, let us denote by $\overline{Sym}(\mathcal{Y})$ the distribution obtained by closing the set $\mathcal{Y}$ under the symmetric product and by $\overline{Lie}(\mathcal{Y})$ the involutive closure of $\mathcal{Y}$. With these ingredients, one can prove the following theorem.

THEOREM 5.3 (see [21]). *The control system* (5.1) *is LCA at* q *if* $\overline{Lie}(\overline{Sym}(\mathcal{Y}))_q = T_q Q$.

If P is a symmetric product of vector fields in $\mathcal{Y}$, we let $\gamma_i(P)$ denote the number of occurrences of Y_i in P. The *degree* of P will be $\gamma_1(P) + \cdots + \gamma_m(P)$. We say that P is *bad* if $\gamma_i(P)$ is even for each $1 \leq i \leq m$. Otherwise, we say that P is *good*. The following theorem gives sufficient conditions for STLCC.

THEOREM 5.4. *Suppose that the system* (5.1) *is LCA at* q *and that every bad symmetric product* P *at* q *in* $\mathcal{Y}$ *can be written as a linear combination of good symmetric products at* q *of lower degree than* P. *Then it is STLCC at* q.

Remark 5.1. This theorem was proved in [21] as an application to mechanical systems of previous work by Sussmann [32] on general control systems with drift. There has been some effort in trying to obtain sufficient *and* necessary conditions for configuration controllability. A conjecture that remains open is that the system (5.1) is STLCC at q if and only if there exists a basis of vector fields generating the input distribution which satisfies the sufficient conditions of the theorem. Lewis [18] proved the validity of the conjecture for the one-input case. Recently, Cortés and Martínez [11] have proved that it is also valid for underactuated systems by one control.

The exposed controllability analysis can be further refined for mechanical control systems with symmetry, taking into account the results of the previous sections. Assume that the control system (5.1) is invariant under the action of a Lie group G. Let us denote by $\mathfrak{B} = \{B_1, \ldots, B_m\}$ the representatives of the input vector fields

$\mathcal{Y} = \{Y_1, \ldots, Y_m\}$ at $\mathfrak{g} \times TM$, that is,

$$Y_i(r,g) = gB_i(r,e) = g \begin{pmatrix} \xi_i(r) \\ v_i \end{pmatrix}, \quad 1 \leq i \leq m.$$

Due to the invariance of the system, we have that $\langle Y_i : Y_j \rangle = \langle gB_i : gB_j \rangle \equiv g\langle B_i : B_j \rangle$ for all $1 \leq i, j \leq m$. The explicit expression in bundle coordinates for this symmetric product is given by Corollaries 3.3 and 4.2. Note also that the Lie brackets $[Y_i, Y_j]$ can be written as

$$[Y_i, Y_j] \equiv g[B_i, B_j] = g \begin{pmatrix} [\xi_i, \xi_j]_{\mathfrak{g}} + \frac{\partial \xi_j}{\partial r} v_i - \frac{\partial \xi_i}{\partial r} v_j \\ [v_i, v_j]_M \end{pmatrix}.$$

As a result, we have the following version of the former results.

THEOREM 5.5. *Let the control system* (5.1) *be invariant under the action of a Lie group G.*

(i) *The system is LCA at $q = (r,g) \in Orb_G(r,e)$ if $\overline{Lie}(\overline{Sym}(\mathfrak{B}))_{(r,e)} = \mathfrak{g} \times T_r M$.*

(ii) *Suppose that the system is LCA at (r,e) and that every bad symmetric product P at (r,e) in $\mathfrak{B}$ can be written as a linear combination of good symmetric products at (r,e) of lower degree than P. Then* (5.1) *is STLCC at $q \in Orb_G(r,e)$.*

These simplified tests of the accessibility and controllability properties of mechanical control systems under symmetry are indeed quite useful in practical examples, since they remove completely the dependence on the Lie group elements $g \in G$ from the computations. In examples such as the blimp, the underwater vehicle, the snakeboard, and the roller racer, where symmetry plays an important role, this property may be a definitive advantage.

An additional important simplification from the computational point of view stems from the fact that, for many dynamic robotic locomotion systems, the set of inputs consists of the full tangent bundle of the shape space M. This essentially corresponds to the observation that the system can adjust its shape as desired. For such problems, we can state the following result.

THEOREM 5.6. *Let the control system* (5.1) *be invariant under the action of a Lie group G. Additionally assume that the system is fully actuated in the shape space; i.e., the set of input forces consists of $F^1 = dr^1, \ldots, F^m = dr^m$, where m now also denotes the dimension of M. Then the locked body angular velocities of the input vector fields all vanish: $\Omega_i = 0$, $1 \leq i \leq m$. Moreover, in the presence of nonholonomic constraints, the projections of the input vector fields to $\mathcal{D}$ also have $\bar{\Omega}_i = 0$, $1 \leq i \leq m$.*

Proof. It is not difficult to verify that the input vector fields are of the form $\left(-gA\Delta^{-1}\dot{r}, \Delta^{-1}\dot{r}\right)$. Then $\Omega_i = 0$ follows. On the other hand, their projections to the constraint distribution $\mathcal{D}$ can be written as $(-g\mathbb{A}\tilde{\Delta}^{-1}\dot{r}, \tilde{\Delta}^{-1}\dot{r})$, which implies that $\bar{\Omega}_i = 0$. ∎

As a consequence of Theorem 5.6, the necessary calculations in the controllability analysis of the successive symmetric products involving the input vector fields (cf. Corollary 3.3) or their projections to $\mathcal{D}$ (cf. Corollary 4.2) are further simplified. In fact, for two vector fields $X = (g\xi, v)$ and $Y = g(\eta, w)$ having vanishing associated locked body angular velocities $\Omega = 0$, $\Psi = 0$, we have by Corollary 3.3 that

$$\langle X : Y \rangle_{\mathcal{G}} = g \begin{pmatrix} -A \langle v : w \rangle_{\Delta} \\ \langle v : w \rangle_{\Delta} \end{pmatrix},$$

which also has vanishing locked body angular velocity. On the other hand, for two vector fields $X = (g\xi, v) \in \mathcal{D}$ and $Y = g(\eta, w) \in \mathcal{D}$ having $\bar{\Omega} = 0$, $\bar{\Psi} = 0$, respectively, we have by Corollary 4.2 that

$$\langle X : Y \rangle = g \begin{pmatrix} -\tilde{I}^{-1}\tilde{\mathbb{L}}^{s} - \mathbb{A}\left(\langle v : w \rangle_{\tilde{\Delta}} - \tilde{\Delta}^{-1}\tilde{\mathbb{S}}^{s}\right) \\ \langle v : w \rangle_{\tilde{\Delta}} - \tilde{\Delta}^{-1}\tilde{\mathbb{S}}^{s} \end{pmatrix},$$

with $\tilde{\mathbb{L}}^{s} = I(\tilde{A}v, \gamma.w - [\cdot, \eta]) + I(\tilde{A}w, \gamma.v - [\cdot, \xi])$ and $\tilde{\mathbb{S}}^{s} = I(\tilde{A}w, \mathbb{B}(v, \cdot)) + I(\tilde{A}v, \mathbb{B}(w, \cdot))$.

Notice also that the tests we have obtained here for principal fiber bundles are the natural extension of the results developed in [7] for mechanical control systems on Lie groups. The major difference is that, on Lie groups, G-invariance implies that the tests are expressed in $\mathfrak{g}$ in a *purely algebraic* way, whereas, on principal fiber bundles, we have to take into account the role of the shape space M, and, therefore, *differentiation* is still required.

Another interesting aspect of this kind of mechanical control system is the adaptation of the concept of *weak controllability* for kinematic systems defined in [15]. This notion essentially means controllability in the fiber, without regard to the intermediate or final positions of the shape variables. This type of controllability is meaningful for locomotion systems, where the group elements correspond to positions and orientation (and therefore are the most interesting variables to control), and one really does not care about the shapes the system is describing. In the following, we discuss it for the second-order dynamical problems we are considering.

Assume then that we are dealing with a *trivial* principal fiber bundle; that is, the decomposition $Q = G \times M$ holds globally. Let V^{τ} denote any subset of Q such that $\tau(V^{\tau})$ is an open subset of G, where $\tau : Q \equiv G \times M \longrightarrow G$ denotes the natural projection. Let $q_0 = (r_0, g_0)$ and $U \subset Q$ as before. Then we have the following definitions.

DEFINITION 5.7. *The system* (5.1) *is* locally fiber configuration accessible (LFCA) at $q_0 \in Q$ *if there exists* $T > 0$ *such that* $\mathcal{R}_Q^U(q_0, \leq t)$ *contains a nonempty subset* V^{τ} *of* Q *for all neighborhoods* U *of* q_0 *and all* $0 \leq t \leq T$. *If this holds for any* $q_0 \in Q$, *then the system is called* LFCA.

DEFINITION 5.8. *The system* (5.1) *is* small-time locally fiber configuration controllable (STLFCC) at $q_0 \in Q$ *if there exists* $T > 0$ *such that* $\mathcal{R}_Q^U(q_0, \leq t)$ *contains a nonempty subset* V^{τ} *of* Q *such that* $g_0 \in \tau(V^{\tau})$ *for all neighborhoods* U *of* q_0 *and all* $0 \leq t \leq T$. *If this holds for any* $q_0 \in Q$, *then the system is called* STLFCC.

From the discussion above, one can prove the following theorem.

THEOREM 5.9. *Let the mechanical control system* (5.1) *be invariant under* G.

(i) *The system is LFCA at* $q = (r, g)$ *if* $\tau_* \overline{Lie}(\overline{Sym}(\mathfrak{B}))_{(r,e)} = \mathfrak{g}$.

(ii) *Suppose that the system is LFCA at* q *and that the projection through* τ *of every bad symmetric product* P *at* q *in* $\mathfrak{B}$, $\tau_* P$, *can be written as a linear combination of projections through* τ *of good symmetric products at* q *of lower degree than* P. *Then* (5.1) *is STLFCC at* q.

Proof. Along the zero section of TQ, $q \longmapsto 0_q$, we have that the decomposition $T_{0_q}TQ \equiv T_qQ \oplus T_qQ$ holds, where the first factor corresponds to configurations and the second one to velocities. Then, from [21], we know that the accessibility distribution $\mathcal{C}$ corresponding to the full control system (that is, considering as states both the configurations and the velocities) can be decomposed as $\mathcal{C}_{0_q} = \mathcal{C}_{hor}(q) \oplus \mathcal{C}_{ver}(q)$, with $\mathcal{C}_{hor}(q) = \overline{Lie}(\overline{Sym}(\mathcal{Y}))_q$ and $\mathcal{C}_{ver}(q) = \overline{Sym}(\mathcal{Y})_q$. If $\tau_* \overline{Lie}(\overline{Sym}(\mathfrak{B}))_{(r,e)} = \mathfrak{g}$, we can conclude that $T_gG \subset \mathcal{C}_{hor}(q)$, and hence the system (5.1) is LFCA at q. The other claim follows from the invariance of the system and Sussmann's result in [32]. □

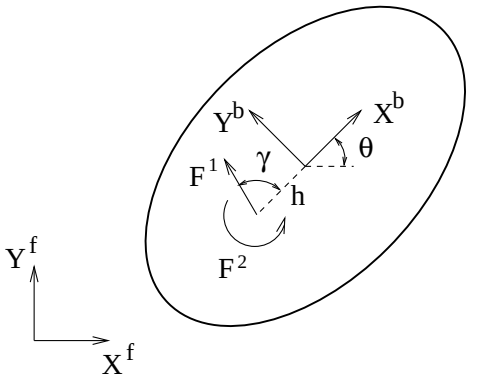


FIG. 6.1. *A planar blimp with rotating thruster.*

6. Examples.

6.1. The blimp. Consider a rigid body moving in $SE(2)$ with a thruster to adjust its pose (see Figure 6.1). The original motivation for this problem is the blimp system developed by Zhang and Ostrowski [35] restricted to the vertical plane. The control inputs are the thruster force F^1 and a torque F^2 that actuates its orientation with respect to the body axis $\{X^b, Y^b\}$. The acting point of the thruster is assumed to be located along the body's long axis, at a distance h from the center of mass.

The configuration of the blimp is determined by a tuple (x, y, θ, γ), where (x, y) is the position of the center of mass, θ is the orientation of the blimp with respect to the fixed basis $\{X^f, Y^f\}$, and γ denotes the orientation of the thrust with respect to the body basis $\{X^b, Y^b\}$. The configuration manifold is then $Q = SE(2) \times \mathbb{S}^1$.

For simplicity, we assume the thruster is massless. Then the Riemannian metric of the system is

$$\mathcal{G} = m(dx \otimes dx + dy \otimes dy) + (J_1 + J_2)d\theta \otimes d\theta + J_2 d\gamma \otimes d\gamma + J_2(d\theta \otimes d\gamma + d\gamma \otimes d\theta),$$

where m denotes the mass of the blimp, J_1 is its moment of inertia, and J_2 is the inertia of the thruster. The Lagrangian of the system is the kinetic energy associated to this metric; that is,

$$L = \frac{1}{2}m(\dot{x}^2 + \dot{y}^2) + \frac{1}{2}J_1\dot{\theta}^2 + \frac{1}{2}J_2(\dot{\gamma} + \dot{\theta})^2.$$

Finally, the input forces are given by

$$F^1 = \cos(\theta + \gamma)dx + \sin(\theta + \gamma)dy - h\sin\gamma d\theta, \quad F^2 = d\gamma.$$

The corresponding input vector fields can be computed to be

$$Y_1 = \frac{1}{m}\cos(\theta + \gamma)\frac{\partial}{\partial x} + \frac{1}{m}\sin(\theta + \gamma)\frac{\partial}{\partial y} - \frac{h}{J_1}\sin\gamma\frac{\partial}{\partial \theta} + \frac{h}{J_1}\sin\gamma\frac{\partial}{\partial \gamma},$$
$$Y_2 = -\frac{1}{J_1}\frac{\partial}{\partial \theta} + \frac{J_1 + J_2}{J_1 J_2}\frac{\partial}{\partial \gamma}.$$

This simple mechanical control system is invariant under the left multiplication of the Lie group $G = SE(2)$,

$$\begin{array}{rcl} \Phi : \quad G \times Q & \longrightarrow & Q \\ ((a, b, \alpha), (x, y, \theta, \gamma)) & \longmapsto & (x\cos\alpha - y\sin\alpha + a, x\sin\alpha + y\cos\alpha + b, \theta + \alpha, \gamma)\,. \end{array}$$

The reduced representation of the input vector fields at $\mathfrak{g} \times TM$ is given by

$$B_1 = \frac{1}{m}\cos\gamma\frac{\partial}{\partial x} + \frac{1}{m}\sin\gamma\frac{\partial}{\partial y} - \frac{h}{J_1}\sin\gamma\frac{\partial}{\partial\theta} + \frac{h}{J_1}\sin\gamma\frac{\partial}{\partial\gamma},$$
$$B_2 = -\frac{1}{J_1}\frac{\partial}{\partial\theta} + \frac{J_1+J_2}{J_1J_2}\frac{\partial}{\partial\gamma}.$$

Let $\{e_x, e_y, e_\theta\}$ be the canonical basis of the Lie algebra $se(2)$. Given the metric $\mathcal{G}$, we can readily identify from its reduced form (2.8) the local form of the mechanical connection and the inertia tensor

$$I = \begin{pmatrix} m & 0 & 0 \\ 0 & m & 0 \\ 0 & 0 & J_1+J_2 \end{pmatrix}, \quad A = \begin{pmatrix} 0 \\ 0 \\ \frac{J_2}{J_1+J_2} \end{pmatrix}.$$

As the shape space is one-dimensional and B^{mech} is skew-symmetric, we deduce that $B^{mech} = 0$. Some computations yields that DI also vanishes. Consequently, $\mathbb{S} = 0$. In addition,

$$D(I\eta)(\xi, v) = \left\{ \begin{pmatrix} m\frac{\partial\eta^1}{\partial\gamma} \\ m\frac{\partial\eta^2}{\partial\gamma} \\ (J_1+J_2)\frac{\partial\eta^3}{\partial\gamma} \end{pmatrix}^T + \frac{mJ_2}{J_1+J_2}\begin{pmatrix} \eta^2 \\ -\eta^1 \\ 0 \end{pmatrix}^T \right\} \begin{pmatrix} \xi^1 \\ \xi^2 \\ \xi^3 \end{pmatrix} v.$$

Note that $\Delta = (\frac{J_1J_2}{J_1+J_2})$. Therefore, the Christoffel symbols of ∇^Δ vanish and

$$\langle v : w \rangle_\Delta = \frac{\partial w}{\partial\gamma}v + \frac{\partial v}{\partial\gamma}w.$$

Summing up, we conclude that $\mathbb{L}^s$ in 3.3 for $X = (g\xi, v)$, $Y = (g\eta, w)$ is given by

$$\mathbb{L}^s = -\begin{pmatrix} m\left(\frac{\partial\xi^1}{\partial\gamma}w + \frac{\partial\eta^1}{\partial\gamma}v\right) \\ m\left(\frac{\partial\xi^2}{\partial\gamma}w + \frac{\partial\eta^2}{\partial\gamma}v\right) \\ (J_1+J_2)\left(\frac{\partial\xi^3}{\partial\gamma}w + \frac{\partial\eta^3}{\partial\gamma}v\right) \end{pmatrix} - \frac{mJ_2}{J_1+J_2}\begin{pmatrix} \Omega^2 w + \Psi^2 v \\ -\Omega^1 w - \Psi^1 v \\ 0 \end{pmatrix}.$$

Following Lemma 3.1, we can compute the symmetric product defined by ∇^I:

$$\langle \Omega : \Psi \rangle_I = \begin{pmatrix} -\Omega^2\Psi^3 - \Omega^3\Psi^2 \\ \Omega^1\Psi^3 + \Omega^3\Psi^1 \\ 0 \end{pmatrix}.$$

With these ingredients, we are now ready to perform the controllability analysis along the lines of section 5. Consider the following symmetric brackets:

$$\langle B_1 : B_1 \rangle_\mathcal{G} = \tfrac{h^2}{J_1^2}\sin(2\gamma)\begin{pmatrix} 0 \\ 0 \\ -1 \\ 1 \end{pmatrix}, \qquad \langle B_1 : B_2 \rangle_\mathcal{G} = \begin{pmatrix} -\frac{1}{mJ_2}\sin\gamma \\ \frac{1}{mJ_2}\cos\gamma \\ -\frac{h(J_1+J_2)}{J_1^2J_2}\cos\gamma \\ \frac{h(J_1+J_2)}{J_1^2J_2}\cos\gamma \end{pmatrix},$$

$$\langle B_2 : B_2 \rangle_\mathcal{G} = 0, \qquad \langle B_2 : \langle B_1 : B_1 \rangle_\mathcal{G} \rangle_\mathcal{G} = 2\tfrac{h^2}{J_1^2}\tfrac{J_1+J_2}{J_1J_2}\cos(2\gamma)\begin{pmatrix} 0 \\ 0 \\ -1 \\ 1 \end{pmatrix}.$$

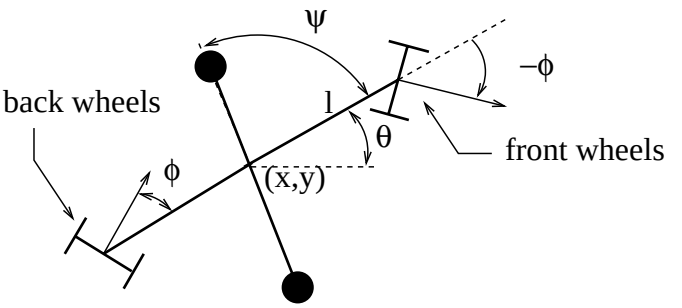


FIG. 6.2. *The snakeboard model.*

Note that $\{B_1, B_2, \langle B_1 : B_2\rangle_{\mathcal{G}}, \langle B_1 : B_1\rangle_{\mathcal{G}}, \langle B_2 : \langle B_1 : B_1\rangle_{\mathcal{G}}\rangle_{\mathcal{G}}\}$ span $\mathfrak{g} \times TM$ at every point (e, r), and hence the system is LCA. However, the bad bracket $\langle B_1 : B_1\rangle_{\mathcal{G}}$ is not in general a linear combination of the lower-order good brackets B_1 and B_2. Therefore, we can not conclude that the system is STLCC. In any case, at $\gamma = 0$, we have that $\langle B_1 : B_1\rangle_{\mathcal{G}}(e, 0) = 0$, and we can assure that the system is STLCC at $(g, 0)$ for all $g \in G$. However, if we restrict our attention to fiber configuration controllability, we can see that $\tau_*\langle B_1 : B_1\rangle_{\mathcal{G}} \in \text{span}\{\tau_* B_2\}$, and, therefore, the blimp is STLFCC. Physically, fiber controllability corresponds to the fact that we can use the shape torque to control the orientation angle θ to a desired value, but not θ and γ simultaneously.

6.2. The snakeboard. The snakeboard [22, 28] is a variant of the skateboard in which the passive wheel assemblies can pivot freely about a vertical axis. By coupling the twisting of the human torso with the appropriate turning of the wheels (where the turning is controlled by the rider's foot movement), the rider can generate a snake-like locomotion pattern without having to kick off the ground.

A simplified model is shown in Figure 6.2. We assume that the front and rear wheel axles move through equal and opposite rotations. This is based on the observations of human snakeboard riders who use roughly the same phase relationship. A momentum wheel rotates about a vertical axis through the center of mass, simulating the motion of a human torso.

The position and orientation of the snakeboard is determined by the coordinates of the center of mass (x, y) and its orientation θ. The shape variables are (ψ, ϕ), and so the configuration space is $Q = SE(2) \times \mathbb{S}^1 \times \mathbb{S}^1$. The physical parameters for the system are the mass of the board, m; the inertia of the rotor, J_r; the inertia of the wheels about the vertical axes, J_w; and the half-length of the board, l. A key component of the snakeboard is the use of the rotor inertia to drive the body. To keep the rotor and body inertias on similar scales, we make the additional simplifying assumption [4, 30] that the inertias of the system satisfy $J + J_r + 2J_w = ml^2$.

The Riemannian metric of this system is

$$\begin{aligned}\mathcal{G} = m(dx \otimes dx + dy \otimes dy) + (J + J_r + 2J_w)d\theta \otimes d\theta \\ + J_r(d\theta \otimes d\psi + d\psi \otimes d\theta) + J_r d\psi \otimes d\psi + 2J_w d\phi \otimes d\phi.\end{aligned}$$

The control torques are assumed to be applied to the rotation of the wheels and the rotor. Hence we consider

$$F^1 = d\psi, \quad F^2 = d\phi.$$

Observe that the snakeboard is an example of the type of dynamic locomotion systems we mentioned earlier, since the set of control inputs fully actuate the shape variables, $\operatorname{span}\{F^1, F^2\} = T^*M$. The corresponding input vector fields via the diffeomorphism $\sharp_{\mathcal{G}}$ are

$$Y_1 = -\frac{1}{J + 2J_w}\frac{\partial}{\partial\theta} + \frac{ml^2}{J_r(J + 2J_w)}\frac{\partial}{\partial\psi}, \quad Y_2 = \frac{1}{2J_w}\frac{\partial}{\partial\phi}.$$

The assumption that the wheels do not slip in the direction of the wheels axles yields the following two nonholonomic constraints:

$$\begin{aligned} &-\sin(\theta+\phi)\dot{x} + \cos(\theta+\phi)\dot{y} - l\cos\phi\,\dot{\theta} = 0,\\ &-\sin(\theta-\phi)\dot{x} + \cos(\theta-\phi)\dot{y} + l\cos\phi\,\dot{\theta} = 0. \end{aligned}$$

A quick set of calculations shows that this constrained mechanical system is invariant under the left multiplication in the Lie group $SE(2)$. The intersection $S = \mathcal{V} \cap \mathcal{D}$ can be seen to be one-dimensional. Moreover, we have that $S_{(e,r)} = e_{1Q}$, where $e_1 = l\cos\phi e_x - \sin\phi e_\theta$. We complete the basis by adding two elements generating $S^{\perp}_{(e,r)}$:

$$e_2 = e_y, \quad e_3 = \frac{1}{l}\tan\phi e_x + e_\theta.$$

Taking into account the discussion of the preceding sections, we can identify the following elements:

$$I = \begin{pmatrix} m & 0 & 0 \\ 0 & m & 0 \\ 0 & 0 & ml^2 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ \frac{J_r}{ml^2} & 0 \end{pmatrix}, \quad \mathbb{A} = \begin{pmatrix} -\frac{J_r}{2ml}\sin(2\phi) & 0 \\ 0 & 0 \\ \frac{J_r}{ml^2}\sin^2\phi & 0 \end{pmatrix}.$$

Our choice of generators of $\mathcal{D}_{(e,r)}$, following section 2.3, is then

$$\mathcal{D}_{(r,e)} = \operatorname{span}\left\{\frac{\partial}{\partial\psi} + \frac{J_r}{ml^2}\sin\phi e_1, \frac{\partial}{\partial\phi}, e_1\right\}.$$

The projections to $\mathcal{D}$ of the input vector fields under the orthogonal decomposition $TQ = \mathcal{D} \oplus \mathcal{D}^{\perp}$ are

$$\begin{aligned} \mathcal{B}_1 &= \mathcal{P}(B_1) = \frac{ml^2}{J_r(ml^2 - J_r\sin^2\phi)}\left(\frac{\partial}{\partial\psi} + \frac{J_r}{ml^2}\sin\phi e_1\right),\\ \mathcal{B}_2 &= \mathcal{P}(B_2) = \frac{1}{2J_w}\frac{\partial}{\partial\phi}. \end{aligned}$$

For the sake of completeness, we have computed the terms $\tilde{\mathbb{L}}^s \in \mathfrak{g}^{\mathcal{D}}$ and $\tilde{\mathbb{S}}^s$ in 4.4 for any G-invariant vector fields $X = (g\xi, v)$ and $Y = (g\eta, w)$, although we already pointed out in section 5 (cf. Theorem 5.6) that the amount of calculations for the controllability tests can be made quite lighter, taking into account the fact that $\bar{\Omega}_i = 0$

for $\mathcal{B}_i$, $i = 1, 2$.

$$\tilde{\mathbb{L}}^s = -\left\{ ml\cos\phi \sum_{\alpha=1}^{2} \left(\frac{\partial \bar{\Psi}^1}{\partial r^\alpha} v^\alpha + \frac{\partial \bar{\Omega}^1}{\partial r^\alpha} w^\alpha \right) - ml^2 \sin\phi \sum_{\alpha=1}^{2} \left(\frac{\partial \bar{\Psi}^3}{\partial r^\alpha} v^\alpha + \frac{\partial \bar{\Omega}^3}{\partial r^\alpha} w^\alpha \right) + J_r \cos\phi (v^2 w^1 + w^2 v^1) + \frac{J_r}{2l} \sin(2\phi) \sin\phi (w^1 \xi^2 + v^1 \eta^2) \right\} e_1^*,$$

$$\tilde{\mathbb{S}}^s = -\frac{J_r^2}{2ml^2} \sin(2\phi) \begin{pmatrix} w^1 v^2 + w^2 v^1 \\ -2 v^1 w^1 \end{pmatrix}.$$

The controllability analysis yields the following results at the point $\mathbf{0} = (0, 0, 0, 0, 0)$:

$$\begin{array}{cc} \langle \mathcal{B}_1 : \mathcal{B}_1 \rangle(\mathbf{0}) = 0, & \langle \mathcal{B}_1 : \mathcal{B}_2 \rangle(\mathbf{0}) = \frac{1}{2J_w ml} e_x, \\ \langle \mathcal{B}_2 : \mathcal{B}_2 \rangle(\mathbf{0}) = 0, & [\mathcal{B}_1, \mathcal{B}_2](\mathbf{0}) = \frac{1}{2J_w ml} e_x, \\ [\mathcal{B}_2, [\mathcal{B}_1, \mathcal{B}_2]](\mathbf{0}) = -\frac{1}{2J_w^2 ml^2} e_\theta, & [\mathcal{B}_2, [\mathcal{B}_1, [\mathcal{B}_2, [\mathcal{B}_1, \mathcal{B}_2]]]](\mathbf{0}) = -\frac{1}{4J_w^3 m^2 l^3} e_y - \frac{1}{2J_w^3 m^2 l^4} e_\theta. \end{array}$$

Note that $\{\mathcal{B}_1, \mathcal{B}_2, \langle \mathcal{B}_1 : \mathcal{B}_2 \rangle, [\mathcal{B}_2, [\mathcal{B}_1, \mathcal{B}_2]], [\mathcal{B}_2, [\mathcal{B}_1, [\mathcal{B}_2, [\mathcal{B}_1, \mathcal{B}_2]]]]\}$ span $\mathfrak{g} \times T_{(0,0)}M$, and so the system is LCA at $(g, 0, 0)$ for all $g \in G$. Moreover, the bad symmetric products $\langle \mathcal{B}_1 : \mathcal{B}_1 \rangle$ and $\langle \mathcal{B}_2 : \mathcal{B}_2 \rangle$ vanish at $\mathbf{0}$, and the remaining ones are either 0 or in span$\{\mathcal{B}_2(\mathbf{0}), \langle \mathcal{B}_1 : \mathcal{B}_2 \rangle(\mathbf{0})\}$, and so we can conclude that the snakeboard is STLCC at $(g, 0, 0)$ for all $g \in G$.

7. Conclusions. We have developed a new set of tools that can be used in the study of simple mechanical systems evolving on principal fiber bundles. These tools have direct application to a large class of problems in robotic locomotion. Using the Lie group symmetries that are associated with an invariant mechanical system on a principal fiber bundle, we have given an explicit formulation of the affine connection in terms of the mechanical and nonholonomic connections for unconstrained and constrained systems, respectively. This formulation can greatly reduce the amount of computation necessary to derive controllability tests, as was observed during the analysis of the snakeboard system.

We have defined a new notion of fiber configuration controllability, which can be used to focus the analysis on the important components of a locomotion system, namely, the fiber variables of position and orientation. The tools developed in this paper were applied to two systems—the planar rigid body and the snakeboard robot.

We are currently working on applying these tools to motion planning for such systems (see [26]). Recent work by Bullo, Leonard, and Lewis [6] suggests an excellent avenue for applying the affine connection in a motion planning framework. We will also explore connections of these tools to steering for dynamic systems as, for example, was done by Ostrowski [29] using the reduced equations for the snakeboard [31].

Acknowledgments. The first three authors wish to thank F. Bullo and A. D. Lewis for interesting and helpful conversations. J. Cortés and S. Martínez wish to thank F. Cantrijn and M. de León for their support. We would also like to thank the anonymous reviewers for their useful comments.

REFERENCES

[1] R. Abraham and J. E. Marsden, *Foundations of Mechanics*, 2nd ed., Benjamin-Cummings, Reading, MA, 1978.

[2] L. Bates and J. Śniatycki, *Nonholonomic reduction*, Rep. Math. Phys., 32 (1992), pp. 99–115.
[3] A. M. Bloch and P. E. Crouch, *Newton's law and integrability of nonholonomic systems*, SIAM J. Control Optim., 36 (1998), pp. 2020–2039.
[4] A. M. Bloch, P. S. Krishnaprasad, J. E. Marsden, and R. M. Murray, *Nonholonomic mechanical systems with symmetry*, Arch. Ration. Mech. Anal., 136 (1996), pp. 21–99.
[5] F. Bullo, *Series expansions for the evolution of mechanical control systems*, SIAM J. Control Optim., 40 (2001), pp. 166–190.
[6] F. Bullo, N. E. Leonard, and A. D. Lewis, *Controllability and motion algorithms for underactuated Lagrangian systems on Lie groups*, IEEE Trans. Automat. Control, 45 (2000), pp. 1437–1454.
[7] F. Bullo and A. D. Lewis, *Configuration controllability of mechanical systems on Lie groups*, in Proceedings of the International Symposium on the Mathematical Theory of Networks and Systems, St. Louis, MO, 1996.
[8] F. Bullo and K. M. Lynch, *Kinematic controllability for decoupled trajectory planning in underactuated mechanical systems*, IEEE Trans. Robot. Automat., 17 (2001), pp. 402–412.
[9] F. Cantrijn, M. de León, J. C. Marrero, and D. Martín de Diego, *Reduction of nonholonomic mechanical systems with symmetries*, Rep. Math. Phys., 42 (1998), pp. 25–45.
[10] J. Cortés and M. de León, *Reduction and reconstruction of the dynamics of nonholonomic systems*, J. Phys. A, 32 (1999), pp. 8615–8645.
[11] J. Cortés and S. Martínez, *Configuration controllability of mechanical systems underactuated by one control*, SIAM J. Control Optim., submitted.
[12] J. Cortés, S. Martínez, and F. Bullo, *On nonlinear controllability and series expansions for Lagrangian systems with damping*, IEEE Trans. Automat. Control, to appear; also available online from http://motion.csl. uiuc.edu/~bullo/papers.
[13] P. E. Crouch, *Geometric structures in systems theory*, Proc. IEE-D, 128 (1981), pp. 242–252.
[14] S. Helgason, *Differential Geometry, Lie Groups and Symmetric Spaces*, Academic Press, New York, 1978.
[15] S. D. Kelly and R. M. Murray *Geometric phases and robotic locomotion*, J. Robotic Systems, 12 (1995), pp. 417–431.
[16] J. Koiller, *Reduction of some classical non-holonomic systems with symmetry*, Arch. Ration. Mech. Anal., 118 (1992), pp. 113–148.
[17] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry*, Volume I, Interscience Publishers, Wiley, New York, 1963.
[18] A. D. Lewis, *Local configuration controllability for a class of mechanical systems with a single input*, in Proceedings of the 1997 European Control Conference, Brussels, Belgium, 1997.
[19] A. D. Lewis, *Affine connections and distributions with applications to nonholonomic mechanics*, Rep. Math. Phys., 42 (1998), pp. 135–164.
[20] A. D. Lewis, *Simple mechanical control systems with constraints*, IEEE Trans. Automat. Control, 45 (2000), pp. 1420–1436.
[21] A. D. Lewis and R. M. Murray, *Configuration controllability of simple mechanical control systems*, SIAM J. Control Optim., 35 (1997), pp. 766–790.
[22] A. D. Lewis, J. P. Ostrowski, R. M. Murray, and J. W. Burdick, *Nonholonomic mechanics and locomotion: The snakeboard example*, in Proceedings of the IEEE International Conference on Robotics and Automation, San Diego, CA, 1994, pp. 2391–2397.
[23] C.-M. Marle, *Reduction of constrained mechanical systems and stability of relative equilibria*, Comm. Math. Phys., 174 (1995), pp. 295–318.
[24] J. E. Marsden, R. Montgomery, and T. S. Ratiu, *Reduction, symmetry and phases in mechanics*, Mem. Amer. Math. Soc., 88 (1990).
[25] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*, Springer-Verlag, New York, 1994.
[26] S. Martínez and J. Cortés, *Motion control algorithms for mechanical systems with symmetries*, Acta Appl. Math., submitted.
[27] H. Nijmeijer and A. J. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.
[28] J. P. Ostrowski, *Geometric Perspectives on the Mechanics and Control of Undulatory Locomotion*, Ph.D. Thesis, California Institute of Technology, Pasadena, CA, 1995.
[29] J. P. Ostrowski, *Steering for a class of dynamic nonholonomic systems*, IEEE Trans. Automat. Control, 45 (2000), pp. 1492–1498.
[30] J. P. Ostrowski and J. W. Burdick, *Controllability for mechanical systems with symmetries and constraints*, Appl. Math. Comput. Sci., 7 (1997), pp. 305–331.
[31] J. P. Ostrowski and J. W. Burdick, *The geometric mechanics of undulatory robotic locomotion*, Int. J. Robotic Research, 17 (1998), pp. 683–702.

[32] H. J. Sussmann, *A general theorem on local controllability*, SIAM J. Control Optim., 25 (1987), pp. 158–194.
[33] J. Z. Synge, *Geodesics in nonholonomic geometry*, Math. Ann., 99 (1928), pp. 738–751.
[34] A. J. van der Schaft, *Linearization of Hamiltonian and gradient systems*, IMA J. Math. Control Inform., 1 (1984), pp. 185–198.
[35] H. Zhang and J. P. Ostrowski, *Visual serving with dynamics: Control of an unmanned blimp*, in Proceedings of the IEEE International Conference on Robotics and Automation, Detroit, MI, 1999, pp. 618–623.
[36] H. Zhang and J. P. Ostrowski, *Control algorithms using affine connections on principal fiber bundles*, in Proceedings of the IFAC Workshop on Lagrangian and Hamiltonian Methods for Nonlinear Control, Princeton, NJ, 2000.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 3, pp. 875–899

AFFINE INVARIANT CONVERGENCE ANALYSIS FOR INEXACT AUGMENTED LAGRANGIAN-SQP METHODS*

S. VOLKWEIN† AND M. WEISER‡

Abstract. An affine invariant convergence analysis for inexact augmented Lagrangian-SQP methods is presented. The theory is used for the construction of an accuracy matching between iteration errors and truncation errors, which arise from the inexact linear system solvers. The theoretical investigations are illustrated numerically by an optimal control problem for the Burgers equation.

Key words. nonlinear programming, multiplier methods, affine invariant norms, Burgers' equation

AMS subject classifications. 49M15, 65N99, 90C55

PII. S0363012900383089

1. Introduction.

This paper is concerned with an optimization problem of the following type:

$$(\mathsf{P}) \qquad \text{minimize } J(x) \text{ subject to } e(x) = 0,$$

where $J : X \to \mathbb{R}$ and $e : X \to Y$ are sufficiently smooth functions and X, Y are real Hilbert spaces. These types of problems occur, for example, in the optimal control of systems described by partial differential equations. To solve (P) we use the augmented Lagrangian-SQP (sequential quadratic programming) technique as developed in [11]. In this method the differential equation is treated as an equality constraint, which is enforced by a Lagrangian term together with a penalty functional. We present an algorithm which has second-order convergence rate and depends upon a second-order sufficient optimality condition. In comparison with SQP methods the augmented Lagrangian-SQP method has the advantage of a more global behavior. For certain examples we found it to be less sensitive with respect to the starting values, and the region for second-order convergence rate was reached earlier; see, e.g., [11, 15, 17]. We shall point out that the penalty term of the augmented Lagrangian functional need not to be implemented but rather that it can be realized by a first-order Lagrangian update.

Augmented Lagrangian-SQP methods applied to problem (P) are essentially Newton-type methods applied to the Kuhn–Tucker equations for an augmented optimization problem. Newton methods and their behavior under different linear transformations were studied by several authors; see [5, 6, 7, 8, 10], for instance. In this paper, we combine both lines of work and present an affine invariant setting for analysis and implementation of augmented Lagrangian-SQP methods in Hilbert spaces. An affine invariant convergence theory for inexact augmented Lagrangian-SQP methods is presented. Then the theoretical results are used for the construction of an accuracy

*Received by the editors December 21, 2000; accepted for publication (in revised form) March 12, 2002; published electronically September 19, 2002.
http://www.siam.org/journals/sicon/41-3/38308.html
†Karl-Franzens-Universität Graz, Institut für Mathematik, Heinrichstraße 36, A-8010 Graz, Austria (stefan.volkwein@uni-graz.at).
‡Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB), Takustraße 7, D-14195 Berlin, Germany (weiser@zib.de). The work of this author was supported by Deutsche Forschungsgemeinschaft (DFG), Sonderforschungsbereich 273.

matching between iteration errors and truncation errors, which arise from the inexact linear system solvers.

The paper is organized as follows. In section 2 the augmented Lagrangian-SQP method is introduced and necessary prerequisites are given. The affine invariance is introduced in section 3. In section 4 an affine invariant convergence result for the augmented Lagrangian-SQP method is presented. Two invariant norms for optimal control problems are analyzed in section 5, and the inexact Lagrangian-SQP method is studied in section 6. In the last section we report on some numerical experiments done for an optimal control problem for the Burgers equation, which is a one-dimensional model for nonlinear convection-diffusion phenomena.

2. The augmented Lagrangian-SQP method. Let us consider the constrained optimal control problem

$$\text{(P)} \qquad \text{minimize } J(x) \text{ subject to } e(x) = 0,$$

where $J : X \to \mathbb{R}$, $e : X \to Y$, and X, Y are real Hilbert spaces. Throughout we do not distinguish between a functional in the dual space and its Riesz representation in the Hilbert space. The Hilbert space $X \times Y$ is endowed with the Hilbert space product topology and, for brevity, we set $Z = X \times Y$.

Let us present an example for (P) that illustrates our theoretical investigations and that is used for the numerical experiments carried out in section 7. For more details we refer the reader to [18].

Example 2.1. Let Ω denote the interval $(0,1)$ and set $Q = (0,T) \times \Omega$ for given $T > 0$. We define the space $W(0,T)$ by

$$W(0,T) = \left\{ \varphi \in L^2(0,T;H^1(\Omega)) : \ \varphi_t \in L^2(0,T;H^1(\Omega)') \right\},$$

which is a Hilbert space endowed with the common inner product. For controls $u, v \in L^2(0,T)$ the state $y \in W(0,T)$ is given by the weak solution of the unsteady Burgers equation with Robin-type boundary conditions, i.e., y satisfies

$$y(0,\cdot) = y_0 \quad \text{in } L^2(\Omega) \tag{2.1a}$$

and

$$\begin{aligned} &\langle y_t(t,\cdot),\varphi\rangle_{(H^1)',H^1} + \sigma_1(t)y(t,1)\varphi(1) - \sigma_0(t)y(t,0)\varphi(0) \\ &+ \int_\Omega \nu y_x(t,\cdot)\varphi' + \big(y(t,\cdot)y_x(t,\cdot) - f(t,\cdot)\big)\varphi \, dx = v(t)\varphi(1) - u(t)\varphi(0) \end{aligned} \tag{2.1b}$$

for all $\varphi \in H^1(\Omega)$ and $t \in (0,T)$ a.e., where $\langle\cdot,\cdot\rangle_{(H^1)',H^1}$ denotes the duality pairing between $H^1(\Omega)$ and its dual. We suppose that $f \in L^2(\Omega)$, $y_0 \in L^\infty(\Omega)$, $\sigma_0, \sigma_1 \in L^\infty(0,T)$, and $\nu > 0$. Recall that $W(0,T)$ is continuously embedded into the space of all continuous functions from $[0,T]$ into $L^2(\Omega)$, denoted by $C([0,T];L^2(\Omega))$; see, e.g., [3, p. 473]. Therefore, (2.1a) makes sense. With controls u, v we associate the cost of tracking type

$$J(y,u,v) = \frac{1}{2}\int_Q |y-z|^2 \, dxdt + \frac{1}{2}\int_0^T \alpha \, |u|^2 + \beta \, |v|^2 \, dt,$$

where $z \in L^2(Q)$ and $\alpha, \beta > 0$ are fixed. Let $X = W(0,T) \times L^2(0,T) \times L^2(0,T)$, $Y = L^2(0,T;H^1(\Omega)) \times L^2(\Omega)$, and $x = (y,u,v)$. We introduce the bounded operator

$$\tilde{e} : X \to L^2(0,T;H^1(\Omega)'),$$

whose action is defined by

$$\begin{aligned}\langle \tilde{e}(y,u,v),\lambda\rangle_{L^2(0,T;H^1(\Omega)'),L^2(0,T;H^1(\Omega))}&\\ = \int_0^T \langle y_t(t,\cdot),\lambda(t,\cdot)\rangle_{(H^1)',H^1}\,dt + \int_Q (\nu y_x\lambda_x + y y_x\lambda - f\lambda)\,dxdt&\\ + \int_0^T \big((\sigma_1 y(\cdot,1)-v)\lambda(\cdot,1) - (\sigma_0 y(\cdot,0)-u)\lambda(\cdot,0)\big)\,dt&\end{aligned}$$

for $\lambda \in L^2(0,T;H^1(\Omega))$. Define $e : X \to Y$ by

$$e(y,u,v) = \big((-\Delta + I)^{-1}\tilde{e}(y,u,v), y(0,\cdot) - y_0\big),$$

where for given $g \in H^1(\Omega)'$ the mapping $(-\Delta + I)^{-1} : H^1(\Omega)' \to H^1(\Omega)$ is the Neumann solution operator associated with

$$\int_\Omega v'\varphi' + v\varphi\,dx = \langle g,\varphi\rangle_{(H^1)',H^1} \quad \text{for all } \varphi \in H^1(\Omega).$$

Now the optimal control problem can be written in the form (P).

For $c \geq 0$ the augmented Lagrange functional $L_c : Z \to \mathbb{R}$ associated with (P) is defined by

$$L_c(x,\lambda) = J(x) + \langle e(x),\lambda\rangle_Y + \frac{c}{2}\,\|e(x)\|_Y^2.$$

The following assumption is rather standard for SQP methods in Hilbert spaces and is supposed to hold throughout the paper.

ASSUMPTION 1. *Let $x^* \in X$ be a reference point such that*

(a) *J and e are twice continuously Fréchet-differentiable, and the mappings J'' and e'' are Lipschitz-continuous in a neighborhood of x^*;*

(b) *the linearization $e'(x^*)$ of the operator e at x^* is surjective;*

(c) *there exists a Lagrange multiplier $\lambda^* \in Y$ satisfying the first-order necessary optimality conditions*

$$L_c'(x^*,\lambda^*) = 0, \quad e(x^*) = 0 \quad \text{for all } c \geq 0, \tag{2.2}$$

where the Fréchet-derivative with respect to the variable x is denoted by a prime; and

(d) *there exists a constant $\kappa > 0$ such that*

$$\langle L_0''(x^*,\lambda^*)\chi,\chi\rangle_X \geq \kappa\,\|\chi\|_X^2 \quad \text{for all } \chi \in \ker e'(x^*),$$

where ker $e'(x^*)$ *denotes the kernel or null space of $e'(x^*)$.*

Remark 2.2. In the context of Example 2.1 we write $x^* = (y^*,u^*,v^*)$. It was proved in [18] that Assumption 1 holds, provided $\|y^* - z\|_{L^2(Q)}$ is sufficiently small.

The next proposition follows directly from Assumption 1. For a proof we refer to [12] and [13], for instance.

PROPOSITION 2.3. *With Assumption* 1 *holding, x^* is a local solution to* (P). *Furthermore, there exists a neighborhood of (x^*,λ^*) such that (x^*,λ^*) is the unique solution of* (2.2) *in this neighborhood.*

The mapping $x \mapsto L_c(x,\lambda^*)$ can be bounded from below by a quadratic function. This fact is referred to as augmentability of L_c and is formulated in the next proposition. For a proof we refer the reader to [11].

PROPOSITION 2.4. *There exist a neighborhood $\hat{U}$ of x^* and a constant $\bar{c} \geq 0$ such that the mapping $x \mapsto L_c''(x, \lambda^*)$ is coercive on the whole space X for all $x \in \hat{U}$ and $c \geq \bar{c}$.*

Remark 2.5. Due to Assumption 1 and Proposition 2.4 there are convex neighborhoods $U(x^*) \subset X$ of x^* and $U(\lambda^*) \subset Y$ of λ^* such that for all $(x, \lambda) \in U = U(x^*) \times U(\lambda^*)$

(a) $J(x)$ and $e(x)$ are twice Fréchet-differentiable, and their second Fréchet-derivatives are Lipschitz-continuous in $\overline{U(x^*)}$;
(b) $e'(x)$ is surjective;
(c) $L_0''(x, \lambda)$ is coercive on the kernel of $e'(x)$;
(d) the point $z^* = (x^*, \lambda^*)$ is the unique solution to (2.2) in U; and
(e) there exist $\tilde{\kappa} > 0$ and $\bar{c} \geq 0$ such that

$$\langle L_c''(x, \lambda)\chi, \chi \rangle_X \geq \tilde{\kappa} \|\chi\|_X^2 \quad \text{for all } \chi \in X \text{ and } c \geq \bar{c}. \tag{2.3}$$

To shorten notation let us introduce the operator

$$F_c(x, \lambda) = \begin{pmatrix} L_c'(x, \lambda) \\ e(x) \end{pmatrix} \quad \text{for all } (x, \lambda) \in U.$$

Then the first-order necessary optimality conditions (2.2) can be expressed as

$$F_c(x^*, \lambda^*) = 0 \quad \text{for all } c \geq 0. \tag{OS}$$

To find x^* numerically we solve (OS) by the Newton method. The Fréchet-derivative of the operator F_c in U is given by

$$\nabla F_c(x, \lambda) = \begin{pmatrix} L_c''(x, \lambda) & e'(x)^\star \\ e'(x) & 0 \end{pmatrix}, \tag{2.4}$$

where $e'(x)^\star : Y \to X$ denotes the adjoint of the operator $e'(x)$.

Remark 2.6. With Assumption 1 holding, there exists a constant $C > 0$ satisfying

$$\|\nabla F_c(x, \lambda)^{-1}\|_{\mathcal{B}(Z)} \leq C \quad \text{for all } (x, \lambda) \in U \tag{2.5}$$

(see, e.g., in [9, p. 114]), where $\mathcal{B}(Z)$ denotes the Banach space of all bounded linear operators on Z.

Now we formulate the augmented Lagrangian-SQP method.

ALGORITHM 1.

(a) *Choose $(x^0, \lambda^0) \in U$, $c \geq 0$, and put $k = 0$.*
(b) *Set $\tilde{\lambda}^k = \lambda^k + ce(x^k)$.*
(c) *Solve for $(\Delta x, \Delta\lambda)$ the linear system*

$$\nabla F_0(x^k, \tilde{\lambda}^k) \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} = -F_0(x^k, \tilde{\lambda}^k). \tag{2.6}$$

(d) *Set $(x^{k+1}, \lambda^{k+1}) = (x^k + \Delta x, \tilde{\lambda}^k + \Delta\lambda)$, $k = k + 1$, and go back to* (b).

Remark 2.7. Since X and Y are Hilbert spaces, (x^{k+1}, λ^{k+1}) can equivalently be obtained by solving the linear system

$$\nabla F_c(x^k, \lambda^k) \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} = -F_c(x^k, \lambda^k) \tag{2.7}$$

and setting $(x^{k+1}, \lambda^{k+1}) = (x^k + \Delta x, \lambda^k + \Delta\lambda)$. Equation (2.7) corresponds to a Newton step applied to (OS). This form of the iteration requires the implementation of $e'(x^k)^\star e'(x^k)$, whereas steps (b) and (c) of Algorithm 1 do not—see [11]. In the case of Example 2.1 this requires at least one additional solve of the Poisson equation.

3. Affine invariance. Let $\tilde{B} : X \to X$ be an arbitrary isomorphism. We transform the x variable by $x = \tilde{B}y$. Thus, instead of (P) we study the whole class of equivalent transformed minimization problems

$$\text{minimize } J(\tilde{B}y) \text{ subject to } e(\tilde{B}y) = 0 \tag{3.1}$$

with the transformed solutions $\tilde{B}y^* = x^*$. Setting

$$B = \begin{pmatrix} \tilde{B} & 0 \\ 0 & I \end{pmatrix} \quad \text{and} \quad G_c(y,\xi) = B^\star F_c(x,\lambda) \quad \text{with } (x,\lambda) = (\tilde{B}y,\xi),$$

the first-order necessary optimality conditions have the form

$$G_c(y,\xi) = 0 \quad \text{for all } c \geq 0. \tag{$\widehat{\mathsf{OS}}$}$$

Applying Algorithm 1 to ($\widehat{\mathsf{OS}}$) we get an equivalent sequence of transformed iterates.

THEOREM 3.1. *Suppose that Assumption* 1 *holds. Let* $(x^0,\lambda^0) \in U$ *and* $(y^0,\xi^0) = (\tilde{B}^{-1}x^0,\lambda^0)$ *be the starting iterates for Algorithm* 1 *applied to the optimality conditions* (OS) *and* ($\widehat{\mathsf{OS}}$)*, respectively. Then both sequences of iterates are well-defined and equivalent in the sense of*

$$(\tilde{B}y^k,\xi^k) = (x^k,\lambda^k) \quad \textit{for } k = 0,1,\ldots . \tag{3.2}$$

Proof. First note that the Fréchet-derivative of the operator G_c is given by

$$\nabla G_c(y,\xi) = B^\star \nabla F_c(x,\lambda)B \quad \text{with } (x,\lambda) = (\tilde{B}y,\xi). \tag{3.3}$$

To prove (3.2) we use an induction argument. By assumption the identity (3.2) holds for $k = 0$. Now suppose that (3.2) is satisfied for $k \geq 0$. This implies $\tilde{B}y^k = x^k$ and $\xi^k = \lambda^k$. Using step (b) of Algorithm 1, it follows that $\tilde{\xi}^k = \xi^k + ce(\tilde{B}y^k) = \tilde{\lambda}^k$. From (3.3),

$$\nabla F_0(x^k,\tilde{\lambda}^k)\begin{pmatrix}\Delta x\\ \Delta\lambda\end{pmatrix} = -F_0(x^k,\tilde{\lambda}^k), \quad \text{and} \quad \nabla G_0(y^k,\tilde{\xi}^k)\begin{pmatrix}\Delta y\\ \Delta\xi\end{pmatrix} = -G_0(y^k,\tilde{\xi}^k),$$

we conclude that $(\Delta y,\Delta\xi) = (\tilde{B}^{-1}\Delta x,\Delta\lambda)$. Utilizing step (d) of Algorithm 1 we get the desired result. ∎

Due to the previous theorem the augmented Lagrangian-SQP method is *invariant under arbitrary transformations* $\tilde{B}$ of the state space X. This nice property should, of course, be inherited by any convergence theory and termination criteria. In section 4 we develop such an invariant theory.

Example 3.2. The usual local Newton–Mysovskii convergence theory (cf. [14, p. 412]) is *not* affine invariant, which leads to an unsatisfactory description of the domain of local convergence. Consider the optimization problem

$$\min \eta^2 + (\xi+\eta)^3 - \xi - \eta \text{ subject to } \xi + 2\eta = 0 \tag{3.4}$$

with unique solution $x^* = (\xi^*,\eta^*) = (2/3,-1/3)$ and associated Lagrange multiplier $\lambda^* = -2/3$. Note that the Jacobian ∇F_0 does not depend on λ here but only on $x = (\xi,\eta)$. In the context of Remark 2.5 we choose the neighborhood

$$U = U(x^*) \times U(\lambda^*) = \{(\xi,\eta) \in \mathbb{R}^2 : |\xi+\eta-1/3| < 0.16\sqrt{2}\} \times \mathbb{R}.$$

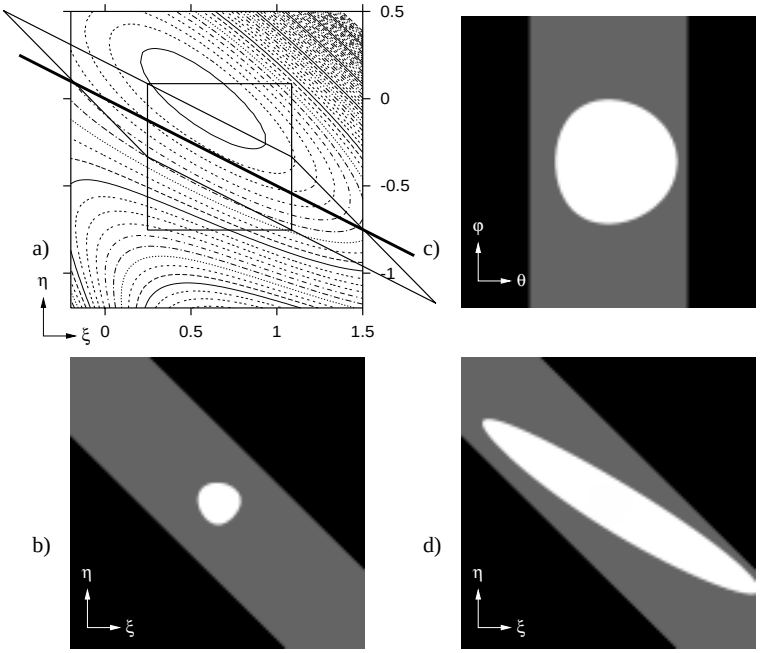


FIG. 3.1. *Illustration for Example* 3.2. (a) *Contour lines of the cost functional, the constraint, and the areas occupied by the other subplots.* (b) *Neighborhood* $U(x^*)$ *(gray) and Kantorovich ball of theoretically assured convergence (white) for the original problem formulation.* (c) $U(x^*)$ *and Kantorovich ball for the "better" formulation.* (d) $U(x^*)$ *and Kantorovich ball for the "better" formulation plotted in coordinates of the original formulation.*

Defining

$$\alpha = \sup_{x\in U(x^*)} \|\nabla F_0(x)^{-1}\| \quad \text{and} \quad \beta = \sup_{x,y\in U(x^*),x\neq y} \frac{\|\nabla F_0(x) - \nabla F_0(y)\|}{\|x-y\|_2},$$

the Newton–Mysovskii theory essentially guarantees convergence for all starting points in the Kantorovich region

$$K := \left\{ z \in U : \|\nabla F_0(z)^{-1} F_0(z)\|_2 \leq \frac{2}{\alpha\beta} \right\}.$$

Here, $\|\cdot\|$ denotes the spectral norm for symmetric matrices and $\|\cdot\|_2$ is the Euclidean norm. For our choice of U, resulting in $\alpha \approx 1.945$ and $\beta = 12\sqrt{2}$, a section of the Kantorovich region at $\lambda = \lambda^*$ is plotted in Figure 3.1(b). A different choice of coordinates, however, yields a significantly different result. With the transformation

$$\xi = 2\vartheta - \phi, \quad \eta = \phi - \vartheta,$$

problem (3.4) can be written as

$$\min(\phi - \vartheta)^2 + \vartheta^3 - \vartheta \text{ subject to } \phi = 0.$$

For the same neighborhood U, the better constants $\alpha \approx 1.859$ and $\beta = 6$ result. Again, a section of the Kantorovich region at $\lambda = \lambda^*$ is shown in Figure 3.1(c).

Transformed back to (ξ, η) space, Figure 3.1(d) reveals a much larger domain of theoretically assured convergence. This "better" formulation of the problem is, however, not at all evident. In contrast, a convergence theory that is invariant under linear transformations automatically includes the "best" formulation.

Remark 3.3. The invariance of Newton's method is not limited to transformations of type (3.1). In fact, Newton's method is invariant under arbitrary transformations of domain and image space; i.e., it behaves exactly the same for $AF_c(B\tilde{z}) = 0$ as for $F_c(z) = 0$—see [5]. Because F_c has a special gradient structure in the optimization context, *meaningful* transformations are coupled due to the chain rule. Meaningful transformations result from transformations of the underlying optimization problem, i.e., transformations of the domain space and the image space of the constraints. Those are of the type

$$\begin{pmatrix} B_1^\star & 0 \\ 0 & B_2^\star \end{pmatrix} F_c \left(\begin{pmatrix} B_1 & 0 \\ 0 & B_2 \end{pmatrix} \begin{pmatrix} \tilde{x} \\ \tilde{\lambda} \end{pmatrix} \right).$$

For such general transformations there is no possibility of defining a norm in an invariant way, since both the domain and the image space of the constraints are transformed independently: $B_2^\star e(B_1\tilde{x})$. For this reason, different types of transformations have been studied for different problems; see, e.g., [6, 7, 10].

4. Affine invariant convergence theory. To formulate the convergence theory and termination criteria in terms of an appropriate norm, we use a norm that is invariant under the transformation (3.1).

DEFINITION 4.1. *Let $z \in U$. Then the norms $\|\cdot\|_z : Z \to \mathbb{R}$, $z \in U$, are called* affine invariant *for* (OS) *if*

$$\|\nabla F_c(\tilde{z})\Delta z\|_z = \|\nabla G_c(B^{-1}\tilde{z})B^{-1}\Delta z\|_{B^{-1}z} \quad \text{for all } \tilde{z} \in U \text{ and } \Delta z \in Z. \tag{4.1}$$

We call $\{\|\cdot\|_z\}_{z\in U}$ a γ-continuous family of invariant norms *for* (OS) *if*

$$\big|\|r\|_{z+\Delta z} - \|r\|_z\big| \le \gamma \, \|\nabla F_c(z)\Delta z\|_z \|r\|_z \tag{4.2}$$

for every $r, \Delta z \in Z$ and $z \in U$ such that $z + \Delta z \in U$.

Using affine invariant norms we are able to present an affine invariant convergence theorem for Algorithm 1.

THEOREM 4.2. *Assume that Assumption* 1 *holds and that there are constants $\omega \ge 0, \gamma \ge 0$, and a γ-continuous family of affine invariant norms $\{\|\cdot\|_z\}_{z\in U}$ such that the operator ∇F_c satisfies*

$$\|(\nabla F_c(z + s\Delta z) - \nabla F_c(z))\Delta z\|_{z+\eta\Delta z} \le s\omega\|\nabla F_c(z)\Delta z\|_z^2 \tag{4.3}$$

for $s, \eta \in [0, 1]$, $z \in U$, and $\Delta z \in Z$ such that $\operatorname{co}\{z, z + \Delta z\} \subset U$, where $\operatorname{co} A$ denotes the convex hull of A. For $k \in \mathbb{N}$ let $h_k = \omega\|F_c(z^k)\|_{z^k}$ and let

$$\mathcal{L}(z) = \Big\{\zeta \in U : \|F_c(\zeta)\|_\zeta \le \Big(1 + \frac{\gamma}{4}\|F_c(z)\|_z\Big)\|F_c(z)\|_z\Big\}. \tag{4.4}$$

Suppose that $h_0 < 2$ and that the level set $\mathcal{L}(z^0)$ is closed. Then the iterates stay in U and the residuals converge to zero at a rate of

$$h_{k+1} \le \frac{1}{2}\, h_k^2.$$

Additionally, we have

$$\|F_c(z^{k+1})\|_{z^k} \le \|F_c(z^k)\|_{z^k}. \tag{4.5}$$

Proof. By induction, assume that $\mathcal{L}(z^k)$ is closed and that $h_k < 2$ for $k \ge 0$. Due to Remark 2.5 the neighborhood U is assumed to be convex, so that $z + \eta\Delta z \in U$ for all $\eta \in [0,1]$. From $\nabla F_c(z^k)\Delta z^k = -F_c(z^k)$ we conclude that

$$\begin{aligned} F_c(z^k + \eta\Delta z^k) &= F_c(z^k) + \int_0^\eta \nabla F_c(z^k + s\Delta z^k)\Delta z^k \, ds \\ &= (1-\eta)F_c(z^k) + \int_0^\eta (\nabla F_c(z^k + s\Delta z^k) - \nabla F_c(z^k))\Delta z^k \, ds \end{aligned}$$

for all $\eta \in [0,1]$. Applying (4.2), (4.3), and $h_k = \omega\|F_c(z^k)\|_{z^k}$, and $h_k < 2$ we obtain

$$\begin{aligned} &\|F_c(z^k + \eta\Delta z^k)\|_{z^k+\eta\Delta z^k} \\ &\quad\le (1-\eta)(1+\eta\gamma\|F_c(z^k)\|_{z^k})\|F_c(z^k)\|_{z^k} + \int_0^\eta s\omega\|\nabla F_c(z^k)\Delta z^k\|_{z^k}^2 \, ds \\ &\quad= \Big((1-\eta)(1+\eta\gamma\|F_c(z^k)\|_{z^k}) + \frac{\eta^2 h_k}{2}\Big)\|F_c(z^k)\|_{z^k} \\ &\quad< \big(1 + (\eta - \eta^2)\gamma\|F_c(z^k)\|_{z^k}\big)\|F_c(z^k)\|_{z^k}. \end{aligned} \tag{4.6}$$

With $\eta \in [0,1]$,

$$\|F_c(z^k + \eta\Delta z^k)\|_{z^k+\eta\Delta z^k} \le \Big(1 + \frac{\gamma}{4}\|F_c(z^k)\|_{z^k}\Big)\|F_c(z^k)\|_{z^k}$$

holds. If $z^k + \Delta z^k \notin \mathcal{L}(z^k)$, there exists an $\bar\eta \in [0,1]$ such that $z^k + \bar\eta\Delta z^k \in U\backslash\mathcal{L}(z^k)$, i.e.,

$$\|F_c(z^k + \bar\eta\Delta z^k)\|_{z^k+\bar\eta\Delta z^k} > \Big(1 + \frac{\gamma}{4}\|F_c(z^k)\|_{z^k}\Big)\|F_c(z^k)\|_{z^k},$$

which is a contradiction. Hence, $z^{k+1} \in \mathcal{L}(z^k)$ and, inserting $\eta = 1$ in (4.6),

$$\|F(z^{k+1})\|_{z^{k+1}} \le \omega\|F_c(z^k)\|_{z^k}^2/2.$$

Thus, we have $h_{k+1} \le h_k^2/2$ and $\mathcal{L}(z^{k+1}) \subset \mathcal{L}(z^k)$. Since $\mathcal{L}(z^k)$ is closed, every Cauchy sequence in $\mathcal{L}(z^{k+1})$ converges to a limit point in $\mathcal{L}(z^k)$, which is, by (4.4) and the continuity of the norm, also contained in $\mathcal{L}(z^{k+1})$. Hence, $\mathcal{L}(z^{k+1})$ is closed. Finally, using $\eta = 1$ in (4.6), the result (4.5) is obtained. ∎

Remark 4.3. We choose simplicity over sharpness here. The definition of the level set $\mathcal{L}(z)$ can be sharpened somewhat by a more careful estimate of the term $(\gamma\|F_c(z^k)\|_{z^k} - 1)\eta + (h_k/2 - \gamma\|F_c(z^k)\|_{z^k})\eta^2$.

Theorem 4.2 guarantees that $\lim_{k\to\infty} h_k = 0$. To ensure that $z^k \to z^*$ in Z as $k \to \infty$ we have to require that the canonical norm $\|\cdot\|_Z$ on Z can be bounded appropriately by the affine invariant norms $\|\cdot\|_z$.

COROLLARY 4.4. *If, in addition to the assumptions of Theorem* 4.2*, there exists a constant* $\tilde C > 0$ *such that*

$$\|\zeta\|_Z \le \tilde C\,\|\nabla F_c(z)\zeta\|_z \quad \textit{for all } \zeta \in Z \textit{ and } z \in U,$$

then the iterates converge to the solution $z^* = (x^*, \lambda^*)$ *of* (OS).

Proof. By assumption and Theorem 4.2 we have

$$\|\Delta z^k\|_Z \le \tilde{C}\,\|F_c(z^k)\|_{z^k} \le \tilde{C}\left(\frac{h_0}{2}\right)^k \|F_c(z^0)\|_{z^0}.$$

Thus, $\{z^k\}_{k\in\mathbb{N}}$ is a Cauchy sequence in $\mathcal{L}(z^0) \subset U$. Since $\mathcal{L}(z^0)$ is closed, the claim follows by Remark 2.5(d). ∎

For actual implementation of Algorithm 1 we need a *convergence monitor* indicating whether or not the assumptions of Theorem 4.2 may be violated and a *termination criterion* deciding whether or not the desired accuracy has been achieved.

From (4.5), a new iterate z^{k+1} is accepted whenever

$$\|F_c(z^{k+1})\|_{z^k} < \|F_c(z^k)\|_{z^k}. \tag{4.7}$$

Otherwise, the assumptions of Theorem 4.2 are violated and the iteration is considered to be nonconvergent. The use of the norm $\|\cdot\|_{z^k}$ for both the old and the new iterate permits an efficient implementation. Since in many cases the norm $\|F_c(z^{k+1})\|_{z^k}$ is defined in terms of $\overline{\Delta z^{k+1}} = \nabla F_c(z^k)^{-1} F_c(z^{k+1})$, the derivative need not be evaluated at the new iterate. If a factorization of $\nabla F_c(z^k)$ is available via a direct solver, it can be reused at negligible cost even if the convergence test fails. If an iterative solver is used, $\overline{\Delta z^{k+1}}$ in general provides a good starting point for computing Δz^{k+1} such that the additional cost introduced by the convergence monitor is minor.

The SQP iteration will be terminated with a solution z^{k+1} as soon as

$$\|F_c(z^{k+1})\|_{z^k} \le \mathrm{TOL}\;\|F_c(z^0)\|_{z^0}$$

with a user-specified tolerance TOL. Again, the use of the norm $\|\cdot\|_{z^k}$ allows an efficient implementation.

5. Invariant norms for optimization problems. What remains to be done is the construction of a γ-continuous family of invariant norms. In this section we introduce two different norms.

5.1. First invariant norm. The first norm takes advantage of the parameter c in the augmented Lagrangian. As we mentioned in Remark 2.5, there exists a $\bar{c} \ge 0$ such that $L_c''(z)$ is coercive on X for all $z \in U$ and $c \ge \bar{c}$. Hence, the operator $L_c''(z)^{-1}$ belongs to $\mathcal{B}(Z)$ for all $c \ge \bar{c}$.

Let us introduce the operator $S_c : U \to \mathcal{B}(Z)$ by

$$S_c(z) = \begin{pmatrix} L_c''(z) & 0 \\ 0 & I \end{pmatrix} \quad \text{for all } z \in U \text{ and } c \ge 0. \tag{5.1}$$

Since $L_c''(z)$ is self-adjoint for all $z \in U$, $S_c(z)$ is self-adjoint as well. Due to (2.3) the operator $S_c(z)$ is coercive for all $z \in U$ and $c \ge \bar{c}$. Thus, for all $z \in U$

$$\|S_c^{1/2}(z)\cdot\| = \sqrt{\langle S_c(z)\cdot,\cdot\rangle_Z} \tag{5.2}$$

is a norm on Z for $c \ge \bar{c}$.

PROPOSITION 5.1. *Let $c \ge \bar{c}$. Then for every $z \in U$ the mapping*

$$\|r\|_z = \|S_c(z)^{1/2}\nabla F_c(z)^{-1} r\| \quad \textit{for } r \in Z \tag{5.3}$$

defines an affine invariant norm for (2.2).

Proof. Let $z \in U$ be arbitrary. Since $\|S_c^{1/2}(z)\cdot\|$ defines a norm on Z for $c \geq \bar{c}$ and $\nabla F_c(z)$ is continuously invertible by Remark 2.6, it follows that $\|\cdot\|_z$ is a norm on Z. Now we prove the invariance property (4.1). Let $\tilde{L}_c$ denote the augmented Lagrangian associated with the transformed problem (3.1). Then we have $\tilde{L}_c''(\zeta) = \tilde{B}^\star L_c''(z)\tilde{B}$ for $z = B\zeta \in U$. Hence, setting $\tilde{S}_c(\zeta) = B^\star S_c(z)B$ we get

$$\|r\|_\zeta = \|\tilde{S}_c(\zeta)^{1/2}\nabla G_c(\zeta)^{-1}r\| \quad \text{for } r \in Z.$$

From (3.3) we conclude that

$$\nabla F_c(z)^{-1}\nabla F_c(\tilde{z}) = B\nabla G_c(\zeta)^{-1}\nabla G_c(\tilde{\zeta})B^{-1} \tag{5.4}$$

with $z = B\zeta, \tilde{z} = B\tilde{\zeta} \in U$. Using (5.3) and (5.4) we obtain

$$\begin{aligned}\|\nabla F_c(\tilde{z})\delta z\|_z &= \|S_c(z)^{1/2}\nabla F_c(z)^{-1}\nabla F_c(\tilde{z})\delta z\| \\ &= \|S_c(z)^{1/2}B\nabla G_c(\zeta)^{-1}\nabla G_c(\tilde{\zeta})B^{-1}\delta z\| \\ &= \|(B^\star S_c(z)B)^{1/2}\nabla G_c(\zeta)^{-1}\nabla G_c(\tilde{\zeta})B^{-1}\delta z\| = \|\nabla G_c(\tilde{\zeta})B^{-1}\delta z\|_\zeta,\end{aligned}$$

which gives the claim. ∎

In order to show the γ-continuity (4.2) required for Theorem 4.2, we need the following lemma.

LEMMA 5.2. *Suppose that $c \geq \bar{c}$ and that there exists a constant $\omega \geq 0$ such that*

$$\|(\nabla F_c(z+\delta z) - \nabla F_c(z))\zeta\|_{z+\delta z} \leq \omega\, \|\nabla F_c(z)\delta z\|_z \|\nabla F_c(z)\zeta\|_z \tag{5.5}$$

for all $\zeta \in Z$, $z \in U$, and $\delta z \in Z$ such that $z + \delta z \in U$. Then we have

$$\|S_c(z+\delta z)^{1/2}\zeta\| \leq \sqrt{1 + \omega(1+C_e)\, \|\nabla F_c(x,\lambda)\delta z\|_z}\, \|S_c(z)^{1/2}\zeta\|,$$

where

$$C_e = \sup\left\{ \frac{\|e'(x)\xi\|_Y^2}{\langle L_c''(x,\lambda)\xi,\xi\rangle_X} : (x,\lambda) \in \overline{U}, \xi \in X\backslash\{0\} \right\} > 0.$$

Proof. Let $\zeta = (\zeta_1,\zeta_2)^\mathsf{T} \in Z$ and $z \in U$. From (5.1) and (5.2) we infer

$$\begin{aligned}\|S_c(z+\delta z)^{1/2}\zeta\|^2 &= \langle S_c(z+\delta z)\zeta,\zeta\rangle_Z \\ &= \langle S_c(z)\zeta,\zeta\rangle_Z + \langle (S_c(z+\delta z) - S_c(z))\zeta,\zeta\rangle_Z \\ &\leq \|S_c(z)^{1/2}\zeta\|^2 + \langle (L_c''(z+\delta z) - L_c''(z))\zeta_1,\zeta_1\rangle_X.\end{aligned} \tag{5.6}$$

By assumption, $S_c(z)$ is continuously invertible. Utilizing the Lipschitz assumption (5.5), the second additive term on the right-hand side can be estimated as

$$\begin{aligned}&\langle (L_c''(z+\delta z) - L_c''(z))\zeta_1,\zeta_1\rangle_X \\ &\quad= \langle (\nabla F_c(z+\delta z) - \nabla F_c(z))(\zeta_1,0)^\mathsf{T}, (\zeta_1,0)^\mathsf{T}\rangle_Z \\ &\quad= \langle \nabla F_c(z)S_c(z)^{-1}S_c(z)\nabla F_c(z)^{-1}(\nabla F_c(z+\delta z) - \nabla F_c(z))(\zeta_1,0)^\mathsf{T}, (\zeta_1,0)^\mathsf{T}\rangle_Z \\ &\quad= \langle S_c(z)\nabla F_c(z)^{-1}(\nabla F_c(z+\delta z) - \nabla F_c(z))(\zeta_1,0)^\mathsf{T}, S_c(z)^{-1}\nabla F_c(z)(\zeta_1,0)^\mathsf{T}\rangle_Z \\ &\quad\leq \|S_c(z)^{1/2}\nabla F_c(z)^{-1}(\nabla F_c(z+\delta z) - \nabla F_c(z))(\zeta_1,0)^\mathsf{T}\| \\ &\qquad\cdot \|S_c(z)^{-1/2}\nabla F_c(z)(\zeta_1,0)^\mathsf{T}\| \\ &\quad= \|(\nabla F_c(z+\delta z) - \nabla F_c(z))(\zeta_1,0)^\mathsf{T}\|_z \|S_c(z)^{-1/2}\nabla F_c(z)(\zeta_1,0)^\mathsf{T}\| \\ &\quad\leq \omega\, \|\nabla F_c(z)\delta z\|_z\, \|\nabla F_c(z)(\zeta_1,0)^\mathsf{T}\|_z\, \|S_c(z)^{-1/2}\nabla F_c(z)(\zeta_1,0)^\mathsf{T}\| \\ &\quad\leq \omega\, \|\nabla F_c(z)\delta z\|_z\, \|S_c(z)^{1/2}\zeta\|\, \|S_c(z)^{-1/2}\nabla F_c(z)(\zeta_1,0)^\mathsf{T}\|.\end{aligned}$$

Note that

$$\begin{aligned}&\|S_c(z)^{-1/2}\nabla F_c(z)(\zeta_1,0)^{\mathsf{T}}\|^2\\&\quad=\langle\nabla F_c(z)(\zeta_1,0)^{\mathsf{T}},S_c(z)^{-1}\nabla F_c(z)(\zeta_1,0)^{\mathsf{T}}\rangle_Z=\langle L_c''(z)\zeta_1,\zeta_1\rangle_X+\|e'(x)\zeta_1\|_Y^2\\&\quad\le(1+C_e)\ \langle L_c''(z)\zeta_1,\zeta_1\rangle_X=(1+C_e)\ \|S_c(z)^{1/2}(\zeta_1,0)^{\mathsf{T}}\|^2.\end{aligned}$$

This implies

$$\langle(L_c''(z+\delta z)-L_c''(z))\zeta_1,\zeta_1\rangle_X\le\omega(1+C_e)\ \|\nabla F_c(z)\delta z\|_z\ \|S_c(z)^{1/2}\zeta\|^2. \tag{5.7}$$

Inserting (5.7) into (5.6), the claim follows. ∎

PROPOSITION 5.3. *Let all hypotheses of Lemma* 5.2 *be satisfied. Then* $\{\|\cdot\|_z\}_{z\in U}$ *is an* $\omega(3+C_e)/2$*-continuous family of invariant norms with*

$$\|\zeta\|_Z\le\frac{1}{\sqrt{\tilde\kappa}}\ \|\nabla F_c(z)\zeta\|_z \tag{5.8}$$

for all $\zeta\in Z$ *and* $z\in U$*, where* $\tilde\kappa>0$ *was introduced in* (2.3).

Proof. From (5.3) it follows that

$$\begin{aligned}\|r\|_{z+\delta z}&\le\|S_c(z+\delta z)^{1/2}\nabla F_c(z)^{-1}r\|\\&\quad+\|S_c(z+\delta z)^{1/2}(\nabla F_c(z+\delta z)^{-1}-\nabla F_c(z)^{-1})r\|.\end{aligned}$$

We estimate the additive terms on the right-hand side separately. Using Lemma 5.2 we find

$$\|S_c(z+\delta z)^{1/2}\nabla F_c(z)^{-1}r\|\le\sqrt{1+\omega(1+C_e)\|\nabla F_c(z)\delta z\|_z}\|r\|_z.$$

Applying (5.3) and (5.5) we obtain

$$\begin{aligned}&\|S_c(z+\delta z)^{1/2}(\nabla F_c(z+\delta z)^{-1}-\nabla F_c(z)^{-1})r\|\\&\quad=\|S_c(z+\delta z)^{1/2}\nabla F_c(z+\delta z)^{-1}(\nabla F_c(z)-\nabla F_c(z+\delta z))\nabla F_c(z)^{-1}r\|\\&\quad=\|(\nabla F_c(z)-\nabla F_c(z+\delta z))\nabla F_c(z)^{-1}r\|_{z+\delta z}\le\omega\ \|\nabla F_c(z)\delta z\|_z\|r\|_z.\end{aligned}$$

Hence, using $\sqrt{1+x}\le1+x/2$ for $x\ge0$,

$$\|r\|_{z+\delta z}\le\Big(1+\frac{\omega}{2}(3+C_e)\|\nabla F_c(z)\delta z\|_z\Big)\|r\|_z,$$

and it follows that $\{\|\cdot\|_z\}_{z\in U}$ is an $\omega(3+C_e)/2$-continuous family of invariant norms. Finally, from

$$\begin{aligned}\|\nabla F_c(z)\zeta\|_z^2&=\langle S_c(z)\nabla F_c(z)^{-1}\nabla F_c(z)\zeta,\nabla F_c(z)^{-1}\nabla F_c(z)\zeta\rangle_Z\\&=\langle S_c(z)\zeta,\zeta\rangle_Z\ge\tilde\kappa\ \|\zeta\|_Z^2\end{aligned}$$

we infer (5.8). ∎

5.2. Second invariant norm. In section 5.1 we introduced an invariant norm, provided the augmentation parameter in Algorithm 1 satisfies $c\ge\bar c$. But in many applications the constant $\bar c$ is not explicitly known. Thus, $L_c''(x,\lambda)^{-1}$ need not to be bounded for $c\in[0,\bar c)$, so that $S_c(x,\lambda)$ given by (5.1) might be singular. To overcome these difficulties we define a second invariant norm that is based on a splitting $X=$

$\ker e'(x) \oplus \bar{X}$ such that at least the coercivity of $L_0''(x,\lambda)$ on $\ker e'(x)$ can be utilized. Even though the thus-defined norm can be used with $c = 0$, a larger value of c may improve the global convergence properties—see [16, section 2.3].

To begin with, let us introduce the bounded linear operator $T_c(x,\lambda) : \ker e'(x) \times Y \to X$ by

$$T_c(x,\lambda) = (L_c''(x,\lambda)\ \ e'(x)^\star) \quad \text{for } (x,\lambda) \in U \text{ and } c \geq 0.$$

LEMMA 5.4. *For every $(x,\lambda) \in U$ and $c \geq 0$ the operator $T_c(x,\lambda)$ is an isomorphism.*

Proof. Let $r \in X$ be arbitrary. Then the equation $T_c(x,\lambda)\zeta = r$ for $\zeta = (\zeta_1,\zeta_2)^\mathsf{T} \in \ker e'(x) \times Y$ is equivalent to

$$\nabla F_c(x,\lambda) \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} = \begin{pmatrix} r \\ 0 \end{pmatrix}. \tag{5.9}$$

Due to Remark 2.6 the operator $\nabla F_c(x,\lambda)$ is continuously invertible for all $(x,\lambda) \in U$ and $c \geq 0$. Thus, ζ is uniquely determined by (5.9), and the claim follows. ∎

We define the bounded linear operator $R_c(x,\lambda) : \ker e'(x) \times Y \to Z$ as

$$R_c(x,\lambda) = \begin{pmatrix} L_c''(x,\lambda) & 0 \\ 0 & I \end{pmatrix} \quad \text{for } (x,\lambda) \in U \text{ and } c \geq 0. \tag{5.10}$$

Note that $R_c(x,\lambda)$ is coercive and self-adjoint. Next we introduce the invariant norm

$$\|(r_1,r_2)^T\|_z^2 = \langle R_c(z)T_c(z)^{-1}r_1, T_c(z)^{-1}r_1\rangle_{Z\times Y} + \|r_2\|_Y^2 \tag{5.11}$$

for $z \in U$ and $(r_1,r_2)^T \in Z$. To shorten notation, we write $\|R_c(z)^{1/2}T_c(z)^{-1}r_1\|^2$ for the first additive term.

PROPOSITION 5.5. *For every $z \in U$ the mapping given by* (5.11) *is an affine invariant norm for* (OS)*, which is equivalent to the usual norm on Z.*

Proof. Let $z \in U$ be arbitrary. Since $R_c(z)$ is coercive and $T_c(z)$ is continuously invertible, it follows that $\|\cdot\|_z$ defines a norm which is indeed equivalent to the usual norm on Z. Now we prove the invariance property (4.1). For $(x,\lambda) = (\tilde{B}y,\xi) \in U$ we have

$$(\tilde{B}^\star L_c''(y,\xi)\tilde{B}\ \ \tilde{B}^\star e'(y)^\star) = B^\star T_c(x,\lambda) \begin{pmatrix} B & 0 \\ 0 & I \end{pmatrix}. \tag{5.12}$$

Utilizing (3.3), (5.11), and (5.12) the invariance property follows. ∎

The following proposition guarantees that $\{\|\cdot\|_z\}_{z\in U}$ is a γ-continuous family of invariant norms for (OS).

PROPOSITION 5.6. *Suppose that there exists a constant $\omega \geq 0$ such that*

$$\|(\nabla F_c(z+\delta z) - \nabla F_c(z))\zeta\|_{z+\delta z} \leq \omega\ \|\nabla F_c(z)\delta z\|_z \|\nabla F_c(z)\zeta\|_z \tag{5.13}$$

for all $\zeta \in Z$, $z \in U$, and $\delta z \in Z$ such that $z + \delta z \in U$. Then we have

$$\|r\|_{z+\delta z} \leq \Big(1 + \frac{3\omega}{2}\ \|\nabla F_c(z)\delta z\|_z\Big)\|r\|_z.$$

For the proof of the previous proposition, we will use the following lemmas.

LEMMA 5.7. *With the assumption of Proposition* 5.6 *holding and* $z = (x, \lambda)$*, it follows that*

$$\|R_c(z+\delta z)^{1/2}\zeta\| \le \sqrt{1+\omega\|\nabla F_c(z)\delta z\|_z}\,\|R_c(z)^{1/2}\zeta\|$$

for all $\zeta \in \ker e'(x) \times Y$ *and* $c \ge 0$.

Proof. Let $z = (x,\lambda) \in U$ and $\zeta = (\zeta_1,\zeta_2)^{\mathsf{T}} \in \ker e'(x) \times Y$. Using (5.10) and (5.11) we obtain

$$\|R_c(z+\delta z)^{1/2}\zeta\|^2 \le \|R_c(z)^{1/2}\zeta\|^2 + \langle (L_c''(z+\delta z) - L_c''(z))\zeta_1, \zeta_1\rangle_X. \tag{5.14}$$

For all $c \ge 0$ the operator $R_c(z)$ is continuously invertible. Furthermore, $R_c(z)$ is self-adjoint. Thus, applying (5.13) and

$$\nabla F_c(z)(\zeta_1,0)^{\mathsf{T}} = T_c(z)(\zeta_1,0)^{\mathsf{T}} = R_c(z)(\zeta_1,0)^{\mathsf{T}},$$

we can estimate the second additive term on the right-hand side of (5.14) as

$$\begin{aligned}
&\langle (L_c''(z+\delta z) - L_c''(z))\zeta_1,\zeta_1\rangle_X \\
&\quad = \langle T_c(z)R_c(z)^{-1}R_c(z)T_c(z)^{-1}(L_c''(z+\delta z) - L_c''(z))\zeta_1,\zeta_1\rangle_Z \\
&\quad = \langle R_c(z)T_c(z)^{-1}(L_c''(z+\delta z) - L_c''(z))\zeta_1, R_c(z)^{-1}T_c(z)^{\star}\zeta_1\rangle_{Z\times Y} \\
&\quad \le \|R_c(z)^{1/2}T_c(z)^{-1}(L_c''(z+\delta z) - L_c''(z))\zeta_1\|\ \|R_c(z)^{-1/2}T_c(z)^{\star}\zeta_1\| \\
&\quad \le \|(\nabla F_c(z+\delta z) - \nabla F_c(z))(\zeta_1,0)^{\mathsf{T}}\|_z\ \|R_c(z)^{-1/2}T_c(z)^{\star}\zeta_1\| \\
&\quad \le \omega\|\nabla F_c(z)\delta z\|_z\|\nabla F_c(z)(\zeta_1,0)^{\mathsf{T}}\|_z\|R_c(z)^{1/2}(\zeta_1,0)^{\mathsf{T}}\| \\
&\quad = \omega\ \|\nabla F_c(z)\delta z\|_z\|R_c(z)^{1/2}\zeta\|^2.
\end{aligned}$$

Inserting this bound in (5.14), the claim follows. ∎

LEMMA 5.8. *Let the assumptions of Theorem* 5.6 *be satisfied. Then*

$$\|((T_c(z+\delta z) - T(z))T_c(z)^{-1}r, 0)^{\mathsf{T}}\|_{z+\delta z} \le \omega\|\nabla F_c(z)\delta z\|_z\|(r,0)^{\mathsf{T}}\|_z \quad \textit{for all } r \in X.$$

Proof. For arbitrary $r \in X$ we set $\zeta = (\zeta_1,\zeta_2)^{\mathsf{T}} = T_c(z)^{-1}r$. Using (5.9) and (5.13) we estimate

$$\begin{aligned}
&\|((T_c(z+\delta z) - T_c(z))T_c(z)^{-1}r,0)^{\mathsf{T}}\|_{z+\delta z} \\
&\quad \le \|(\nabla F_c(z+\delta z) - \nabla F_c(z))(\zeta_1,\zeta_2)^{\mathsf{T}}\|_{z+\delta z} \le \omega\ \|\nabla F_c(z)\delta z\|_z\|\nabla F_c(z)(\zeta_1,\zeta_2)^{\mathsf{T}}\|_z \\
&\quad = \omega\ \|\nabla F_c(z)\delta z\|_z\|(r,0)^{\mathsf{T}}\|_z,
\end{aligned}$$

so that the claim follows. ∎

Proof of Proposition 5.6. Let $z, z+\delta z \in U$. Utilizing (5.11) and Lemmas 5.7 and 5.8 we find

$$\begin{aligned}
\|(r_1,0)^{\mathsf{T}}\|_{z+\delta z} &= \|R_c(z+\delta z)^{1/2}T_c(z+\delta z)^{-1}r_1\| \\
&\le \|R_c(z+\delta z)^{1/2}T_c(z)^{-1}r_1\| + \|R_c(z+\delta z)^{1/2}(T_c(z+\delta z)^{-1} - T_c(z)^{-1})r_1\| \\
&\le \sqrt{1+\omega\|\nabla F_c(z)\delta z\|_z}\|(r_1,0)^{\mathsf{T}}\|_z + \|((T_c(z) - T_c(z+\delta z))T_c(z)^{-1}r_1,0)^{\mathsf{T}}\|_{z+\delta z} \\
&\le \sqrt{1+\omega\|\nabla F_c(z)\delta z\|_z}\|(r_1,0)^{\mathsf{T}}\|_z + \omega\|\nabla F_c(z)\delta z\|_z\|(r_1,0)^{\mathsf{T}}\|_z \\
&\le \Big(1 + \frac{3\omega}{2}\ \|\nabla F_c(z)\delta z\|_z\Big)\|(r_1,0)^{\mathsf{T}}\|_z,
\end{aligned}$$

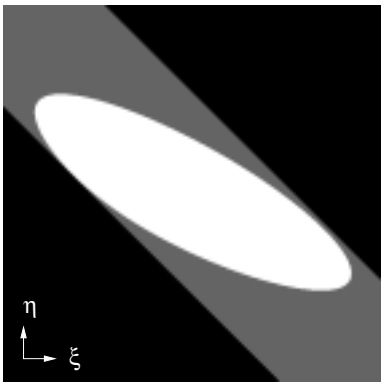


FIG. 5.1. *Illustration for Examples* 3.2 *and* 5.10. *Neighborhood* $U(x^*)$ *(gray) and affine invariant domain of theoretically assured convergence (white).*

and therefore

$$\begin{aligned}\|r\|_{z+\delta z}^2 &= \|(r_1,0)^{\mathsf{T}}\|_{z+\delta z}^2 + \|r_2\|^2 \\ &\le \Big(1+\frac{3\omega}{2}\,\|\nabla F_c(z)\delta z\|_z\Big)^2 \|(r_1,0)^{\mathsf{T}}\|_z^2 + \|r_2\|^2 \\ &\le \Big(1+\frac{3\omega}{2}\,\|\nabla F_c(z)\delta z\|_z\Big)\|r\|_z^2.\end{aligned}$$

Hence, $\{\|\cdot\|_z\}_{z\in U}$ is a $3\omega/2$-continuous family of invariant norms. ∎

Remark 5.9. Note that the Lipschitz constant of the second norm does not involve C_e and hence is independent of the choice of c. In contrast, choosing c too small may lead to a large Lipschitz constant of the first norm and thus can affect the algorithm.

Example 5.10. Let us return to Example 3.2. Using the second norm with $c = 0$, the theoretically assured, affine invariant domain of convergence is shown in Figure 5.1, to be compared with Figures 3.1(b) and (d). Its shape and size is clearly more similar to the noninvariant domain of convergence for the "better" formulation and, by definition, does not change when the coordinates change.

5.3. Computational efficiency. The affine invariance of the two norms developed in the previous sections does not come free: the evaluation of the norms is more involved than the evaluation of some standard norm.

Nevertheless, the computational overhead of the first norm defined in section 5.1 is almost negligible, since it can in general be implemented by one additional matrix vector multiplication. It requires, however, a sufficiently large parameter c.

On the other hand, the second norm defined in section 5.2 works for arbitrary $c \ge 0$ but requires one additional system solve with the same Jacobian but different right-hand side. In the case in which a factorization of the matrix is available, the computational overhead is negligible—compare the CPU times of the exact Newton method in section 7. If, however, the system is solved iteratively, the additional system solve may incur a substantial cost, in which case the first norm should be preferred.

5.4. Connection to the optimization problem. When solving optimization problems of type (P), feasibility $e(x) = 0$ and optimality are the relevant quantities. This is well reflected by the proposed norms $\|\cdot\|_z$. Let $z = (x, \lambda)$ and

$\Delta z = (\Delta x, \Delta\lambda)^{\mathsf{T}} = -\nabla F_c(z)^{-1} F_c(z)$. Using Taylor's theorem (see [19, p. 148]) and the continuity of L_0'', we obtain for the first norm

$$\begin{aligned}
\|F_c(z)\|_z^2 &= \langle S_c(z)\Delta z, \Delta z\rangle_Z \\
&= \langle L_c''(z)\Delta x, \Delta x\rangle_X + \|\Delta\lambda\|_Y^2 \\
&= \langle L_0''(z)\Delta x, \Delta x\rangle_X + c\|e'(x)\Delta x\|_Y^2 + \|\Delta\lambda\|_Y^2 \\
&= \langle L_0''(z^*)\Delta x, \Delta x\rangle_X + o(\|z^* - z\|_Z^2) + c\|e'(x)\Delta x\|_Y^2 + \|\Delta\lambda\|_Y^2 \\
&= 2(L_0(z) - L_0(z^*)) + o(\|z^* - z\|_Z^2) + c\|e(x)\|_Y^2 + \|\Delta\lambda\|_Y^2 \\
&= 2\big(J(x) - J(x^*) - \langle\lambda, e(x)\rangle_Y\big) + c\|e(x)\|_Y^2 + \|\Delta\lambda\|_Y^2 + o(\|z^* - z\|_Z^2).
\end{aligned}$$

The second norm is based on the partitioning $F_c(x,\lambda) = (L_c'(x,\lambda), e(x))^{\mathsf{T}}$ and correspondingly on a splitting of the Newton correction into an *optimizing* direction $\nabla F_c(x,\lambda)(\zeta_1,\zeta_2)^{\mathsf{T}} = -(L_c'(x,\lambda), 0)^{\mathsf{T}}$ tangential to the constraints manifold and a *feasibility* direction $\nabla F_c(x,\lambda)(\xi_1,\xi_2)^{\mathsf{T}} = -(0, e(x))^{\mathsf{T}}$. Since $e'(x)\zeta_1 = 0$, we have for $z = (x,\lambda)$

$$\begin{aligned}
\|F_c(z)\|_z^2 &= \langle L_c''(z)\zeta_1, \zeta_1\rangle_X + \|\zeta_2\|_Y^2 + \|e(x)\|_Y^2 \\
&= \langle L_0''(z)\zeta_1, \zeta_1\rangle_X + \|\zeta_2\|_Y^2 + \|e(x)\|_Y^2 \\
&= \langle L_0''(z^*)\zeta_1, \zeta_1\rangle_X + o(\|z^* - z\|_Z^2) + \|\zeta_2\|_Y^2 + \|e(x)\|_Y^2 \\
&= 2(L_0(z) - L_0(z^*)) + \|\zeta_2\|_Y^2 + \|e(x)\|_Y^2 + o(\|z^* - z\|_Z^2) \\
&= 2\big(J(x) - J(x^*) - \langle\lambda, e(x)\rangle_Y\big) + \|e(x)\|_Y^2 + \|\zeta_2\|_Y^2 + o(\|z^* - z\|_Z^2).
\end{aligned}$$

Recall that $\Delta\lambda = \zeta_2 + \xi_2$. Thus, in the proximity of the solution, both affine invariant norms measure the quantities we are interested in when solving optimization problems, in addition to the error in the Lagrange multiplier and the optimizing direction's Lagrange multiplier component, respectively.

6. Inexact augmented Lagrangian-SQP methods. Taking into account discretization errors or truncation errors resulting from iterative solution of linear systems, we have to consider inexact Newton methods, where an inner residual remains:

$$\begin{aligned}
\nabla F_c(z^k)\delta z^k &= -F_c(z^k) + r^k, \\
z^{k+1} &= z^k + \delta z^k.
\end{aligned} \tag{6.1}$$

Such inexact Newton methods have been studied in a nonaffine invariant setting by Dembo, Eisenstat, and Steihaug [4] and Bank and Rose [1].

With slightly stronger assumptions than before and a suitable control of the inner residual, a similar convergence theory can be established as in section 4.

Note that exact affine invariance is preserved only in the case in which the inner iteration is affine invariant, too.

THEOREM 6.1. *Assume that Assumption* 1 *holds and that there are constants* $\omega \geq 0$, $\gamma \geq 0$, *and a* γ*-continuous family of affine invariant norms* $\{\|\cdot\|_z\}_{z\in U}$ *such that the operator* ∇F_c *satisfies*

$$\|(\nabla F_c(z + s\delta z) - \nabla F_c(z))\delta z\|_{z+\eta\delta z} \leq \omega s\|\nabla F_c(z)\delta z\|_z^2 \tag{6.2}$$

for $s, \eta \in [0,1]$, $z \in U$, *and* $\delta z \in Z$ *such that* $z + \delta z \in U$. *Choose some* $0 < \Theta < 1$ *and define the level sets*

$$\mathcal{L}(z) = \Big\{\zeta \in U : \|F_c(\zeta)\|_\zeta \leq \Big(1 + \frac{\gamma\Theta}{2\omega}\Big)\|F_c(z)\|_z\Big\}.$$

Suppose that $z^0 \in U$ and that $\mathcal{L}(z^0)$ is closed. If the inner residual r^k resulting from the inexact solution of the Newton correction (6.1) *is bounded by*

$$\|r^k\|_{z^k} \le \delta_k \, \|F_c(z^k)\|_{z^k}, \tag{6.3}$$

where

$$(1+\gamma\|\nabla F_c(z^k)\delta z^k\|_{z^k})\delta_k + (1+\delta_k)\frac{\omega}{2}\|\nabla F_c(z^k)\delta z^k\|_{z^k} \le \Theta, \tag{6.4}$$

then the iterates stay in U and the residuals converge to zero as $k \to \infty$ at a rate of

$$\|F_c(z^{k+1})\|_{z^{k+1}} \le \Theta \|F_c(z^k)\|_{z^k}. \tag{6.5}$$

Proof. Analogously to the proof of Theorem 4.2, one obtains

$$F_c(z^k+\eta\delta z^k) = (1-\eta)F_c(z^k) + \eta r^k + \int_0^\eta (\nabla F_c(z^k+s\delta z^k) - \nabla F_c(z^k))\delta z^k \, ds \tag{6.6}$$

for all $\eta \in [0,1]$. Using (6.6), (6.2), (4.2), and (6.3), we find for $\phi \in [0,1]$

$$\begin{aligned}
&\|F_c(z^k+\eta\delta z^k)\|_{z^k+\phi\delta z^k} \\
&\quad \le (1-\eta)\|F_c(z^k)\|_{z^k+\phi\delta z^k} + \eta\|r^k\|_{z^k+\phi\delta z^k} + \int_0^\eta s\omega\|\nabla F_c(z^k)\delta z^k\|_{z^k}^2 \, ds \\
&\quad \le (1+\gamma\phi\|\nabla F_c(z^k)\delta z^k\|_{z^k}) \left((1-\eta)\|F_c(z^k)\|_{z^k} + \eta\|r^k\|_{z^k}\right) \\
&\qquad + \frac{\omega\eta^2}{2} \, \|\nabla F_c(z^k)\delta z^k\|_{z^k}^2 \\
&\quad \le (1+\gamma\phi\|\nabla F_c(z^k)\delta z^k\|_{z^k})(1-\eta+\delta_k\eta)\|F_c(z^k)\|_{z^k}) \\
&\qquad + \frac{\omega\eta^2}{2} \, \|\nabla F_c(z^k)\delta z^k\|_{z^k}^2.
\end{aligned} \tag{6.7}$$

From (6.1) and (6.3) we have

$$(1-\delta_k)\|F_c(z^k)\|_{z^k} \le \|\nabla F_c(z^k)\delta z^k\|_{z^k} = \|F_c(z^k) - r^k\|_{z^k} \le (1+\delta_k)\|F_c(z^k)\|_{z^k} \tag{6.8}$$

and thus, setting $\phi = \eta$ in (6.7) and $\chi = \gamma\|\nabla F_c(z^k)\delta z^k\|_{z^k}$ and using (6.4), it follows that

$$\begin{aligned}
&\frac{\|F_c(z^k+\eta\delta z^k)\|_{z^k+\eta\delta z^k}}{\|F_c(z^k)\|_{z^k}} \\
&\quad \le (1+\gamma\eta\|\nabla F_c(z^k)\delta z^k\|_{z^k})(1-\eta+\delta_k\eta) + (1+\delta_k)\frac{\omega\eta^2}{2}\|\nabla F_c(z^k)\delta z^k\|_{z^k} \\
&\quad \le (1+\eta\chi)(1-\eta) + (1+\eta\chi)\delta_k\eta + (1+\delta_k)\frac{\omega\eta^2}{2}\|\nabla F_c(z^k)\delta z^k\|_{z^k} \\
&\quad \le (1+\eta\chi)(1-\eta) + \eta\Theta \\
&\quad \le 1 - \eta(1-\Theta) + \eta(1-\eta)\chi.
\end{aligned} \tag{6.9}$$

From (6.4) we have $\|\nabla F_c(z^k)\delta z^k\|_{z^k} \le 2\Theta/\omega$. Since $1-\Theta > 0$, we conclude that

$$\|F_c(z^k+\eta\delta z^k)\|_{z^k+\eta\delta z^k} \le \Big(1+\frac{\gamma\Theta}{2\omega}\Big)\|F_c(z^k)\|_{z^k}. \tag{6.10}$$

If $z^{k+1} \notin U$, then there is some $\eta^* \in [0,1]$ such that $\mathrm{co}\{z^k, z^k+\eta\delta z^k\} \subset U$ for $\eta \in [0,\eta^*)$ and $z^k+\eta^*\delta z^k \notin \mathcal{L}(z^k)$, i.e., $\|F_c(z^k+\eta^*\delta z^k)\|_{z^k+\eta^*\delta z^k} > (1+\gamma\Theta/(2\omega))\|F_c(z^k)\|_{z^k}$, which contradicts (6.10). Thus, $z^{k+1} \in U$. Furthermore, inserting $\eta = 1$ into (6.9) yields

$$\|F_c(z^{k+1})\|_{z^{k+1}} \le \Theta\|F_c(z^k)\|_{z^k},$$

and therefore $\mathcal{L}(z^{k+1}) \subset \mathcal{L}(z^k)$ is closed. ∎

The next corollary follows analogously as Corollary 4.4.

COROLLARY 6.2. *If, in addition to the assumptions of Theorem* 6.1*, there exists a constant* $\hat{C} > 0$ *such that*

$$\|\zeta\|_Z \le \hat{C}\|\nabla F_c(z)\zeta\|_z$$

for all $\zeta \in Z$ *and* $z \in U$*, then the iterates converge to the solution* $z^* = (x^*, \lambda^*)$ *of* (OS).

For actual implementation of an inexact Newton method following Theorem 6.1, we need to satisfy the accuracy requirement (6.4). Thus, we need not only an error estimator for the inner iteration computing δ_k but also easily computable estimates $[\omega]$ and $[\gamma]$ for the Lipschitz constants ω and γ in case no suitable theoretical values can be derived. Setting $\eta = 1$ in (6.6), we readily obtain

$$\|F_c(z^{k+1}) - r^k\|_{z^k} \le \frac{\omega}{2}\|\nabla F_c(z^k)\delta z^k\|^2_{z^k}$$

and hence a lower bound

$$\frac{2\ \|F_c(z^{k+1}) - r^k\|_{z^k}}{\|\nabla F_c(z^k)\delta z^k\|^2_{z^k}} \le \omega.$$

Unfortunately, the norms involve solutions of Newton-type systems and therefore cannot be computed exactly. Assuming the relative accuracy of evaluating the norms are $\hat{\delta}_k$ and $\tilde{\delta}_k$, respectively, we define the actually computable estimate

$$[\omega]_k = 2\frac{1-\hat{\delta}_k}{(1+\tilde{\delta}_k)^2}\frac{\|F_c(z^{k+1}) - r^k\|_{z^k}}{\|\nabla F_c(z^k)\delta z^k\|^2_{z^k}} \le \omega.$$

We would like to select a δ_k such that the accuracy matching condition (6.4) is satisfied. Unfortunately, due to the local sampling of the global Lipschitz constant ω and the inexact computation of the norms, the estimate $[\omega]_k$ is possibly too small, translating into a possibly too large tolerance for the inexact Newton correction. In order to compensate for that, we introduce a safety factor $\rho < 1$ and require the approximate accuracy matching condition

$$(1+[\gamma]_k\|\nabla F_c(z^k)\delta z^k\|_{z^k})\delta_k + (1+\delta_k)\frac{[\omega]_k}{2}\|\nabla F_c(z^k)\delta z^k\|_{z^k} \le \rho\Theta \tag{6.11}$$

to hold. An obvious choice for ρ would be $(1-\hat{\delta}_k)/(1+\tilde{\delta}_k)$. From Propositions 5.3 and 5.6 we infer that γ is of the same order of magnitude as ω. Thus we take the estimate

$$[\gamma]_k = 3[\omega]_k/2,$$

currently ignoring C_e when using the first norm.

Again, the convergence monitor (4.7) can be used to detect nonconvergence. In the inexact setting, however, the convergence monitor may also fail due to δ_k chosen too large. Therefore, whenever (4.7) is violated and a reduction of δ_k is promising (e.g., $(1+([\gamma]_k+[\omega]_k/2)\|\nabla F_c(z^k)\delta z^k\|_{z^k})\delta_k \geq [\omega]_k/10\|\nabla F_c(z^k)\delta z^k\|_{z^k}$), the Newton correction should be recomputed with reduced δ_k.

Remark 6.3. If an inner iteration is used for approximately solving the Newton equation (6.1) which provides the *orthogonality relation* $(\delta z^k, \Delta z^k - \delta z^k)_{z^k} = 0$ in a scalar product $(\cdot,\cdot)_{z^k}$ that induces the affine invariant norm, the estimate (6.11) can be tightened by substituting $(1+\delta_k)^2$ by $1+\delta_k^2$. Furthermore, the norm $\|\Delta z^k\|_{z^k}$ of the exact Newton correction is computationally available, which permits the construction of algorithms that are robust even for large inaccuracies δ_k. The application of a conjugate gradient method that is confined to the null space of the linearized constraints [2] to augmented Lagrangian-SQP methods can be the focus of future research.

7. Numerical experiments. This section is devoted to presenting numerical tests for Example 2.1 that illustrate the theoretical investigations of the previous sections. To solve (P) we apply the so-called optimize-then-discretize approach: we compute an approximate solution by discretizing Algorithm 1, i.e., by discretizing the associated system (2.6). In the context of Example 2.1 we have $x^k = (y^k, u^k, v^k)$, $\delta x = (\delta y, \delta u, \delta v) \in W(0,T)\times L^2(0,T)\times L^2(0,T)$. To reduce the size of the system we take advantage of a relationship between the SQP steps δu, δv for the controls and the SQP step $\delta\lambda$ for the Lagrange multiplier. In fact, from

$$\begin{aligned}
\frac{\partial^2 L_0}{\partial u^2}(x^k,\tilde{\lambda}^k)\delta u + \frac{\partial e}{\partial u}(x^k)^\star \delta\lambda &= -\frac{\partial L_0}{\partial u}(x^k,\tilde{\lambda}^k),\\
\frac{\partial^2 L_0}{\partial v^2}(x^k,\tilde{\lambda}^k)\delta v + \frac{\partial e}{\partial v}(x^k)^\star \delta\lambda &= -\frac{\partial L_0}{\partial v}(x^k,\tilde{\lambda}^k)
\end{aligned}$$

we infer that

$$\begin{aligned}
\delta u &= -\frac{1}{\alpha}\Big(\tilde{\lambda}^k(\cdot,0) - \lambda^k(\cdot,0) + \delta\lambda(\cdot,0)\Big) \quad \text{in } (0,T),\\
\delta v &= \frac{1}{\beta}\Big(\tilde{\lambda}^k(\cdot,1) - \lambda^k(\cdot,1) + \delta\lambda(\cdot,1)\Big) \qquad \text{in } (0,T),
\end{aligned} \tag{7.1}$$

where $\tilde{\lambda}^k = \lambda^k + ce(x^k)$ by step (b) of Algorithm 1. Inserting (7.1) into (2.6) we obtain a system only in the unknowns $(\delta y, \delta\lambda)$. Note that the second Fréchet-derivative of the Lagrangian is given by

$$\langle L_0''(x^k,\tilde{\lambda}^k)\zeta,\xi\rangle_X = \int_Q \zeta_1\xi_1\big(1+2\tilde{\lambda}^k\big)\,dx + \int_0^T \alpha\zeta_2\xi_2 + \beta\zeta_3\xi_3\,dt$$

for $\zeta = (\zeta_1, \zeta_2, \zeta_3)$, $\xi = (\xi_1, \xi_2, \xi_3) \in X$. The solution $(\delta y, \delta u, \delta v, \delta\lambda)$ of (2.6) is computed as follows: First we solve

$$
\begin{aligned}
y_t - \nu y_{xx} + (y^k y)_x &= -e^k && \text{in } Q,\\
\nu y_x(\cdot,0) + \sigma_0 y(\cdot,0) + \frac{\lambda(\cdot,0)}{\alpha} &= \frac{1}{\alpha}\left(\lambda^k(\cdot,0) - \tilde\lambda^k(\cdot,0)\right) && \text{in } (0,T),\\
\nu y_x(\cdot,1) + \sigma_1 y(\cdot,1) - \frac{\lambda(\cdot,1)}{\beta} &= \frac{1}{\beta}\left(\tilde\lambda^k(\cdot,1) - \lambda^k(\cdot,1)\right) && \text{in } (0,T),\\
y(0,\cdot) &= 0 && \text{in } \Omega,\\
(1-\tilde\lambda_x^k)y - \lambda_t - \nu\lambda_{xx} - y^k\lambda_x &= y^k - z && \text{in } Q,\\
\nu\lambda_x(\cdot,0) + (y(\cdot,0)+\sigma_0)\lambda(\cdot,0) &= 0 && \text{in } (0,T),\\
\nu\lambda_x(\cdot,1) + (y(\cdot,1)+\sigma_1)\lambda(\cdot,1) &= 0 && \text{in } (0,T),\\
\lambda(T,\cdot) &= 0 && \text{in } \Omega,
\end{aligned}
\tag{7.2}
$$

where $e^k = y_t^k - \nu y_{xx}^k + y^k y_x^k - f$, and set $\delta y = y$ and $\delta\lambda = \lambda$. Then we obtain δu and δv from (7.1). For more details we refer the reader to [18].

For the time integration we use the backward Euler scheme while the spatial variable is approximated by piecewise linear finite elements. The programs were written in MATLAB, version 5.3, and executed on a Pentium III 550 MHz personal computer.

Run 7.1 (Neumann control). In the first example we choose $T = 1$, $\nu = 0.1$, $\sigma_0 = \sigma_1 = 0$, $f = 0$, and

$$
y_0 = \begin{cases} 1 & \text{in } (0, 0.5],\\ 0 & \text{otherwise.}\end{cases}
$$

The grid is given by

$$
x_i = \frac{i}{50} \quad \text{for } i = 0,\dots,50 \quad \text{and} \quad t_j = \frac{jT}{50} \quad \text{for } j = 0,\dots,50.
$$

To solve (2.1) for $u = v = 0$ we apply the Newton method at each time step. The algorithm needs one second CPU time. The value of the cost functional is 0.081.

Now we turn to the optimal control problem. We choose $\alpha = \beta = 0.01$, and the desired state is $z(t,\cdot) = y_0$ for $t \in (0,T)$. In view of the choice of z and the nonlinear convection term yy_x in (2.1b) we can interpret this problem as determining u in such a way that it counteracts the uncontrolled dynamics which smooths the discontinuity at $x = 0.5$ and transports it to the left as t increases. The discretization of (7.2) leads to an indefinite system,

$$
H^k \begin{pmatrix}\delta y\\ \delta\lambda\end{pmatrix} = r^k \quad \text{with} \quad H^k = \begin{pmatrix} A^k & (B^k)^{\mathsf T}\\ B^k & C^k\end{pmatrix}. \tag{7.3}
$$

As starting values for Algorithm 1 we take $y^0 = 0$, $u^0 = v^0 = 0$, and $\lambda^0 = 0$.

(i) First we solve (7.3) by an LU-factorization (MATLAB routine `lu`) so that the theory of section 4 applies. According to section 4 we stop the SQP iteration if

$$
\|F_c(z^{k+1})\|_{z^k} \le 10^{-3} \cdot \|F_c(z^0)\|_{z^0}. \tag{7.4}
$$

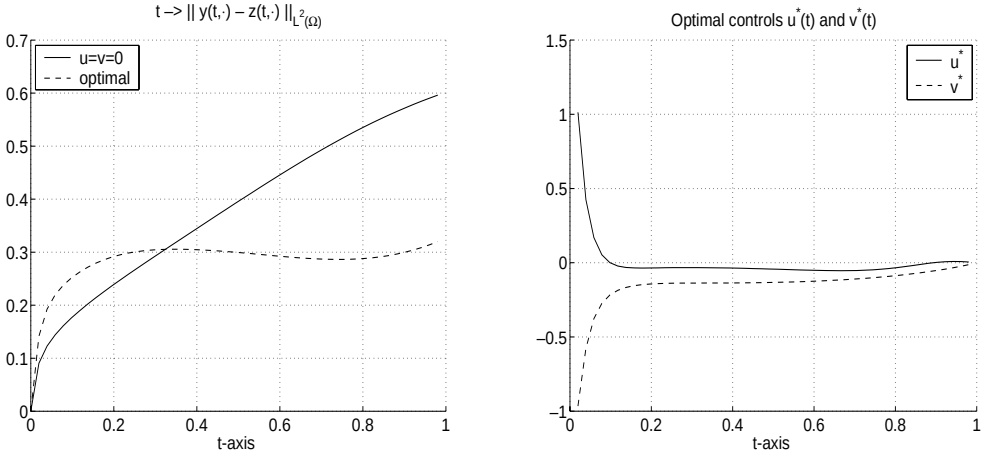


FIG. 7.1. *Run* 7.1*: residuum* $t \mapsto \|y(t,\cdot) - z(t,\cdot)\|_{L^2(\Omega)}$ *and optimal controls.*

In case $\|F_c(z^0)\|_{z^0}$ is very large, the factor 10^{-3} on the right-hand side of (7.4) might be too big. To avoid this situation Algorithm 1 is terminated if (7.4) and, in addition,

$$\|F_c(z^{k+1})\|_{z^k} < 10^{-3}$$

hold. The augmented Lagrangian-SQP method stops after four iterations. The CPU times for different values of c can be found in Tables 7.6 and 7.7. Let us mention that for $c = 0.1$ the algorithm needs 102.7 seconds and for $c = 1$ we observe divergence of Algorithm 1. As it was proved in [15] the set of admissible starting values reduces whenever c enlarges. The value of the cost functional is 0.041. In Figure 7.1 the residuum $t \mapsto \|y(t,\cdot) - z(t,\cdot)\|_{L^2(\Omega)}$ for the solution of (2.1) for $u = v = 0$ as well as for the optimal state is plotted. Furthermore, the optimal controls are presented. The decay of $\|F_c(z^{k+1})\|_{z^k}$, $k = 0, \ldots, 3$, for the first invariant norm given by (5.3) and for different values of c is shown in Table 7.1. Recall that the invariant norm is defined only for $c \geq \bar{c}$. Unfortunately, the constant $\bar{c} \geq 0$ is unknown. We proceed as follows: Choose a fixed value for c and compute

$$[\kappa]_k = \frac{\langle L_c''(x^k, \lambda^k)\delta x, \delta x\rangle_X}{\|\delta x\|_X^2}$$

in each level of the SQP iteration. Whenever $[\kappa]_k$ is greater than zero, we have coercivity in the direction of the SQP step. Otherwise, c needs to be increased. In

TABLE 7.1
Run 7.1(i): *decay of* $\|F_c(z^{k+1})\|_{z^k}$ *for the first norm.*

	$c = 0$	$c = 10^{-3}$	$c = 10^{-2}$
$\|F_c(z^0)\|_{z^0}$	4.636278	4.630344	4.642807
$\|F_c(z^1)\|_{z^0}$	1.635481	1.625800	1.581022
$\|F_c(z^2)\|_{z^1}$	0.210650	0.202490	0.184842
$\|F_c(z^3)\|_{z^2}$	0.003625	0.003234	0.002663
$\|F_c(z^4)\|_{z^3}$	0.000002	0.000001	0.000001

TABLE 7.2
Run 7.1(i): *values of* $[\kappa]_k$ *for different* c.

	$c = 0$	$c = 10^{-3}$	$c = 10^{-2}$
$[\kappa]_0$	0.020	0.020	0.020
$[\kappa]_1$	0.019	0.019	0.020
$[\kappa]_2$	0.004	0.023	0.024
$[\kappa]_3$	0.021	0.022	0.025

TABLE 7.3
Run 7.1(i): *decay of* $\|F_c(z^{k+1})\|_{z^k}$ *for the second norm.*

$\|F_0(z^0)\|_{z^0}$	$\|F_0(z^1)\|_{z^0}$	$\|F_0(z^2)\|_{z^1}$	$\|F_0(z^3)\|_{z^2}$	$\|F_0(z^4)\|_{z^3}$
26.77865	4.91492	0.63812	0.0105	0.00002

Table 7.2 we present the values for $[\kappa]_k$. We observed numerically that $[\kappa]_k$ is positive for $k = 0, \ldots, 3$. Moreover, $[\kappa]_k$ increased if c increased.

Next we tested the second norm introduced in (5.11) for $c = 0$. Again, the augmented Lagrangian-SQP method stops after four iterations and needs 97.4 seconds CPU time. Thus, both invariant norms lead to a similar performance of Algorithm 1. The decay of $\|F_c(z^{k+1})\|_{z^k}$ can be found in Table 7.3.

(ii) Now we solve (7.3) by an inexact generalized minimum residual (GMRES) method (MATLAB routine `gmres`). For a preconditioner for the GMRES method we took an incomplete LU-factorization of the matrix

$$D = \begin{pmatrix} I & P^{\mathsf{T}} \\ P & 0 \end{pmatrix} \tag{7.5}$$

by utilizing the MATLAB function `luinc(`D`,1e-03)`. Here, the matrix P is the discretization of the heat operator $y_t - \nu y_{xx}$ with the homogeneous Robin boundary conditions $\nu y_x(\cdot, 0) + \sigma_0 y(\cdot, 0) = \nu y_x(\cdot, 1) + \sigma_1 y(\cdot, 1) = 0$ in $(0, T)$. The same preconditioner is used for all Newton steps.

We chose $\Theta_k = 0.6$ for all k. In section 6 we introduced estimators for the constants ω and γ, denoted by $[\omega]_k$ and $[\gamma]_k$, respectively. Thus, for $k \geq 0$ we calculate $[\omega]_k$ and $[\gamma]_k$, and then we determine δ_{k+1} as follows:

$\delta_{k+1} = \Theta_k$;
while $(1 + [\gamma]_k \|\nabla F_c(z^k)\delta z^k\|_{z^k})\delta_{k+1} + (1 + \delta_{k+1})\dfrac{[\omega]_k}{2}\|\nabla F_c(z^k)\delta z^k\|_{z^k} > \rho_k \Theta_k$ **do**
  $\delta_{k+1} = \dfrac{\delta_{k+1}}{2}$;
end (while);

where $\rho_k = (1 + \hat{\delta}_k)/(1 + \tilde{\delta}_k)$ for all $k \geq 0$. For the first norm $\|\nabla F_c(z^k)\delta z^k\|_{z^k}$ is already determined by the computation of the previous Newton correction. Thus we have $\tilde{\delta}_k = \delta_k$, but in the case of the second norm, $\|\nabla F_c(z^k)\delta z^k\|_{z^k}$ has to be calculated with a given tolerance $\tilde{\delta}_k$. In our tests we take $\tilde{\delta}_k = \hat{\delta}_k$ for all $k \geq 0$. As starting values we choose $\delta_0 = 10^{-10}$ and $\hat{\delta}_0 = \delta_0$. We test four strategies for the choice of $\hat{\delta}_k$ for $k \geq 1$: $\hat{\delta}_k = 0.1$, $\hat{\delta}_k = 0.01$, $\hat{\delta}_k = 0.001$, and $\hat{\delta}_k = \delta_k$. It turns out that for $\hat{\delta}_k = 0.1$ we obtain the best performance with respect to CPU times. Hence, in the following test examples we take $\delta_k = 0.1$ for $k \geq 1$.

TABLE 7.4
Run 7.1(ii): *decay of* $\|F_c(z^k)\|_{z^k}$ *for the first norm with* $\Theta_k = 0.5$.

	$c = 0$	$c = 10^{-3}$	$c = 10^{-2}$
$\|F_c(z^0)\|_{z^0}$	4.63628	4.70140	5.50174
$\|F_c(z^1)\|_{z^1}$	1.24009	1.27674	1.29441
$\|F_c(z^2)\|_{z^2}$	0.07190	0.20773	0.18324
$\|F_c(z^3)\|_{z^3}$	0.01348	0.01451	0.01102
$\|F_c(z^4)\|_{z^4}$	0.00524	0.00548	0.00739
$\|F_c(z^5)\|_{z^5}$	0.00217	0.00218	0.00004
$\|F_c(z^6)\|_{z^6}$	0.00121	0.00077	—
$\|F_c(z^7)\|_{z^7}$	0.00033	—	—

TABLE 7.5
Run 7.1(ii): *values of* $[\omega]_k$ *for* $\Theta_k = 0.5$.

	$c = 0$	$c = 10^{-3}$	$c = 10^{-2}$
$[\omega]_0$	1.35e–01	1.29e–01	1.18e–01
$[\omega]_1$	2.40e–01	2.20e–01	3.03e–01
$[\omega]_2$	1.16e–01	1.23e–01	6.20e–01
$[\omega]_3$	1.12e–01	1.19e+00	5.27e+01
$[\omega]_4$	7.12e–02	3.81e+00	1.95e+00
$[\omega]_5$	9.91e–02	4.06e+00	—
$[\omega]_6$	8.86e–02	—	—

TABLE 7.6
Run 7.1(ii): *CPU times in seconds for the first norm.*

	$c = 0$	$c = 10^{-3}$	$c = 10^{-2}$
exact	97.5	96.8	96.9
inexact, $\Theta_k = 0.3$	46.8	45.2	47.2
inexact, $\Theta_k = 0.4$	46.0	46.8	45.0
inexact, $\Theta_k = 0.5$	47.8	46.7	44.8
inexact, $\Theta_k = 0.6$	49.3	50.5	47.2
inexact, $\Theta_k = 0.7$	49.2	52.9	48.7
inexact, $\Theta_k = 0.8$	47.2	52.6	46.1
inexact, $\Theta_k = 0.9$	53.0	56.7	46.2
inexact, $\Theta_k = \Theta_{k-1}/2$, $\Theta_0 = 0.9$	42.7	42.6	45.4

The decay of $\|F(z^k)\|_{z^k}$ is presented in Table 7.4. Algorithm 1 stops after at most seven iterations. Let us mention that for $c \in \{0, 10^{-3}, 10^{-2}\}$ the estimates $[\kappa]_k$ for the coercivity constant are positive. In particular, for $c = 10^{-2}$ the augmented Lagrangian-SQP method has the best performance. In Table 7.5 the values of the estimators $[\omega]_k$ are presented. In Table 7.6 the CPU times for the first norm are presented. It turns out that the performance of the inexact method does not change significantly for different values of Θ_k. Since we have to solve an additional linear system at each level of the SQP iteration in order to compute the second norm, the first norm leads to a better performance of the inexact method with respect to the CPU time. Compared to part (i) the CPU time is reduced by about 50% if one takes the first norm. In the case of the second norm the reduction is about 45% for $\Theta_k \in \{0.3, 0.4, 0.5, 0.6, 0.6\}$; see Table 7.7. Finally we test the inexact method using

TABLE 7.7
Run 7.1(ii): *CPU times in seconds for both norms and* $c = 0$.

	First norm	Second norm
exact	97.5	97.4
inexact, $\Theta_k = 0.3$	46.8	54.5
inexact, $\Theta_k = 0.4$	46.0	54.0
inexact, $\Theta_k = 0.5$	47.8	53.9
inexact, $\Theta_k = 0.6$	49.3	53.9
inexact, $\Theta_k = 0.7$	49.2	59.6
inexact, $\Theta_k = 0.8$	47.2	74.4
inexact, $\Theta_k = 0.9$	53.0	77.0
inexact, $\Theta_k = \Theta_{k-1}/2$, $\Theta_0 = 0.9$	42.7	51.7

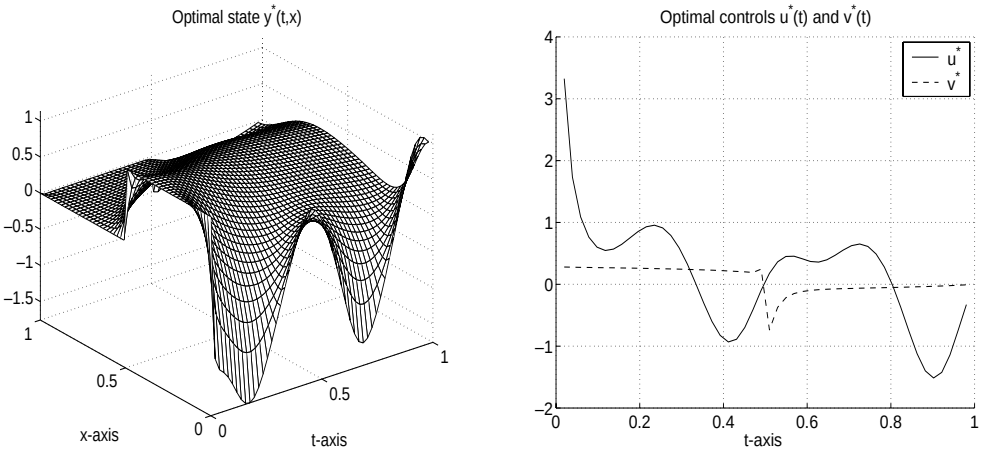


FIG. 7.2. *Run* 7.2*: optimal state and controls.*

decreasing Θ_k. We choose $\Theta_0 = 0.9$ and $\Theta_k = \Theta_{k-1}/2$ for $k \geq 1$. It turns out that this strategy speeds up the inexact method for both norms, as can be expected from the theoretical complexity model developed in [7].

Run 7.2 (Robin control). We choose $T = 1$, $\nu = 0.05$, $\sigma_0(t) = \sin(4\pi t)$, $f = 0$, $\alpha = \beta = 0.01$,

$$\sigma_1 = \begin{cases} -10 & \text{in } \left(0, \frac{T}{2}\right), \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad y_0 = \begin{cases} 1 & \text{in } \left(0, \frac{1}{2}\right), \\ 0 & \text{otherwise.} \end{cases}$$

The desired state was taken to be $z(t, \cdot) = y_0 \cos(4\pi t)$ for $t \in [0, T]$.

(i) First we again solve (7.3) by an LU-factorization. We take the same starting values and stopping criteria as in Run 7.1. The augmented Lagrangian-SQP method stops after four iteration and needs 105 seconds CPU time. The discrete optimal solution is plotted in Figure 7.2. From Table 7.8 it follows that (4.7) is satisfied numerically. Let us mention that $[\kappa]_0, \ldots, [\kappa]_3$ are positive for $c \in \{0, 10^{-3}, 10^{-2}\}$. For the needed CPU times we refer to Tables 7.10 and 7.11.

(ii) Now we solve (7.3) by an inexact GMRES method. For a preconditioner we take the same as in Run 7.1. We choose $\Theta_k = 0.5$ for all k. The decay of $\|F(z^k)\|_{z^k}$ is presented in Table 7.9. As in part (i) we find that $[\kappa]_k > 0$ for all test runs. The needed

TABLE 7.8
Run 7.2(i): *decay of* $\|F_c(z^{k+1})\|_{z^k}$ *for different* c.

	$c = 0$	$c = 10^{-3}$	$c = 10^{-2}$
$\|F_c(z^0)\|_{z^0}$	3.11799	3.12494	3.15978
$\|F_c(z^1)\|_{z^0}$	1.25420	1.29953	1.75698
$\|F_c(z^2)\|_{z^1}$	0.18289	0.18768	0.26507
$\|F_c(z^3)\|_{z^2}$	0.01361	0.00849	0.01200
$\|F_c(z^4)\|_{z^3}$	0.00009	0.00003	0.00006

TABLE 7.9
Run 7.2(ii): *decay of* $\|F_c(z^k)\|_{z^k}$ *for* $\Theta_k = 0.5$.

	$c = 0$	$c = 10^{-3}$	$c = 10^{-2}$
$\|F_c(z^0)\|_{z^0}$	3.117994	3.296494	4.457908
$\|F_c(z^1)\|_{z^1}$	1.171467	1.285089	2.181491
$\|F_c(z^2)\|_{z^2}$	0.187818	0.199968	0.330048
$\|F_c(z^3)\|_{z^3}$	0.012231	0.024343	0.033398
$\|F_c(z^4)\|_{z^4}$	0.013299	0.003203	0.001044
$\|F_c(z^5)\|_{z^5}$	0.001202	0.002336	0.000132
$\|F_c(z^6)\|_{z^6}$	0.000390	0.000441	—

TABLE 7.10
Run 7.2(ii): *CPU times in seconds for the first norm.*

	$c = 0$	$c = 10^{-3}$	$c = 10^{-2}$
exact	105.1	105.7	105.7
inexact, $\Theta_k = 0.3$	44.6	43.2	48.5
inexact, $\Theta_k = 0.4$	43.6	49.9	48.5
inexact, $\Theta_k = 0.5$	43.0	43.8	50.5
inexact, $\Theta_k = 0.6$	48.1	45.3	45.5
inexact, $\Theta_k = 0.7$	44.2	45.3	45.2
inexact, $\Theta_k = 0.8$	44.5	47.2	45.0
inexact, $\Theta_k = 0.9$	44.5	45.0	45.2
inexact, $\Theta_k = \Theta_{k-1}/2$, $\Theta_0 = 0.9$	40.3	40.3	48.2

TABLE 7.11
Run 7.2(ii): *CPU times in seconds for both norms and* $c = 0$.

	First norm	Second norm
exact	105.1	105.5
inexact, $\Theta_k = 0.3$	44.6	50.7
inexact, $\Theta_k = 0.4$	43.6	53.0
inexact, $\Theta_k = 0.5$	43.0	53.0
inexact, $\Theta_k = 0.6$	48.1	55.3
inexact, $\Theta_k = 0.7$	44.2	55.2
inexact, $\Theta_k = 0.8$	44.5	65.7
inexact, $\Theta_k = 0.9$	44.5	65.9
inexact, $\Theta_k = \Theta_{k-1}/2$, $\Theta_0 = 0.9$	40.3	48.0

CPU times are shown in Tables 7.10 and 7.11. As we can see, the inexact augmented Lagrangian-SQP method with GMRES is much faster than the exact one using the *LU*-factorization. For the first norm the CPU time is reduced by about 55%, and for the second norm by about 50% for $\Theta_k \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$. Moreover, for our example the best choice for c is $c = 10^{-3}$. For smaller values of Θ_k the method does not speed up significantly. As in Run 7.1 we test the inexact method using decreasing Θ_k. Again we choose $\Theta_0 = 0.9$ and $\Theta_k = \Theta_{k-1}/2$ for $k \geq 1$. As in Run 7.1, this strategy speeds up the inexact method significantly for both norms. The reduction is about 9% compared to the CPU times for fixed Θ_k; see Table 7.11.

REFERENCES

[1] R. E. Bank and D. J. Rose, *Global approximate newton methods*, Numer. Math., 37 (1981), pp. 279–295.

[2] D. Braess, P. Deuflhard, and K. Lipnikov, *A subspace cascadic multigrid method for mortar elements*, Numer. Math., to appear.

[3] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology, Vol.* 5: *Evolution Problems* I, Springer, Berlin, 1992.

[4] R. S. Dembo, S. C. Eisenstat, and T. Steihaug, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.

[5] P. Deuflhard, *Newton Methods for Nonlinear Problems. Affine Invariance and Adaptive Algorithms*, Springer, to appear.

[6] P. Deuflhard and G. Heindl, *Affine invariant convergence theorems for Newton's method and extensions to related methods*, SIAM J. Numer. Anal., 16 (1979), pp. 1–10.

[7] P. Deuflhard and M. Weiser, *Local inexact Newton multilevel FEM for nonlinear elliptic problems*, in Computational Science for the 21st Century, M.-O. Bristeau, G. Etgen, W. Fitzigibbon, J.-L. Lions, J. Périaux, and M. Wheeler, eds., Wiley, New York, 1997, pp. 129–138.

[8] P. Deuflhard and M. Weiser, *Global inexact Newton multilevel FEM for nonlinear elliptic problems*, in Multigrid Methods V, Lecture Notes in Comput. Sci. and Engrg. 3, W. Hackbusch and G. Wittum, eds., Springer, Berlin, 1998, pp. 71–89.

[9] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms*, Springer, Berlin, 1980.

[10] A. Hohmann, *Inexact Gauss Newton Methods for Parameter Dependent Nonlinear Problems*, Ph.D. thesis, Dept. Math. and Comput. Sci., Free University of Berlin, Germany, 1994.

[11] K. Ito and K. Kunisch, *Augmented Lagrangian-SQP-methods in Hilbert spaces and application to control in the coefficients problems*, SIAM J. Optim., 6 (1996), pp. 96–125.

[12] D. G. Luenberger, *Optimization by Vector Space Methods*, Wiley, New York, 1969.

[13] H. Maurer and J. Zowe, *First and second-order necessary and sufficient optimality conditions for infinite-dimensional programming problems*, Math. Programming, 16 (1979), pp. 98–110.

[14] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[15] S. Volkwein, *Mesh-independence for an augmented Lagrangian-SQP method in Hilbert spaces*, SIAM J. Control Optim., 38 (2000), pp. 767–785.

[16] S. Volkwein, *Optimal and Suboptimal Control of Partial Differential Equations: Augmented Lagrange-SQP Methods and Reduced-Order Modeling with Proper Orthogonal Decomposition*, Grazer Math. Ber. 343, Karl-Franzens-Universität Graz, Graz, Austria, 2001.

[17] S. Volkwein, *Mesh-independence of Lagrange-SQP methods with Lipschitz-continuous Lagrange multiplier updates*, Optim. Methods Softw., 17 (2002), pp. 77–111.

[18] S. Volkwein, *Second-order conditions for boundary control problems of the Burgers equation*, Control Cybernet., 30 (2001), pp. 249–278.

[19] E. Zeidler, *Nonlinear Functional Analysis and Its Applications*, Vol. I, Springer, New York, 1986.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 3, pp. 900–931

ROBUST OPTIMAL SWITCHING CONTROL FOR NONLINEAR SYSTEMS*

JOSEPH A. BALL†, JERAWAN CHUDOUNG‡, AND MARTIN V. DAY†

Abstract. We formulate a robust optimal control problem for a general nonlinear system with finitely many admissible control settings and with costs assigned to switching of controls. We formulate the problem both in an L_2-gain/dissipative system framework and in a game-theoretic framework. We show that, under appropriate assumptions, a continuous switching-storage function is characterized as a viscosity supersolution of the appropriate system of quasi-variational inequalities (the appropriate generalization of the Hamilton–Jacobi–Bellman–Isaacs equation for this context) and that the minimal such switching-storage function is equal to the continuous switching lower-value function for the game. Finally, we show how a prototypical example with one-dimensional state space can be solved by a direct geometric construction.

Key words. running cost, switching cost, worst-case disturbance attenuation, differential game, state-feedback control, nonanticipating strategy, storage function, lower-value function, system of quasi-variational inequalities, viscosity solution

AMS subject classifications. Primary, 49J35; Secondary, 49L20, 49L25, 49J35, 93B36, 93B52

PII. S0363012900372611

1. Introduction.

We consider a state-space system Σ_{sw}

$$\dot{y} = f(y, a, b), \tag{1.1}$$

$$z = h(y, a, b), \tag{1.2}$$

where $y(t) \in \mathbb{R}^N$ is the state, $a(t) \in A$ is the control input, $b(t) \in B \subset \mathbb{R}^M$ is the deterministic unknown disturbance, and $z(t) \in \mathbb{R}$ is the cost function. We assume that the set A of admissible control values is a finite set, $A = \{a^1, \dots, a^r\}$. The control signals $a(t)$ are then necessarily piecewise constant with values in A. We normalize control signals $a(t)$ to be right continuous and refer to the value $a(t)$ as the *new current control* and $a(t^-)$ as the *old current control* at time t. We assume that there is a distinguished input index i_0 for which $f(0, a^{i_0}, 0) = 0$ and $h(0, a^{i_0}, 0) = 0$ so that 0 is an equilibrium point for the autonomous system induced by setting $a(t) = a^{i_0}$ and $b(t) = 0$. In addition, we assume that a cost $k(a^i, a^j) \geq 0$ is assigned at each time instant τ_n at which the controller switches from the old current control $a(\tau_n^-) = a^i$ to the new current control $a(\tau_n) = a^j$. For a given old initial control $a(0^-)$, the associated control decision is to choose switching times

$$0 \leq \tau_1 < \tau_2 < \cdots, \qquad \lim_{n\to\infty} \tau_n = \infty,$$

and controls

$$a(\tau_1),\ a(\tau_2),\ a(\tau_3), \dots$$

*Received by the editors May 26, 2000; accepted for publication (in revised form) February 21, 2002; published electronically September 19, 2002.
http://www.siam.org/journals/sicon/41-3/37261.html

†Department of Mathematics, Virginia Tech, Blacksburg, VA 24061 (ball@math.vt.edu, day@math.vt.edu).

‡Department of General Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801 (chudoung@uiuc.edu).

such that the controller switches from the old current control $a(\tau_n^-)$ to the new current control $a(\tau_n) \neq a(\tau_n^-)$ at time τ_n, where we set

$$a(t) = \begin{cases} a(0^-), & t \in [0, \tau_1), \\ a(\tau_n), & t \in [\tau_n, \tau_{n+1}),\ n = 1, 2, \dots, \end{cases}$$

if $\tau_1 > 0$ and

$$a(t) = a(\tau_n), \qquad t \in [\tau_n, \tau_{n+1}),\ n = 1, 2, \dots,$$

otherwise. We assume that the state $y(\cdot)$ of (1.1) does not jump at the switching time τ_n; i.e., the solution $y(\cdot)$ is assumed to be absolutely continuous. The cost of running the system up to time $T \geq 0$ with initial state $y(0) = x$, old initial control $a(0^-) = a^j$, control signal a for $t \geq 0$, and disturbance signal b is given by

$$C_{T^-}(x, a^j, a, b) = \int_0^T h(y_x(t, a, b), a(t), b(t))\ dt + \sum_{\tau:\ 0 \leq \tau < T} k(a(\tau^-), a(\tau)).$$

We have used the notation $y_x(\cdot, a, b)$ for the unique solution of (1.1) corresponding to the choices of the initial condition $y(0) = x$, the control $a(\cdot)$, and the disturbance $b(\cdot)$. In what follows, we will often abbreviate $y_x(\cdot, a, b)$ to $y_x(\cdot)$ or $y(\cdot)$; the precise meaning should be clear from the context.

As the running cost $h(y(t), a(t), b(t)) + k(a(t^-), a(t))$, where $a(t^-) = a^j$ if $t = 0$, involves not only the value $y(t)$ of the state along with the value of the control $a(t)$ and the value of the disturbance $b(t)$ at time t but also the value of the old current control $a(t^-)$, it makes sense to think of the old current control $a(t^-)$ at time t as part of an augmented state vector $y^{aug}(t) = (y(t), a(t^-))$ at time t. This can be done formally by including $a(t^-)$ as part of the state vector, in which case a switching control problem becomes an *impulse* control problem (see [10], where problems of this sort are set in the general framework of hybrid systems). We shall keep the switching-control formalism here; however, in implementing optimization algorithms, we shall see that it is natural to consider augmented state-feedback controls $(x, a^j) \to a(x, a^j)$ rather than merely state-feedback controls $x \to a(x)$ in order to obtain solutions. We shall refer to such augmented state-feedback controls $(x, a^j) \to a(x, a^j) \in A$ as simply *switching state-feedback* controllers. Note that, while the augmented state is required to compute the instantaneous running cost at time t, only the (nonaugmented) state vector $y(t)$ at time t is needed to determine the state trajectory past time t for a given input signal $(a(\cdot), b(\cdot))$ past time t.

The precise formulation of our optimal control problem is as follows. First, for a prescribed attenuation level $\gamma > 0$ and a given augmented initial state (x, a^j), we seek an admissible control signal $a(\cdot) = a_{x,j}(\cdot)$ with $a(0^-) = a^j$ so that

$$C_{T^-}(x, a^j, a, b) \leq \gamma^2 \int_0^T |b(t)|^2\ dt + U_\gamma^j(x) \tag{1.3}$$

for all locally L_2 disturbances b, all positive real numbers T, and some nonnegative-valued bias function $U_\gamma^j(x)$ with $U_\gamma^{i_0}(0) = 0$. Note that this inequality corresponds to an input-output system having L_2-gain at most γ, where C_{T^-} replaces the L_2-norm of the output signal over the time interval $[0, T]$, and where the equilibrium point is taken to be $(0, a^{i_0})$ in the augmented state space. The dissipation inequality (1.3) then can be viewed as an L_2-gain inequality, and our problem can be viewed as

the analogue of the nonlinear H^∞-control problem for systems with switching costs (see [20]). In the *switching state-feedback* version of the problem, $a(\cdot)$ is a function of the current state and the current old control; i.e., one decides what control to use at time t based on knowledge of the current augmented state $(y(t), a(t^-))$. In the standard game-theoretic formulation of the problem, $a(\cdot)$ is a nonanticipating function $a(\cdot) = \alpha_x^j[b](\cdot)$ (called a *strategy*) of the disturbance b depending also on the initial state x and initial old control value a^j; i.e., for a given augmented initial state (x, a^j), the computation of the control value $\alpha_x^j[b](t)$ at time t uses knowledge only of the past and current values of the disturbance $b(\cdot)$. Second, we ask for the admissible control a with $a(0^-) = a^j$ (with whatever information structure) which gives the best system performance in the sense that the nonnegative functions $U_\gamma^j(x)$ are as small as possible. A closely related problem formulation is to view the switching-control system as a game with payoff function

$$J_{T^-}(x, a^j, a, b) = \int_{[0,T)} l(y_x(t), a^j, a(t), b(t)), \qquad a(0^-) = a^j,\ j = 1, \dots, r,$$

where we view $l(y_x, a^j, a, b)$ as the measure given by

$$\begin{aligned} l(y(t), a^j, a(t), b(t)) &= [h(y(t), a(t), b(t)) - \gamma^2 |b(t)|^2]\ dt + k(a(t^-), a(t))\delta_t, \\ &\text{with } a(0^-) = a^j, \end{aligned}$$

where δ_t is the unit point-mass distribution at the point t. In this game setting, the disturbance player seeks to use $b(t)$ and T to maximize the payoff, while the control player seeks to use the choice of piecewise-constant right-continuous function $a(t)$ to minimize the payoff. The *switching lower value* $V_\gamma = (V_\gamma^1, \dots, V_\gamma^r)$ of this game is then given by

$$V_\gamma^j(x) = \inf_\alpha \sup_{b,\, T} J_{T^-}(x, a^j, \alpha_x^j[b], b), \quad j = 1, \dots, r, \tag{1.4}$$

where the supremum is over all nonnegative real numbers T and all locally L_2-disturbance signals b, while the infimum is over all nonanticipating control strategies $b \to \alpha_x^j[b]$ depending on the initial augmented state (x, a^j). By letting T tend to 0, we see that each component of the switching lower value $V_\gamma(x) = (V_\gamma^1(x), \dots, V_\gamma^r(x))$ is nonnegative. Then, by construction, $(V_\gamma^1, \dots, V_\gamma^r)$ gives the smallest possible value which can satisfy (1.3) (with V_γ^j in place of U_γ^j) for some nonanticipating strategy $(x, a^j, b) \to \alpha_x^j[b](\cdot) = a(\cdot)$.

In the standard theory of nonlinear H^∞-control, the notion of *storage function* for a dissipative system plays a prominent role (see [20]). For our setting with switching costs, we say that a nonnegative vector function $S_\gamma = (S_\gamma^1, \dots, S_\gamma^r)$ on $\mathbb{R}^N$ is a *switching-storage function* for the system (1.1)–(1.2) with strategy α if, for all $y(0) = x \in \mathbb{R}^N$, b measurable with values in B and $0 \le t_1 < t_2$, the following inequality holds:

$$\begin{aligned} S_\gamma^{j(t_2)}(y_x(t_2, \alpha_x^j[b], b)) - S_\gamma^{j(t_1)}(y_x(t_1, \alpha_x^j[b], b)) \\ \le \int_{t_1}^{t_2} [\gamma^2 |b(s)|^2 - h(y_x(s), \alpha_x^j[b](s), b(s))]\ ds \\ - \sum_{t_1 \le \tau < t_2} k(\alpha_x^j[b](\tau^-), \alpha_x^j[b](\tau)) \end{aligned} \tag{1.5}$$

(where $j(t)$ is specified by $\alpha_x^j[b](t^-) = a^{j(t)}$). The control problem then is to find the switching strategy $\alpha : (x, a^j, b) \to \alpha_x^j[b](\cdot)$ which gives the best performance, as measured by obtaining the minimal possible $S_\gamma(x) = (S_\gamma^1(x), \dots, S_\gamma^r(x))$ as the associated closed-loop switching-storage function. Note that any switching-storage function may serve as the vector bias function $U_\gamma = (U_\gamma^1, \dots, U_\gamma^r)$ in the L_2-gain inequality (1.3) if, in addition, $S_\gamma^{i_0}(0) = 0$. This suggests that the *available switching-storage function* (i.e., the minimal possible switching-storage function over all possible switching strategies) should equal the switching lower-value V_γ (1.4) for the game described above. We shall see that this is indeed the case with appropriate hypotheses imposed.

Our main results concerning the robust optimal switching-cost problem are as follows: *Under minimal smoothness assumptions on the problem data and compactness of the set B, the following hold:*

(i) $V_\gamma^j(x) \le \min_{i \ne j}\{V_\gamma^i(x) + k(a^j, a^i)\},\ x \in \mathbb{R}^N,\ j = 1, \dots, r.$

(ii) *If continuous, V_γ is a viscosity solution in $\mathbb{R}^N$ of the system of quasi-variational inequalities (SQVI) defined in section* 2 *(see* (2.5)*).* (The precise definition of viscosity subsolution, supersolution, and solution will be given in section 2.)

(iii) *If $S_\gamma = (S_\gamma^1, \dots, S_\gamma^r)$ is a continuous switching-storage function for some strategy α, then S_γ is a nonnegative continuous viscosity supersolution of the SQVI* (2.5).

(iv) *If $U_\gamma = (U_\gamma^1, \dots, U_\gamma^r)$ is a nonnegative, continuous viscosity supersolution of the SQVI* (2.5) *and U_γ has the property* (i)*, then there is a canonical choice of switching state-feedback control strategy $\alpha_{U_\gamma} : (x, a^j, b) \to \alpha_{U_\gamma, x}^j[b]$ such that U_γ is a switching-storage function for the closed-loop system formed by using the strategy α_{U_γ}; thus*

$$U_\gamma^j(x) \ge \sup_{b,\ T}\left\{\int_{[0,T)} l(y_x(s), a^j, \alpha_{U_\gamma, x}^j[b](s), b(s))\right\} \ge V_\gamma^j(x).$$

The switching lower-value V_γ, if continuous, is characterized as the minimal, nonnegative continuous viscosity supersolution of (2.5) *having property* (i) *above as well as the minimal continuous function satisfying property* (i)*, which is a switching-storage function for the closed-loop system associated with some nonanticipating strategy α.*

In the precise formulation of our problem, for technical convenience, we impose the condition that the disturbance signals $b(t)$ take values in a bounded subset B of $\mathbb{R}^M$; hence our setup technically does not include the linear-quadratic case (where f is linear and h is quadratic). In general, this issue has been a stumbling block for application of the nonlinear dynamic programming formalism to this class of problems. In [21], this difficulty was overcome by an ad hoc reparametrization technique, whereby the general unbounded case was reduced to the bounded case. This would be one approach to removing the boundedness assumption which we have imposed here; however, see also Remark 1 in section 3 below.

The usual formulation of the H^∞-control problem also involves a stability constraint. We also prove that, under appropriate conditions, the closed-loop system associated with switching strategy α_{U_γ} corresponding to the nonnegative continuous supersolution U_γ of the SQVI is stable. The main idea is to use the supersolution U_γ as a Lyapunov function for trajectories of the closed-loop system. Related stability problems for systems with control switching are discussed, e.g., by Branicky in [11].

Infinite-horizon optimal switching-control problems are discussed in [6, Chapter III, section 4.4] but with a discount factor in the running cost and no disturbance

term. Differential games with switching strategies and switching costs for the case of finite horizon problems is discussed in [23], while the case of an infinite horizon with both control and competing disturbance but with a discount factor in the running cost is discussed in [24]. These authors, under their various assumptions, were able to show that the value function is continuous and is the unique solution of the appropriate system of quasi-variational inequalities. However, our formulation has no discount factor in the running cost, so the running cost is not guaranteed to be integrable over the infinite interval $[0, \infty)$. This forces the introduction of the extra "disturbance player" T in (1.4). We establish a dynamic programming principle (DPP) for this setting and derive from it the appropriate system of quasi-variational inequalities (SQVI) to be satisfied by V_γ. While elements of our derivation of the DPP closely follow the known proofs for other cases (see [23], [24]), these proofs do not carry over directly due to a lack of positive discount factor and the presence of the extra disturbance player T. Our lower-value function V_γ probably in general is not continuous and, moreover, cannot be characterized simply as the unique solution of the SQVI as is the case for finite-horizon problems and problems with a positive discount factor. Our formulation of the optimal switching-cost problem is a precise analogue of the standard nonlinear H^∞-control problem; our results (particularly the characterization of the switching lower value as the minimal viscosity supersolution of the appropriate SQVI) parallel those of Soravia [21] obtained for the standard nonlinear H^∞-control problem (see also [13], [22], and [6, Appendix B] for later, closely related refinements of the nonlinear H_∞ results).

Another approach to the derivation of the Hamilton–Jacobi–Bellman–Isaacs (HJBI) equation satisfied by the value function for a differential game is as an application of a comparison principle for the HJBI equation (see [12] or [6]). In [1], this approach was adapted to provide an alternative derivation of the SQVI satisfied by the lower-value function for the robust switching-control problem studied here.

In our companion paper [2], we present a parallel analysis for another analogue of the nonlinear H^∞-control problem, namely, a *robust stopping-time control problem*, where the only control is a decision as to when to stop the system, and there is an instantaneous cost for stopping (dependent on the final state) in addition to the running cost. In this setting, the storage function (or value function if one uses the game interpretation) is a solution of a single variational inequality rather than a coupled system of quasi-variational inequalities as is the case here. The results and general techniques from [2] parallel those of the present paper, but specific details necessarily differ due to the differences in settings. A connection between the two problems is explained in Remark 2 in section 3.

More general types of impulse-control problems have been studied in the literature (see, e.g., [7], [17], [18]) where a general (not necessarily discrete) measure is allowed to enter both the dynamics and the running cost. Such generality leads to a number of complications, such as what is meant by a trajectory of the closed-loop system, how to implement the DPP for discontinuous Hamiltonians, etc. Again, these authors' formulations focus on a finite horizon or assume a discount factor in the running cost. Our purpose here is to work out the details for the switching-control analogue of the standard nonlinear H^∞-control problem, where there is an infinite horizon with no discount factor in the running cost for the simpler situation where the singularities in the control are simple jumps.

Original motivation for our work arose from the problem of designing a real-time feedback control for traffic signals at a highway intersection (see [3], [4]), where the

size of the cost imposed on switching can be used as a tuning parameter to lead to more desirable types of traffic-light signalization. Also, a positive switching cost eliminates the chattering present in the solution otherwise.

The paper is organized as follows. In section 2, we discuss assumptions and definitions. Section 3 presents the main results on the connection between value functions (and storage functions) with systems of quasi-variational inequalities. Section 4 presents stability of the closed-loop switching control system. Finally, section 5 presents an example with one-dimensional state-space, where the value function and associated robust state-feedback control are explicitly computable; a similar example for the setting of the robust stopping-time problem is presented in [1].

2. Preliminaries. Let $A = \{a^1, a^2, \dots, a^r\}$ be a finite set, and let B be a compact subset of $\mathbb{R}^M$ containing the origin 0. We consider a general nonlinear system Σ_{sw} (see (1.1)–(1.2)) with a switching-cost function k. We make the following assumptions on problem data $f,\ h,\ k$:

(A1) $f : \mathbb{R}^N \times A \times B \to \mathbb{R}^N$ and $h : \mathbb{R}^N \times A \times B \to \mathbb{R}$ are continuous;

(A2) f and h are bounded on $B(0,R) \times A \times B$ for all $R > 0$;

(A3) there are moduli ω_f and ω_h such that

$$|f(x,a,b) - f(y,a,b)| \le \omega_f(|x-y|, R),$$
$$|h(x,a,b) - h(y,a,b)| \le \omega_h(|x-y|, R)$$

for all $x, y \in B(0,R)$, $R > 0$, $a \in A$, and $b \in B$;

(A4) $(f(x,a,b) - f(y,a,b)) \cdot (x-y) \le L|x-y|^2$ for all $x, y \in \mathbb{R}^N,\ a \in A$, and $b \in B$;

(A5) $k : A \times A \to \mathbb{R}$ and

$$k(a^j, a^i) < k(a^j, a^d) + k(a^d, a^i),$$
$$k(a^j, a^i) > 0,$$
$$k(a^j, a^j) = 0$$

for all $a^d,\ a^i,\ a^j \in A,\ d \ne i \ne j$;

(A6) $h(x,a,0) \ge 0$ for all $x \in \mathbb{R}^N,\ a \in A$.

The set of admissible controls for our problem is the set

$$\mathcal{A} = \left\{ a(\cdot) = \sum_{i \ge 1} a_{i-1} 1_{[\tau_{i-1}, \tau_i)}(\cdot) \mid a_i \in A,\ a_i \ne a_{i-1} \text{ for } i \ge 1, \right.$$
$$\left. 0 = \tau_0 \le \tau_1 < \tau_2 < \cdots,\ \tau_i \uparrow \infty \right\}$$

consisting of piecewise-constant right-continuous functions on $[0, \infty)$ with values in the control set A, where we denote by $\tau_1, \tau_2, \dots$ the points at which control switchings occur. The set of admissible disturbances is $\mathcal{B}$, which consists of measurable functions on $[0, \infty)$ with values in the set B:

$$\mathcal{B} = \{b : [0, \infty) \to B \mid b \text{ is measurable on } [0, \infty)\}.$$

Note that any admissible disturbance b is then locally integrable by the assumption that the disturbance set B is bounded. A *strategy* is a map $\alpha \colon \mathbb{R}^N \times A \times \mathcal{B} \to \mathcal{A}$

with value at (x, a^j, b) denoted by $\alpha_x^j[b](\cdot)$. The strategy α assigns control function $a(t) = \alpha_x^j[b](t)$ if the augmented initial condition is (x, a^j) and the disturbance is $b(\cdot)$. Thus, if it happens that $\tau_1 > \tau_0 = 0$, then $a(t) = a_0 = a^j$ for $t \in [\tau_0, \tau_1)$. Otherwise, $a(t) = a_1 \neq a^j$ for $t \in [0, \tau_2) = [\tau_1, \tau_2)$, and an instantaneous charge of $k(a^j, a(0))$ is incurred at time 0 in the cost function. A strategy α is said to be *nonanticipating* if, for each $x \in \mathbb{R}^N$ and $j \in \{1, \dots, r\}$, for any $T > 0$ and $b,\ \bar{b} \in \mathcal{B}$, with $b(s) = \bar{b}(s)$ for all $s \leq T$, it follows that $\alpha_x^j[b](s) = \alpha_x^j[\bar{b}](s)$ for all $s \leq T$. We denote by Γ the set of all nonanticipating strategies:

$$\Gamma = \{\alpha \colon \mathbb{R}^N \times A \times \mathcal{B} \to \mathcal{A} \mid \alpha_x^j \text{ is nonanticipating for each } x \in \mathbb{R}^N \text{ and } j = 1, \dots, r\}.$$

We consider trajectories of the nonlinear system

$$\begin{cases} \dot{y}(t) = f(y(t), a(t), b(t)), \\ y(0) = x. \end{cases} \tag{2.1}$$

Under the assumptions (A1), (A2), and (A4), for given $x \in \mathbb{R}^N,\ a \in \mathcal{A}$, and $b \in \mathcal{B}$, the solution of (2.1) exists uniquely for all $t \geq 0$. We denote by $y_x(\cdot, a, b)$ or simply $y_x(\cdot)$ the unique solution of (2.1) corresponding to the choice of the initial condition $x \in \mathbb{R}^N$, the control $a(\cdot) \in \mathcal{A}$, and the disturbance $b(\cdot) \in \mathcal{B}$. We also have the usual estimates on the trajectories (see, e.g., [6, pp. 97–99]:

$$|y_x(t, a, b) - y_z(t, a, b)| \leq e^{Lt}|x - z|, \quad t > 0, \tag{2.2}$$

$$|y_x(t, a, b) - x| \leq M_x t, \quad t \in [0, 1/M_x], \tag{2.3}$$

$$|y_x(t, a, b)| \leq (|x| + \sqrt{2Kt})e^{Kt} \tag{2.4}$$

for all $a \in \mathcal{A},\ b \in \mathcal{B}$, where

$$M_x = \max\{|f(z, a, b)| \mid |x - z| \leq 1,\ a \in A,\ b \in B\},$$
$$K = L + \max\{|f(0, a, b)| \mid a \in A,\ b \in B\}.$$

For a specified gain tolerance $\gamma > 0$, we define the Hamiltonian function $H^j : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ by setting

$$H^j(y, p) = \min_{b \in B}\{-p \cdot f(y, a^j, b) - h(y, a^j, b) + \gamma^2|b|^2\}, \quad j = 1, \dots, r.$$

Note that $H^j(y, p) < +\infty$ for all $y, p \in \mathbb{R}^N$ by (A2). Under assumptions (A1)–(A4), one can show that the Hamiltonian H^j is continuous on $\mathbb{R}^N \times \mathbb{R}^N$ and satisfies

$$|H^j(x, p) - H^j(y, p)| \leq L|x - y||p| + \omega_h(|x - y|, R)$$
$$\text{for all } p \in \mathbb{R}^N,\ x, y \in B(0, R),\ R > 0, \text{ and}$$
$$|H^j(x, p) - H^j(x, q)| \leq L(|x| + 1)|p - q| \text{ for all } x, p, q \in \mathbb{R}^N.$$

We now introduce the system of quasi-variational inequalities (SQVI): for $j = 1, 2, \dots, r$,

$$\max\left\{H^j(x, Du^j(x)),\ u^j(x) - \min_{i \neq j}\{u^i(x) + k(a^j, a^i)\}\right\} = 0, \quad x \in \mathbb{R}^N. \tag{2.5}$$

DEFINITION 1. *A vector function $u = (u^1, u^2, \dots, u^r)$, where $u^j \in C(\mathbb{R}^N)$, is a viscosity subsolution of the SQVI* (2.5) *if, for any $\varphi^j \in C^1(\mathbb{R}^N)$,*

$$\max\left\{H^j(x_0, D\varphi^j(x_0)),\ u^j(x_0) - \min_{i\neq j}\{u^i(x_0) + k(a^j, a^i)\}\right\} \le 0, \quad j = 1, 2, \dots, r,$$

at any local maximum point $x_0 \in \mathbb{R}^N$ of $u^j - \varphi^j$. Similarly, u is a viscosity supersolution of the SQVI (2.5) *if, for any $\varphi^j \in C^1(\mathbb{R}^N)$,*

$$\max\left\{H^j(x_1, D\varphi^j(x_1)),\ u^j(x_1) - \min_{i\neq j}\{u^i(x_1) + k(a^j, a^i)\}\right\} \ge 0, \quad j = 1, 2, \dots, r,$$

at any local minimum point $x_1 \in \mathbb{R}^N$ of $u^j - \varphi^j$. Finally, u is a viscosity solution of the SQVI (2.5) *if it is simultaneously a viscosity sub- and supersolution.*

3. Main results. In this section, we show the connection of the lower value function $V_\gamma = (V_\gamma^1, \dots, V_\gamma^r)$ (see (1.4)) (and a switching-storage function) with the SQVI (2.5).

We begin with the application of the DPP to this setting and then derive some properties of the lower-value vector function V_γ (see (1.4)). We then use these properties to show that V_γ, if continuous, is a viscosity solution of the SQVI (2.5). Throughout this section, we assume that V_γ is finite.

PROPOSITION 3.1. *Assume* (A1)–(A5). *Then, for $j = 1, 2, \dots, r$ and $x \in \mathbb{R}^N$, the lower-value vector function $V_\gamma = (V_\gamma^1, \dots, V_\gamma^r)$ given by* (1.4) *satisfies*

$$V_\gamma^j(x) \le \min_{i\neq j}\{V_\gamma^i(x) + k(a^j, a^i)\}.$$

Proof. Fix a pair of indices $i, j \in \{1, \dots, r\}$ with $i \neq j$. For a given $x \in \mathbb{R}^n$, $\alpha \in \Gamma$, $b \in \mathcal{B}$, and $T > 0$, we have

$$\int_{[0,T)} l(y_x(s), a^j, \alpha_x^j[b](x), b(s)) = k(a^j, \alpha_x^j[b](0)) + \int_{[0,T)} l(y_x(s), \alpha_x^j[b](0), \alpha_x^j[b](s), b(s)).$$

Note that there are three cases to consider: (i) $\alpha_x^j[b](0) = j$, (ii) $\alpha_x^j[b](0) = i$, (iii) $\alpha_x^j[b](0) \neq j \neq i$. If (i) or (ii) occurs, then

$$\begin{aligned}\int_{[0,T)} & l(y_x(s), a^j, \alpha_x^j[b](x), b(s)) \\ &< k(a^j, a^i) + k(a^i, \alpha_x^j[b](0)) + \int_{[0,T)} l(y_x(s), \alpha_x^j[b](0), \alpha_x^j[b](s), b(s)) \\ &= k(a^j, a^i) + \int_{[0,T)} l(y_x(s), a^i, \alpha_x^j[b](s), b(s)). \end{aligned} \tag{3.1}$$

If (iii) occurs, then

$$
\begin{aligned}
&\int_{[0,T)} l(y_x(s), a^j, \alpha_x^j[b](x), b(s)) \\
&\quad = k(a^j, \alpha_x^j[b](0)) - k(a^i, \alpha_x^j[b](0)) \\
&\qquad + k(a^i, \alpha_x^j[b](0)) + \int_{[0,T)} l(y_x(s), \alpha_x^j[b](0), \alpha_x^j[b](s), b(s)) \\
&\quad = k(a^j, \alpha_x^j[b](0)) - k(a^i, \alpha_x^j[b](0)) + \int_{[0,T)} l(y_x(s), a^i, \alpha_x^j[b](s), b(s)) \\
&\quad < k(a^j, a^i) + \int_{[0,T)} l(y_x(s), a^i, \alpha_x^j[b](s), b(s)),
\end{aligned} \tag{3.2}
$$

where the last inequality follows from (A5). By the definition of $V_\gamma^j(x)$, we have

$$V_\gamma^j(x) \le \sup_{b\in\mathcal{B}, T\ge 0} \int_{[0,T)} l(y_x(s), a^j, \alpha_x^j[b](s), b(s))$$

for all $\alpha \in \Gamma$. Taking the supremum over $b \in \mathcal{B}$ and $T \ge 0$ on the right-hand side of (3.1) or (3.2) therefore gives

$$V_\gamma^j(x) \le k(a^j, a^i) + \sup_{b\in\mathcal{B}, T\ge 0} \int_{[0,T)} l(y_x(s), a^i, \alpha_x^j[b](s), b(s)). \tag{3.3}$$

Given any strategy $\alpha \in \Gamma$, we can always find another $\widetilde{\alpha} \in \Gamma$ with $\widetilde{\alpha}_x^i[b] = \alpha_x^j[b]$ for each $b \in \mathcal{B}$, and, conversely, for any $\widetilde{\alpha} \in \Gamma$, there is an $\alpha \in \Gamma$ so that $\widetilde{\alpha}_x^i$ is determined by α in this way. Hence, taking the infimum over all $\alpha \in \Gamma$ in the last terms on the right-hand side of (3.3) leaves us with $V_\gamma^i(x)$. Thus

$$V_\gamma^j(x) \le k(a^j, a^i) + V_\gamma^i(x).$$

Since $i \ne j$ is arbitrary, the result follows. □

THEOREM 3.2 (DPP). *Assume* (A1)–(A4). *Then, for* $j = 1, 2, \dots, r$, $t > 0$, *and* $x \in \mathbb{R}^N$, *we have*

$$
\begin{aligned}
V_\gamma^j(x) = \inf_{\alpha\in\Gamma} \sup_{b\in\mathcal{B},\ T>0} \Bigg\{ &\int_{[0,t\wedge T)} l(y_x(s, a^j, \alpha_x^j[b], b), \alpha_x^j[b](s), b(s)) \\
&+ 1_{[0,T)}(t) V_\gamma^i(y_x(t, \alpha_x^j[b], b)) \text{ such that } \alpha_x^j[b](t^-) = a^i \Bigg\},
\end{aligned} \tag{3.4}
$$

where

$$l(y(s), a^j, a(s), b(s)) = [h(y(s), a(s), b(s)) - \gamma^2 |b(s)|^2] ds + k(a(s^-), a(s))\delta_s$$

with $a(0^-) = a^j$.

Proof. Fix $x \in \mathbb{R}^N$, $j \in \{1, 2, \dots, r\}$, and $t > 0$. We denote by $\omega(x)$ the right-hand side of (3.4). Let $\epsilon > 0$. For any $z \in \mathbb{R}^N$ and any $a^\ell \in A$, we pick $\bar{\alpha} \in \Gamma$ such that

$$V_\gamma^\ell(z) + \epsilon \ge \int_{[0,T)} l(y_z(s), a^\ell, \bar{\alpha}_z^\ell[b](s), b(s)) \quad \text{for all } b \in \mathcal{B}, \text{ for all } T > 0. \tag{3.5}$$

We first want to show that $\omega(x) \geq V_\gamma^j(x)$. Choose $\hat{\alpha} \in \Gamma$ such that

$$\omega(x) + \epsilon \geq \sup_{b \in \mathcal{B},\, T \geq 0} \left\{ \int_{[0,t\wedge T)} l(y_x(s), a^j, \hat{\alpha}_x^j[b](s), b(s)) + 1_{[0,T)}(t) V_\gamma^i(y_x(t)),\ \hat{\alpha}_x^j[b](t^-) = a^i \right\}. \tag{3.6}$$

For each $b \in \mathcal{B}$ and $T > 0$, choose $\delta \in \Gamma$ so that

$$\delta_x^j[b](s) = \begin{cases} \hat{\alpha}_x^j[b](s), & s < t \wedge T, \\ \bar{\alpha}_z^i[b(\cdot + t \wedge T)](s - (t \wedge T)), & s \geq t \wedge T, \end{cases}$$

with $z = y_x(t\wedge T, \hat{\alpha}_x^j[b], b)$ and $a^i = \hat{\alpha}_x^j[b](t\wedge T)$. Clearly, δ_x^j is nonanticipating because $\hat{\alpha}_x^j$ and $\bar{\alpha}_z^i$ are. Note that

$$y_x(s + t \wedge T, \delta_x^j[b], b) = y_z(s, \bar{\alpha}_z^i[b(\cdot + t \wedge T)], b(\cdot + t \wedge T)) \text{ for } s \geq 0.$$

Thus, by the change of variables $\tau = s + t \wedge T$, we have

$$\int_{[0,T-(t\wedge T))} l(y_z(s), a^i, \bar{\alpha}_z^i[b(\cdot + t \wedge T)](s), b(s + t \wedge T)) = \int_{[t\wedge T,T)} l(y_x(\tau), a^j, \delta_x^j[b](\tau), b(\tau)). \tag{3.7}$$

As a consequence of (3.5), (3.6), and (3.7), we have

$$\begin{aligned} 2\omega(x) + 2\epsilon &\geq \sup_{b \in \mathcal{B},\, T>0} \left\{ \int_{[0,t\wedge T)} l(y_x(s), a^j, \hat{\alpha}_x^j[b](s), b(s)) + 1_{[0,T)}(t) \int_{[t\wedge T,T)} l(y_z(s), a^i, \bar{\alpha}_z^i[b](s), b(s)) \right\} \\ &= \sup_{b \in \mathcal{B},\, T>0} \left\{ \int_{[0,T)} l(y_x(s), a^j, \delta_x^j[b](s), b(s)) \right\} \\ &\geq \inf_{\alpha \in \Gamma} \sup_{b \in \mathcal{B},\, T>0} \left\{ \int_{[0,T)} l(y_x(s), a^j, \alpha_x^j[b](s), b(s)) \right\} \\ &= V_\gamma^j(x). \end{aligned}$$

Since $\epsilon > 0$ is arbitrary, we conclude that $\omega(x) \geq V_\gamma^j(x)$.

Next we want to show that $\omega(x) \leq V_\gamma^j(x)$. From the definition of $\omega(x)$, choose $b_1 \in \mathcal{B}$ and $T_1 \geq 0$ such that

$$\omega(x) - \epsilon \leq \int_{[0,T_1\wedge t)} l(y_x(s), a^j, \bar{\alpha}_x^j[b_1](s), b_1(s)) + 1_{[0,T_1)}(t) V_\gamma^i(y_x(t)), \tag{3.8}$$

where $\bar{\alpha}_x^j$ is defined as in (3.5) and $\bar{\alpha}_x^j[b_1](t^-) = a^i$ for some $a^i \in A$. If $t \geq T_1$, we

have

$$\begin{aligned}\omega(x)-\epsilon &\le \int_{[0,T_1)} l(y_x(s),a^j,\bar\alpha_x^j[b_1](s),b_1(s))\\ &\le \sup_{b\in\mathcal{B},\,T>0}\left\{\int_{[0,T)} l(y_x(s),a^j,\bar\alpha_x^j[b](s),b(s))\right\}\\ &\le V_\gamma^j(x)+\epsilon,\end{aligned}$$

where the last inequality follows from (3.5). If $t<T_1$, we have

$$\omega(x)-\epsilon \le \int_{[0,t)} l(y_x(s),\bar\alpha_x^j[b_1](s),b_1(s)) + V_\gamma^i(y_x(t)). \tag{3.9}$$

Set $z=y_x(t,\bar\alpha_x^j[b_1],b_1)$. For each $b\in\mathcal{B}$, define $\tilde b\in\mathcal{B}$ by

$$\tilde b(s)=\begin{cases} b_1(s), & s<t,\\ b(s-t), & s\ge t,\end{cases}$$

and choose $\widehat\alpha\in\Gamma$ so that

$$\widehat\alpha_z[b](s)=\bar\alpha_x^j[\tilde b](s+t)\qquad \text{for } s\ge 0.$$

By definition of V_γ^i, choose $b_2\in\mathcal{B}$ and $T_2>0$ such that

$$V_\gamma^i(z)-\epsilon \le \int_{[0,T_2)} l(y_z(s),a^i,\widehat\alpha_z[b_2](s),b_2(s)).$$

Then, by change of variable $\tau=s+t$, we have

$$V_\gamma^i(z)-\epsilon \le \int_{[t,t+T_2)} l(y_x(\tau),a^j,\bar\alpha_x^j[\tilde b_2](\tau),\tilde b_2(\tau)). \tag{3.10}$$

As a consequence of (3.9) and (3.10), we have

$$\begin{aligned}\omega(x)-2\epsilon &\le \int_{[0,t)} l(y_x(s),a^j,\bar\alpha_x^j[b_1](s),b_1(s)) + \int_{[t,t+T_2)} l(y_x(\tau),a^j,\bar\alpha_x^j[\tilde b_2](\tau),\tilde b_2(\tau))\\ &= \int_{[0,t+T_2)} l(y_x(\tau),a^j,\bar\alpha_x^j[\tilde b_2](\tau),\tilde b_2(\tau))\\ &\le \sup_{b\in\mathcal{B},\,T>0}\left\{\int_{[0,T)} l(y_x(\tau),a^j,\bar\alpha_x^j[b](\tau),b(\tau))\right\}\\ &\le V_\gamma^j(x)+\epsilon,\end{aligned}$$

where the last inequality follows from (3.5). Since $\epsilon>0$ is arbitrary, for both cases we have $\omega(x)\le V_\gamma^j(x)$ as required. □

COROLLARY 3.3. *Assume* (A1)–(A4) *and* (A6). *Then, for each* $j\in\{1,\dots,r\}$, $x\in\mathbb{R}^N$, *and* $t>0$, *we have*

$$V_\gamma^j(x) \le \sup_{b\in\mathcal{B},\,T>0}\left\{\int_0^{t\wedge T}[h(y_x(s),a^j,b(s))-\gamma^2|b(s)|^2]ds + 1_{[0,T)}(t)V_\gamma^j(y_x(t))\right\} \tag{3.11}$$

$$\le \sup_{b\in\mathcal{B}}\int_0^t [h(y_x(s),a^j,b(s))-\gamma^2|b(s)|^2]\,ds + V_\gamma^j(y_x(t)). \tag{3.12}$$

Proof. Fix $j \in \{1, \dots, r\}$, $x \in \mathbb{R}^N$, and $t > 0$. Define $\alpha \in \Gamma$ by setting $\alpha_x^j[b](s) = a^j$ for all $s \geq 0$ for each $b \in \mathcal{B}$. By Theorem 3.2, we have

$$V_\gamma^j(x) \leq \sup_{b \in \mathcal{B},\, T>0} \left\{ \int_0^{t \wedge T} [h(y_x(s), a^j, b(s)) - \gamma^2 |b(s)|^2] ds + 1_{[0,T)}(t) V_\gamma^j(y_x(t)) \right\},$$

and (3.11) follows.

To prove the second inequality (3.12), consider any $b \in \mathcal{B}$ and T with $0 < T < t$. Define a new $\bar{b} \in \mathcal{B}$ by $\bar{b}(s) = b(s)$ for $s \leq T$ and $\bar{b}(s) = 0$ for $s > T$. It follows that $y_x(s, a^j, b) = y_x(s, a^j, \bar{b})$ for $0 \leq s \leq T$. For $s > T$, we have

$$h(y_x(s), a^j, b(s)) - \gamma^2 |b(s)|^2 = h(y_x(s), a^j, 0) \geq 0$$

(by (A6)). Since we also know that $V_\gamma^j(x) \geq 0$ for all $x \in \mathbb{R}^N$ (take $T = 0$ in the definition (1.4)), we get

$$\begin{aligned} &\int_0^{t \wedge T} [h(y_x(s), a^j, b(s)) - \gamma^2 |b(s)|^2] \, ds + 1_{[0,T)}(t) V_\gamma^j(y_x(t)) \\ &\leq \int_0^t [h(y_x(s), a^j, \bar{b}(s)) - \gamma^2 |b(s)|^2] \, ds + V_\gamma^j(y_x(t, a^j, \bar{b})), \end{aligned}$$

and (3.12) follows as well. ∎

PROPOSITION 3.4. *Assume* (A1)–(A5). *Suppose that, for each* $j \in \{1, \dots, r\}$, V^j *is continuous. If* $V_\gamma^j(x) < \min_{i \neq j}\{V_\gamma^i(x) + k(a^j, a^i)\}$, *then there exists* $\tau = \tau_x > 0$ *such that, for* $0 < t < \tau_x$,

$$V_\gamma^j(x) = \sup_{b \in \mathcal{B},\, T>0} \left\{ \int_0^{t \wedge T} [h(y_x(s), a^j, b(s)) - \gamma^2 |b(s)|^2] ds + 1_{[0,T)}(t) V_\gamma^j(y_x(t)) \right\}.$$

Proof. We assume $V_\gamma^j(x) < \min_{i \neq j}\{V_\gamma^i(x) + k(a^j, a^i)\}$. From Corollary 3.3, we know that, for all $t > 0$,

$$V_\gamma^j(x) \leq \sup_{b \in \mathcal{B},\, T>0} \left\{ \int_0^{t \wedge T} [h(y_x(s), a^j, b(s)) - \gamma^2 |b(s)|^2] ds + 1_{[0,T)}(t) V_\gamma^j(y_x(t)) \right\}.$$

Suppose there is a sequence $\{t_n\}$ with $0 < t_n < \frac{1}{n}$ for $n = 1, 2, \dots$ such that

(3.13)
$$V_\gamma^j(x) < \sup_{b \in \mathcal{B},\, T>0} \left\{ \int_0^{t_n \wedge T} [h(y_x(s), a^j, b(s)) - \gamma^2 |b(s)|^2] ds + 1_{[0,T)}(t_n) V_\gamma^j(y_x(t_n)) \right\}.$$

Let $w(x, t_n)$ be the right-hand side of (3.13). For each t_n, define $\epsilon_n = \frac{1}{3}[w(x, t_n) - V_\gamma^j(x)]$. As $t_n \to 0$ as $n \to \infty$, from (3.13) we see that $w(x, t_n) \to V_\gamma^j(x)$, and hence $\epsilon_n \to 0$ as $n \to \infty$. It follows that

$$V_\gamma^j(x) + \epsilon_n < w(x, t_n) - \epsilon_n. \tag{3.14}$$

Choose $b_n \in \mathcal{B}$ and $T_n \geq 0$ such that

(3.15)
$$w(x, t_n) - \epsilon_n \leq \int_0^{t_n \wedge T_n} [h(y_x(s), a^j, b_n(s)) - \gamma^2 |b_n(s)|^2] ds + 1_{[0,T_n)}(t_n) V_\gamma^j(y_x(t_n)).$$

By Theorem 3.2, choose $\alpha_n \in \Gamma$ such that

(3.16)
$$V_\gamma^j(x) + \epsilon_n \geq \int_{[0,t_n\wedge T_n]} l(y_x(s), a^j, (\alpha_n)_x^j[b_n](s), b_n(s)) + 1_{[0,T_n)}(t_n)V_\gamma^{i_n}(y_x(t_n)),$$

where $(\alpha_n)_x^j[b_n](t_n^-) = a^{i_n} \in A$. From (3.14), (3.15), and (3.16), we have

(3.17)
$$\begin{aligned} &\int_{[0,t_n\wedge T_n)} l(y_x(s), a^j, (\alpha_n)_x^j[b_n](s), b_n(s)) + 1_{[0,T_n)}(t_n)V_\gamma^{i_n}(y_x(t_n)) \\ &< \int_0^{t_n\wedge T_n} [h(y_x(s), a^j, b_n(s)) - \gamma^2|b_n(s)|^2]ds + 1_{[0,T_n)}(t_n)V_\gamma^j(y_x(t_n)). \end{aligned}$$

This implies that $(\alpha_n)_x^j[b_n]$ jumps in the interval $[0, t_n\wedge T_n]$. Without loss of generality, assume the number of switchings is equal to d_n. If $t_n < T_n$ for infinitely many n, by going down to a subsequence we may assume that $t_n \leq T_n$ for all n. From (3.16), we have

$$\begin{aligned} V_\gamma^j(x) &\geq \limsup_{n\to\infty} \left\{ \int_{[0,t_n\wedge T_n)} l(y_x(s), a^j, \alpha_{x,n}^j[b_n](s), b_n(s)) \right. \\ &\qquad \left. + 1_{[0,T_n)}(t_n)V_\gamma^{i_n}(y_x(t_n)),\ \alpha_{x,n}^j[b_n](t_n^-) = a^{i_n} \in A \right\} \\ &= \limsup_{n\to\infty} \left\{ \int_0^{t_n} [h(y_x(s), \alpha_{x,n}^j[b_n](s), b_n(s)) - \gamma^2|b_n(s)|^2]ds \right. \\ &\qquad \left. + \sum_{m=1}^{d_n} k(a_{m-1}, a_m) + V_\gamma^{i_n}(y_x(t_n)),\ \alpha_{x,n}^j[b_n](t_n) = a^{i_n} \in A \right\} \\ &= \limsup_{n\to\infty} \left\{ \sum_{m=1}^{d_n} k(a_{m-1}, a_m) + V_\gamma^{i_n}(y_x(t_n)),\ \alpha_{x,n}^j[b_n](t_n^-) = a^{i_n} \in A \right\}. \end{aligned}$$

By using the continuity of $V_\gamma^{i_n}$ and $\sum_{m=1}^{d_n} k(a_{m-1}, a_m) > k(a^j, a^{i_n})$, we have

$$V_\gamma^j(x) \geq \min_{i\neq j}\{V_\gamma^i(x) + k(a^j, a^i)\},$$

which contradicts one of the assumptions. If $t_n \geq T_n$ for infinitely many n, again without loss of generality we may assume that $t_n \geq T_n$ for all n. From (3.17), we have

$$\begin{aligned} &\liminf_{n\to\infty} \left\{ \int_{[0,T_n]} l(y_x(s), a^j\alpha_{x,n}^j[b_n](s), b_n(s)) \right\} \\ &\leq \limsup_{n\to\infty} \left\{ \int_0^{T_n} [h(y_x(s), a^j, b_n(s)) - \gamma^2|b_n(s)|^2]ds \right\}, \end{aligned}$$

or, equivalently,

$$\begin{aligned} &\liminf_{n\to\infty} \left\{ \int_0^{T_n} [h(y_x(s), \alpha_{x,n}^j[b_n](s), b_n(s)) - \gamma^2|b_n(s)|^2]ds + \sum_{m=1}^{d_n} k(a_{m-1}, a_m) \right\} \\ &\leq \limsup_{n\to\infty} \left\{ \int_0^{T_n} [h(y_x(s), a^j, b_n(s)) - \gamma^2|b_n(s)|^2]ds \right\}. \end{aligned}$$

Thus

$$\liminf_{n\to\infty}\left\{\sum_{m=1}^{d_n} k(a_{m-1},a_m)\right\} \le \limsup_{n\to\infty}\left\{\int_0^{T_n} h(y_x(s),a^j,b_n(s))ds\right\} - \liminf_{n\to\infty}\left\{\int_0^{T_n} h(y_x(s),\alpha^j_{x,n}[b_n](s),b_n(s))ds\right\},$$

and in this case $T_n \to 0$ as $n \to \infty$. Note that the integral terms tend to 0 uniformly with respect to $b_n \in \mathcal{B}$ as $T_n \to 0$ by assumption (A2), the uniform estimate (2.3), and the continuity assumption (A1) on h. Thus we have

$$\liminf_{n\to\infty}\left\{\sum_{m=1}^{d_n} k(a_{m-1},a_m)\right\} \le 0,$$

which contradicts (A5). □

LEMMA 3.5. *Assume* (A1)–(A6) *and* $V^j_\gamma \in C(\mathbb{R}^N)$, $j = 1,\dots,r$. *If* $V^j_\gamma(x) < \min_{i\neq j}\{V^i_\gamma(x) + k(a^j,a^i)\}$, *then there exists* $\tau = \tau_x > 0$ *such that*

$$V^j_\gamma(x) \ge \sup_{b\in\mathcal{B}}\left\{\int_0^t [h(y_x(s),a^j,b(s)) - \gamma^2|b(s)|^2]ds + V^j_\gamma(y_x(t))\right\} \quad \textit{for all } t \in (0,\tau_x).$$

Proof. From Proposition 3.4, choose $\tau = \tau_x > 0$ such that, for all $t \in (0,\tau)$,

$$V^j_\gamma(x) = \sup_{b\in\mathcal{B},\, T>0}\left\{\int_0^{t\wedge T} [h(y_x(s),a^j,b(s)) - \gamma^2|b(s)|^2]ds + 1_{[0,T)}(t)V^j_\gamma(y_x(t))\right\}.$$

Thus

$$\begin{aligned} V^j_\gamma(x) &\ge \sup_{b\in\mathcal{B},\, T>t}\left\{\int_0^{t\wedge T} [h(y_x(s),a^j,b(s)) - \gamma^2|b(s)|^2]ds + 1_{[0,T)}(t)V^j_\gamma(y_x(t))\right\} \\ &= \sup_{b\in\mathcal{B}}\left\{\int_0^t [h(y_x(s),a^j,b(s)) - \gamma^2|b(s)|^2]ds + V^j_\gamma(y_x(t))\right\}. \end{aligned}$$ □

THEOREM 3.6. *Assume* (A1)–(A6) *and* $V^j_\gamma \in C(\mathbb{R}^N)$, $j = 1,\dots,r$. *Then* V_γ *is a viscosity solution of the SQVI* (2.5)

(3.18)
$$\max\left\{H^j(x,DV^j_\gamma(x)),\ V^j_\gamma(x) - \min_{i\neq j}\{V^i_\gamma(x) + k(a^j,a^i)\}\right\} = 0, x \in \mathbb{R}^N,\ j = 1,\dots,r.$$

Proof. We first show that V^j_γ is a viscosity supersolution of the SQVI (3.18). Fix $x_0 \in \mathbb{R}^N$ and $a^j \in A$. Let $\varphi^j \in C^1(\mathbb{R}^N)$, and x_0 is a local minimum of $V^j_\gamma - \varphi^j$. We want to show that

$$\max\left\{H^j(x_0,D\varphi^j(x_0)),\ V^j_\gamma(x_0) - \min_{i\neq j}\{V^i_\gamma(x_0) + k(a^j,a^i)\}\right\} \ge 0. \tag{3.19}$$

We have two cases to consider.

Case 1. $V^j_\gamma(x_0) = \min_{i\neq j}\{V^i_\gamma(x_0) + k(a^j,a^i)\}$.

Case 2. $V_\gamma^j(x_0) < \min_{i\neq j}\{V_\gamma^i(x_0) + k(a^j, a^i)\}$.
If Case 1 occurs, we have

$$\begin{aligned}&\max\left\{H^j(x_0, D\varphi^j(x_0)),\ V_\gamma^j(x_0) - \min_{i\neq j}\{V_\gamma^i(x_0) + k(a^j, a^i)\}\right\}\\ &\geq V_\gamma^j(x_0) - \min_{i\neq j}\{V_\gamma^i(x_0) + k(a^j, a^i)\}\\ &\geq 0.\end{aligned}$$

If Case 2 occurs, we want to show that $H^j(x_0, D\varphi^j(x_0)) \geq 0$. Fix $b \in B$, and set $b(s) = b$ for all $s \geq 0$. From Lemma 3.5, choose $\bar{t}_0 > 0$ such that, for $t \in (0, \bar{t}_0)$,

$$V_\gamma^j(x_0) - V_\gamma^j(y_{x_0}(t)) \geq \int_0^t [h(y_{x_0}(s), a^j, b) - \gamma^2|b|^2]\, ds. \tag{3.20}$$

Since x_0 is a local minimum of $V_\gamma^j - \varphi^j$, by (2.3) there exists $\hat{t}_0 > 0$ such that

$$\varphi^j(x_0) - \varphi^j(y_{x_0}(s), a^j, b(s)) \geq V_\gamma^j(x_0) - V_\gamma^j(y_{x_0}(s), a^j, b(s)), \quad 0 < s < \hat{t}_0. \tag{3.21}$$

Set $t_0 = \min\{\bar{t}_0,\ \hat{t}_0\}$. As a consequence of (3.20) and (3.21), we have

$$\varphi^j(x_0) - \varphi^j(y_{x_0}(t)) \geq \int_0^t [h(y_{x_0}(s), a^j, b) - \gamma^2|b|^2]ds, \quad 0 < t < t_0. \tag{3.22}$$

Divide both sides by t, and let $t \to 0$ to get

$$-D\varphi^j(x_0) \cdot f(x_0, a^j, b) - h(x_0, a^j, b) + \gamma^2|b|^2 \geq 0.$$

Since $b \in B$ is arbitrary, we have $H^j(x_0, D\varphi^j(x_0)) \geq 0$.

We next show that V_γ^j is a viscosity subsolution of the SQVI (3.18). Fix $x_1 \in \mathbb{R}^N$ and $a^j \in A$. Let $\varphi^j \in C^1(\mathbb{R}^N)$, and x_1 is a local maximum of $V_\gamma^j - \varphi^j$. We want to show that

$$\max\left\{H^j(x_1, D\varphi^j(x_1)),\ V_\gamma^j(x_1) - \min_{i\neq j}\{V_\gamma^i(x_1) + k(a^j, a^i)\}\right\} \leq 0. \tag{3.23}$$

From Proposition 3.1, $V_\gamma^j(x_1) \leq \min_{i\neq j}\{V_\gamma^i(x_1) + k(a^j, a^i)\}$. Thus we want to show that $H^j(x_1, D\varphi^j(x_1)) \leq 0$.

Let $t > 0$ and $\epsilon > 0$. From (3.12) in Corollary 3.3, we may choose $\hat{b} = \hat{b}_{t,\epsilon} \in \mathcal{B}$ such that

$$V_\gamma^j(x_1) \leq \int_0^t [h(y_{x_1}(s), a^j, \hat{b}(s)) - \gamma^2|\hat{b}(s)|^2]\, ds + V_\gamma^j(y_{x_1}(t, \hat{b})) + \epsilon t, \tag{3.24}$$

and hence

$$V_\gamma^j(x_1) - V_\gamma^j(y_{x_1}(t, \hat{b})) \leq \int_0^t [h(y_{x_1}(s), a^j, \hat{b}(s)) - \gamma^2|\hat{b}(s)|^2]\, ds + \epsilon t. \tag{3.25}$$

Since x_1 is a local maximum of $V_\gamma^j - \varphi^j$, by (2.3) we may assume that

$$\varphi^j(x_1) - \varphi^j(y_{x_1}(s), a^j, \hat{b}(s)) \leq V_\gamma^j(x_1) - V_\gamma^j(y_{x_1}(s), a^j, \hat{b}(s)), \quad 0 < s \leq t. \tag{3.26}$$

Combine (3.25) and (3.26) to get

$$\varphi^j(x_1) - \varphi^j(y_{x_1}(t, a^j, \hat{b}(t))) \le \int_0^t [h(y_{x_1}(s), a^j, \hat{b}(s)) - \gamma^2 |\hat{b}(s)|^2]\, ds + \epsilon t. \tag{3.27}$$

Observe that (2.3) and (A3) imply

$$|f(y_{x_1}(s), a^j, \hat{b}(s)) - f(x_1, a^j, \hat{b}(s))| \le \omega_f(M_x s, |x| + M_x t_0) \text{ for } 0 < s < t_0 \tag{3.28}$$

and

$$|h(y_x(s), a^j, \hat{b}(s)) - h(x_1, a^j, \hat{b}(s))| \le \omega_h(M_{x_1} s, |x| + M_{x_1} t_0) \text{ for } 0 < s < t_0, \tag{3.29}$$

where t_0 does not depend on ϵ, t, or $\hat{b}$. By (3.29), the integral on the right-hand side of (3.27) can be written as

$$\int_0^t [h(x_1, a^j, \hat{b}(s)) - \gamma^2 |\hat{b}(s)|^2]\, ds + \mathrm{o}(t) \text{ as } t \to 0.$$

Thus

$$\varphi^j(x_1) - \varphi^j(y_{x_1}(t, a^j, \hat{b}(t))) \le \int_0^t [h(x_1, a^j, \hat{b}(s)) - \gamma^2 |\hat{b}(s)|^2]\, ds + \epsilon t + \mathrm{o}(t). \tag{3.30}$$

Moreover,

$$\begin{aligned} \varphi^j(x_1) - \varphi^j(y_{x_1}(t, a^j, \hat{b})) &= -\int_0^t \frac{d}{ds} \varphi^j(y_{x_1}(s, a^j, \hat{b}))\, ds \\ &= -\int_0^t D\varphi^j(y_{x_1}(s, a^j, \hat{b})) \cdot f(y_{x_1}(s), a^j, \hat{b}(s))\, ds \\ &= -\int_0^t D\varphi^j \cdot f(x_1, a^j, \hat{b}(s))\, ds + \mathrm{o}(t), \end{aligned} \tag{3.31}$$

where we used (2.3), (3.28), and $\varphi^j \in C^1$ in the last equality to estimate the difference between $D\varphi^j \cdot f$ computed at $y_{x_1}(s)$ and at x_1, respectively. Plugging (3.31) into (3.30) gives

$$\int_0^t -D\varphi^j(x_1) \cdot f(x_1, a^j, \hat{b}(s))\, ds \le \int_0^t [h(x_1, a^j, \hat{b}(s)) - \gamma^2 |\hat{b}|^2]\, ds + \epsilon t + \mathrm{o}(t).$$

Thus

$$\int_0^t [-D\varphi^j(x_1) \cdot f(x_1, a^j, \hat{b}(s)) - h(x_1, a^j, \hat{b}(s)) + \gamma^2 |\hat{b}(s)|^2]\, ds \le \epsilon t + \mathrm{o}(t). \tag{3.32}$$

We estimate the left-hand side of this inequality from below next to get

$$\inf_{b \in B} \{-D\varphi^j(x_1) \cdot f(x_1, a^j, b) - h(x_1, a^j, b) + \gamma^2 |b|^2\} \cdot t \le \epsilon t + \mathrm{o}(t). \tag{3.33}$$

Divide by t, and pass to the limit as $t \to 0$ to get

$$\inf_{b \in B} \{-D\varphi^j(x) \cdot f(x, a^j, b) - h(x, a^j, b) + \gamma^2 |b|^2\} \le \epsilon.$$

Since $\epsilon > 0$ is arbitrary, we conclude that $H^j(x, D\varphi^j(x)) \le 0$. ∎

We next give a connection of a switching storage (vector) function with the SQVI (3.18).

THEOREM 3.7. *Assume* (A1)–(A5), *and assume that* $S = (S^1, \dots, S^r)$ *is a continuous switching-storage function for the closed-loop system formed by the nonanticipating strategy* $\alpha \in \Gamma$. *Then* S *is a viscosity supersolution of SQVI* (3.18).

Proof. The proof follows exactly as the proof that the lower-value function V^γ is a viscosity subsolution in the proof of Theorem 3.6 once we verify the following analogue of Lemma 3.5 for switching-storage functions. ∎

LEMMA 3.8. *Assume* (A1)–(A6), *and assume that* $S = (S^1, \dots, S^r)$ *is a continuous switching-storage function. If* $S^j(x) < \min_{i \ne j}\{S^i(x) + k(a^j, a^i)\}$, *then there exists* $\tau = \tau_x > 0$ *such that*

$$S^j(x) \ge \sup_{b \in \mathcal{B}} \left\{ \int_0^t [h(y_x(s), a^j, b(s)) - \gamma^2 |b(s)|^2] ds + S^j(y_x(t)) \right\} \quad \text{for all } t \in (0, \tau_x).$$

Proof. By the defining condition (1.5) for a storage function (1.5) (with attenuation level γ), we have

$$S^j(x) \ge \int_0^t [h(y_x(s), \alpha_x^j[b](s), b(s)) - \gamma^2 |b(s)|^2] \, ds + \sum_{0 \le \tau < t} k(\alpha_x^j[b](\tau^-), \alpha_x^j[b](\tau)) + S^{j(t)}(y_x(t, \alpha_x^j[b], b)). \tag{3.34}$$

Due to the assumed boundedness of B and the boundedness of h (see (A2)), it is clear that, given $\epsilon > 0$, we may choose $\tau = \tau_\epsilon$ so that

$$\sup_{b \in \mathcal{B}} \left\{ \int_0^t [h(y_x(s), a^j, b(s)) - \gamma^2 |b(s)|^2] ds \right\} > -\epsilon$$

for all $t \in [0, \tau_\epsilon)$. We conclude that, for any such t,

$$S^j(x) \ge -\epsilon + \sum_{0 \le \tau < t} k(\alpha_x^j[b](\tau^-), \alpha_x^j[b](\tau)) + S^{j(t)}(y_x(t, \alpha_x^j[b], b)). \tag{3.35}$$

If we now choose $\epsilon = \frac{1}{2}[\min_{i \ne j}\{S^i(x) + k(a^j, a^i)\} - S^j(x)] > 0$ and use the continuity of S^j for each j, the estimate (3.35) in the presence of any jumps in the interval $[0, \tau_\epsilon)$ leads to a contradiction. Since we are now assured that there are no jumps, (3.34) collapses to

$$S^j(x) \ge \int_0^t [h(y_x(s), \alpha_x^j[b](s), b(s)) - \gamma^2 |b(s)|^2] \, ds + S^j(y_x(t, \alpha_x^j[b], b))$$

for $0 \le t < \tau_\epsilon$ and for all $b \in \mathcal{B}$. Taking the supremum of $b \in \mathcal{B}$ now leads to the desired result. This concludes the proof of Lemma 3.8 and of Theorem 3.7. ∎

We now proceed to the synthesis of a switching-control strategy achieving the dissipation inequality for a given viscosity supersolution $U = (U^1, \dots, U^r)$ of SQVI (3.18). Given a continuous nonnegative vector function $U = (U^1, \dots, U^r)$ on $\mathbb{R}^N$ satisfying the condition

$$U^j(x) \le \min_{i \ne j}\{U^i(x) + k(a^j, a^i)\} \text{ for all } x \in \mathbb{R}^N, \ j = 1, \dots, r,$$

we associate a state-feedback switching strategy $\alpha_U : (y(t), a^j) \to \alpha^j(y(t))$ by the rule

$$\alpha^j(y(t)) = \begin{cases} a^j & \text{if } U^j(y(t)) < \min_{i\neq j}\{U^i(y(t)) + k(a^j, a^i)\}, \\ a^\ell & \text{for any } \ell \in \arg\min_{i\neq j}\{U^i(y(t)) + k(a^j, a^i)\} \text{ otherwise.} \end{cases} \tag{3.36}$$

In other words, the associated feedback switching strategy is as follows: *if the current state is $y(t)$ and the current old control is $a(t^-) = a^j$, then set $a(t) = \alpha^j(y(t))$.* Such a strategy can also be expressed as a nonanticipating strategy $\alpha_U : (x, a^j, b) \to \alpha^j_{U,x}[b]$; explicitly, for this particular case α_U, we have that $\alpha^j_{U,x}[b]$ is given by

$$\alpha^j_{U,x}[b](t) = \sum_{n\geq 1} a_{n-1} 1_{[\tau_{n-1},\tau_n)}(t) \text{ for } t \geq 0 \tag{3.37}$$

and $\alpha^j_{U,x}[b](0^-) = a_0$, where

$$\tau_0 = 0, \quad a_0 = a^{j_0} = a^j,$$

and, for $n = 1, 2, 3, \ldots$, $\tau_n[b]$ is the infimum over $t > \tau_{n-1}$ for which

$$\begin{aligned} &U^{j_{n-1}}(y_{y(\tau_{n-1})}(t - \tau_{n-1}, a^{j_{n-1}}, b(\cdot - \tau_{n-1}))) \\ &\qquad = \min_{i\neq j_{n-1}} \{U^i(y_{y(\tau_{n-1})}(t - \tau_{n-1}, a^{j_{n-1}}, b(\cdot - \tau_{n-1}))) + k(a^{j_{n-1}}, a^i)\}\}, \end{aligned}$$

or $+\infty$ if the preceding set is empty; and $a_n = a^{j_n} =$ any $a^l \neq a^{j_{n-1}}$ for which

$$\begin{aligned} &\min_{i\neq j_{n-1}} \{U^i(y_{y(\tau_{n-1})}(\tau_n - \tau_{n-1}, a^{j_{n-1}}, b(\cdot - \tau_{n-1}))) + k(a^{j_{n-1}}, a^i)\} \\ &\qquad = U^l(y_{y(\tau_{n-1})}(\tau_n - \tau_{n-1}, a^{j_{n-1}}, b(\cdot - \tau_{n-1}))) + k(a^{j_{n-1}}, a^l) \end{aligned}$$

if $\tau_n < \infty$ or undefined if $\tau_n = \infty$. Note that, if $\tau_1 = \tau_0 = 0$, there is an immediate switch from a_0 to a_1 at time 0, and the $n = 1$ term in (3.37) is vacuous. Moreover, by (A5), $\tau_n > \tau_{n-1}$ for $\tau_{n-1} < \infty$, and $n > 1$. To see this, we assume that $\tau_n = \tau_{n-1} < \infty$ for some $n > 1$. From the definition of τ_{n-1} and τ_n, we would have

$$\begin{aligned} U^{j_{n-2}}(y(\tau_{n-1})) &= U^{j_{n-1}}(y(\tau_{n-1})) + k(a^{j_{n-2}}, a^{j_{n-1}}) \\ &= U^{j_n}(y(\tau_{n-1})) + k(a^{j_{n-1}}, a^{j_n}) + k(a^{j_{n-2}}, a^{j_{n-1}}) \text{ (hence } j_n \neq j_{n-2}) \\ &> U^{j_n}(y(\tau_{n-1})) + k(a^{j_{n-2}}, a^{j_n}) \\ &\geq \min_{i\neq j_{n-2}} \{U^i(y(\tau_{n-1})) + k(a^{j_{n-2}}, a^i)\}, \end{aligned}$$

which gives a contradiction. Moreover, as shown in the proof of the next theorem, if $\tau_n < \infty$ for all n, it still holds that $\lim_{n\to\infty} \tau_n = \infty$, so the closed-loop trajectory is defined for all $t > 0$.

THEOREM 3.9. *Assume the following.*

(i) (A1)–(A5) *hold.*

(ii) $U = (U^1, \ldots, U^r)$ *is a nonnegative continuous viscosity supersolution in* $\mathbb{R}^N$ *of the SQVI* (3.18)

$$\max\left\{H^j(x, DU^j(x)),\ U^j(x) - \min_{i\neq j}\{U^i(x) + k(a^j, a^i)\}\right\} = 0,\ x \in \mathbb{R}^N,\ j = 1, \ldots, r.$$

(iii) $U^j(x) \le \min_{i \neq j}\{U^i(x) + k(a^j, a^i)\},\ x \in \mathbb{R}^N,\ j \in \{1, \dots, r\}$.

Let α_U be the state-feedback strategy defined by (3.36) *or, equivalently, the nonanticipating disturbance-feedback strategy α_U defined by* (3.37). *Then $U = (U^1, \dots, U^r)$ is a storage function for the closed-loop system formed by the strategy α_U. In particular, we have*

$$U^j(x) \ge \sup_{b \in \mathcal{B},\ T \ge 0} \left\{ \int_{[0,T)} l(y_x(s), a^j, \alpha^j_{U,x}[b](s), b(s)) \right\} \ge V^j_\gamma(x)$$

for each $x \in \mathbb{R}^N$ and $a^j \in A$. Thus V_γ, if continuous, is characterized as the minimal, nonnegative continuous viscosity supersolution of the SQVI (3.18) *satisfying condition* (iii) *as well as the minimal continuous switching-storage function satisfying condition* (iii) *for the closed-loop system associated with some nonanticipating strategy α_{V_γ}.*

Proof. Let $\alpha^j_{U,x}[b](t)$ be the switching strategy defined as in (3.37). We claim that

$$\tau_n \to \infty \ \text{ as } \ n \to \infty.$$

If $\tau_n = \infty$ for some n, then it is trivially true. Otherwise, since we observed just before the statement of Theorem 3.9 that $\{\tau_n\}$ is a nondecreasing sequence, it would follow that

$$\lim_{n \to \infty} \tau_n = T < \infty \tag{3.38}$$

with $0 \le \tau_n < T$ for all n. From (3.38), we have that $\{\tau_n\}$ is a Cauchy sequence, and hence for all $\nu > 0$ there is some n such that $\tau_n < \tau_{n-1} + \nu$. By the definition of τ_n,

$$U^{j_{n-1}}(y_x(\tau_n)) = U^l(y_x(\tau_n)) + k(a^{j_{n-1}}, a^l) \text{ for some } a^l \neq a^{j_{n-1}}. \tag{3.39}$$

(We have written $y_x(t)$ for $y_x(t, \alpha^j_x[b], b)$.) By definition of τ_{n-1}, we have

$$U^{j_{n-2}}(y_x(\tau_{n-1})) = U^{j_{n-1}}(y_x(\tau_{n-1})) + k(a^{j_{n-2}}, a^{j_{n-1}}). \tag{3.40}$$

By (iii), we have

$$\begin{aligned} U^{j_{n-2}}(y_x(\tau_{n-1})) &\le \min_{i \neq j_{n-2}} \{U^i(y_x(\tau_{n-1})) + k(a^{j_{n-2}}, a^i)\} \\ &\le U^l(y_x(\tau_{n-1})) + k(a^{j_{n-2}}, a^l) \text{ if } l \neq j_{n-2}, \end{aligned}$$

and hence

$$U^{j_{n-2}}(y_x(\tau_{n-1})) \le U^l(y_x(\tau_{n-1})) + k(a^{j_{n-2}}, a^l) \tag{3.41}$$

if $l \neq j_{n-2}$. If $l = j_{n-2}$, (3.41) holds with equality (by (A5)), and hence (3.41) in fact holds without restriction. From (3.40) and (3.41), we have

$$k(a^{j_{n-2}}, a^{j_{n-1}}) - k(a^{j_{n-2}}, a^l) \le U^l(y_x(\tau_{n-1})) - U^{j_{n-1}}(y_x(\tau_{n-1})). \tag{3.42}$$

As a consequence of (3.39) and (3.42), we have

$$\begin{aligned} 0 &< k(a^{j_{n-2}}, a^{j_{n-1}}) + k(a^{j_{n-1}}, a^l) - k(a^{j_{n-2}}, a^l) \\ &\le U^l(y_x(\tau_{n-1})) - U^l(y_x(\tau_n)) + U^{j_{n-1}}(y_x(\tau_n)) - U^{j_{n-1}}(y_x(\tau_{n-1})) \\ &\le \omega_l(\nu) + \omega_{j_{n-1}}(\nu), \end{aligned}$$

and hence (by the strict triangle inequality in (A5))

$$0 < \min_{i,j,l:\ i\neq j\neq l} \left\{k(a^i,a^j)+k(a^j,a^l)-k(a^i,a^l)\right\} \le \omega_\ell(\nu)+\omega_j(\nu),$$

where, in general, ω_j is a modulus of continuity for $U^j(y_x(\cdot))$ on the interval $[0,T]$. Letting ν tend to zero now leads to a contradiction, and the claim follows.

Hence $\alpha_x^j[b](t) = \sum a_{n-1} 1_{[\tau_{n-1},\tau_n)}(t) \in \Gamma$. Since U is a viscosity supersolution of the SQVI (3.18), we have $H^{j_n}(y_x(s), DU^{j_n}(y_x(s))) \ge 0$ in the viscosity-solution sense for $\tau_n < s < \tau_{n+1}$. Thus (see [6, section II.5.5])

$$U^{j_n}(y_x(s)) - U^{j_n}(y_x(t)) \ge \int_s^t [h(y_x(s), a^{j_n}, b(s)) - \gamma^2 |b(s)|^2] ds \tag{3.43}$$

for all $b \in \mathcal{B}$, $\tau_n < s \le t < \tau_{n+1}$. (This argument uses the boundedness of the disturbance set B.) Letting $s \to \tau_n^+$ and $t \to \tau_{n+1}^-$, we get

$$U^{j_n}(y_x(\tau_n)) - U^{j_n}(y_x(\tau_{n+1})) \ge \int_{\tau_n}^{\tau_{n+1}} [h(y_x(s), a^{j_n}, b(s)) - \gamma^2 |b(s)|^2] ds \quad \text{for all } b \in \mathcal{B}. \tag{3.44}$$

We also have

$$U^{j_n}(y_x(\tau_{n+1})) = U^{j_{n+1}}(y_x(\tau_{n+1})) + k(a^{j_n}, a^{j_{n+1}}) \text{ for } \tau_{n+1} < \infty. \tag{3.45}$$

Adding (3.44) over $\tau_n \le T$ and using (3.45), we have

$$\begin{aligned} U^{j_0}(x) &\ge \int_0^T [h(y_x(s), \alpha_x^j[b](s), b(s)) - \gamma^2 |b(s)|^2] ds + \sum_{\tau_n \le T} k(a_{n-1}, a_n) + U^{j_n}(y_x(T)) \\ &\ge \int_0^T [h(y_x(s), \alpha_x^j[b](s), b(s)) - \gamma^2 |b(s)|^2] ds + \sum_{\tau_n \le T} k(a_{n-1}, a_n). \end{aligned}$$

Since this inequality holds for arbitrary $b \in \mathcal{B}$ and $T \ge 0$, we have

$$U^j(x) \ge \sup_{b\in\mathcal{B},\ T\ge 0} \left\{ \int_{[0,T]} l(y_x(s), a^j, \alpha_x^j[b](s), b(s)) \right\}.$$

Thus $U^j(x) \ge V_\gamma^j(x)$. By Theorem 3.6, we know that V_γ is a viscosity supersolution of the SQVI (3.18) if it is continuous. (Note that the proof of the viscosity-supersolution property of V_γ in Theorem 3.6 does not use the assumption (A6).) Also, V_γ has the property (iii) by Proposition 3.1. Thus we conclude that, if continuous, V_γ is the minimal, nonnegative continuous viscosity supersolution of SQVI (3.18) which satisfies condition (iii)

The first part of Theorem 3.9, already proved, then implies that V_γ is a switching-storage function. Moreover, if S is any continuous, switching-storage function for some nonanticipating strategy α_{V_γ}, from Theorem 3.7 we see that S is a viscosity supersolution of the SQVI (3.18). Again, from the first part of this theorem, already proved, we then see that $S \ge V_\gamma$ if S has the property (iii), and hence V_γ is also the minimal continuous switching-storage function satisfying the condition (iii), as asserted. □

Remark 1. The proof of Theorem 3.9 required deduction of an integral inequality (3.43) from knowledge of an inequality of the form $H(y(x), DU(y(x)) \geq 0$ holding in the viscosity sense; the proof of this fact from [6, section II.5.5] ultimately uses the boundedness of the disturbance set B. However, the paper [13] obtains such an integral inequality without a boundedness assumption by using other tools of nonsmooth analysis (e.g., "contingent epiderivative" and viability theory). By using this alternative nonsmooth framework rather than restricting oneself to "viscosity-sense supersolutions," one can get a version of Theorem 3.9 which does not rely on the boundedness of B.

Remark 2. The results of this section reduce the computation of robust state-feedback switching strategy α to computing the solution $(U = U^1, \ldots, U^r)$ (or, more precisely, the minimal viscosity supersolution) of the SQVI of the form

(SQVI)
$$\max\left\{H^j(x, DU^j(x)),\, U^j(x) - \min_{i \neq j}\{U^i(x) + k(a^j, a^i)\}\right\} = 0, \quad j = 1, \ldots, r.$$

This leaves open the issue of how one computes such a solution of an SQVI. A connection can be made with the easier problem of solving a single variational inequality as follows.

If $U = (U^1, \ldots, U^r)$ (with $U^j \in C(\mathbb{R}^N)$ for $j = 1, \ldots, r$) is the minimal viscosity supersolution of (SQVI), then each U^j can be interpreted as the minimal viscosity supersolution of the variational inequality (VI)

$$\max\{H(x, DU(x)), U(x) - \Phi(x)\} = 0 \tag{VI}$$

with Hamiltonian H equal to H^j and with stopping cost Φ equal to $\Phi^j = \min_{i \neq j}\{U^i + k(a^j, a^i)\}$. This suggests defining an iteration map F as follows. Given an r-tuple $U = (U^1, \ldots, U^r)$ of nonnegative real-valued functions, define a new r-tuple $F(U) = (F(U)^1, \ldots, F(U)^r)$ of nonnegative real-valued functions by

$$F(U)^j = \text{ the minimal viscosity supersolution of (VI) with } H = H^j \text{ and } \Phi = \Phi^j.$$

Note that U is the minimal viscosity supersolution of (SQVI) if and only if $F(U) = U$, i.e., if and only if U is a fixed point of F. Formally, one can solve the fixed point problem by guessing a starting point $U_0 = (U_0^1, \ldots, U_0^r)$ and then iterating

$$U_{n+1} = F(U_n),\ n = 0, 1, 2, \ldots, .$$

If $U_n \to U_\infty$ and F is continuous, then, from $U_{n+1} = F(U_n)$, one can take the limit to get $U_\infty = F(U_\infty)$, from which we see that U_∞ is a fixed point for F. For finite horizon problems or problems with a positive discount factor in the running cost, the connection is a little cleaner; in this situation, one has a uniqueness theorem for solutions of the relevant SQVI.

A similar remark giving a connection between the impulsive control problem and the stopping time problem is given in [6, Chapter III, section 4.3], where some convergence results are also given. It would be of interest to develop similar convergence results for the SQVI associated with an optimal switching-control problem.

4. Stability for switching-control problems. In this section, we show how the solution of the SQVI (3.18) can be used for stability analysis.

We consider the system (1.1)–(1.2) with some control strategy α plugged in to get a closed-loop system with the disturbance signal as the only input:

$$\Sigma_{sw}\left\{\begin{array}{rcl}\dot{y} &=& f(y,\alpha_x^j[b],b),\ y(0)=x,\ a(0^-)=a^j,\\ z &=& h(y,\alpha_x^j[b],b).\end{array}\right.$$

An example of such a strategy α is the canonical strategy α_U (see (3.36) or (3.37)) determined by a continuous supersolution of the SQVI (3.18). Moreover, if $V_\gamma = (V_\gamma^1,\dots,V_\gamma^r)$ is the vector lower-value function for the associated game as in (1.4) and we assume that 0 is an equilibrium point for the autonomous system formed from (1.1)–(1.2) by taking $a(s)=a^{i_0}$ and $b(s)=0$ (so $f(0,a^{i_0},0)=0$ and $h(0,a^{i_0},0)=0$), then it is easy to check that $V_\gamma^{i_0}(0)=0$. Furthermore, the associated strategy $\alpha=\alpha_{V_\gamma}$ has the property that

$$\alpha_0^{i_0}[0]=a^{i_0}, \tag{4.1}$$

so 0 is an equilibrium point of the closed-loop system Σ_{sw} with $\alpha=\alpha_{V_\gamma}$ and $a(0^-)=a^{i_0}$ as well. Our goal is to give conditions which guarantee a sort of converse, starting with any continuous supersolution U of the SQVI (3.18).

We first need a few preliminaries. The following elementary result can be found, e.g., in [19].

LEMMA 4.1. *If $\phi(\cdot):\mathbb{R}\to\mathbb{R}$ is a nonnegative, uniformly continuous function such that $\int_0^\infty \phi(s)\,ds<\infty$, then $\lim_{t\to\infty}\phi(t)=0$.*

We say that the closed-loop switching system Σ_{sw} is *zero-state observable* for initial control setting a^j if whenever $h(y_x(t),\alpha_x^j[0](t),0)=0$ for all $t\ge 0$, then $y_x(t)=y_x(t,\alpha_x^j[0],0)=0$ for all $t\ge 0$. We say that the closed-loop system Σ_{sw} is *zero-state detectable* for initial control setting a^j if

$$\lim_{t\to\infty} h(y_x(t),\alpha_x^j[0](t),0)=0 \text{ implies that } \lim_{t\to\infty} y_x(t,\alpha_x^j[0],0)=0.$$

The following proposition gives conditions which guarantee that a particular component U^j of a viscosity supersolution $U=(U^1,\dots,U^r)$ is positive-definite, a conclusion which will be needed as a hypothesis in the stability theorem to follow.

PROPOSITION 4.2. *Assume the following:*

(i) (A1)–(A6) *hold;*

(ii) Σ_{sw} *is zero-state observable for some initial control setting a^j;*

(iii) $U=(U^1,\dots,U^r)$ *is a nonnegative continuous viscosity supersolution of the SQVI* (3.18)

$$\max\left\{H^j(x,DU^j(x)),U^j(x)-\min_{i\ne j}\{U^i(x)+k(a^j,a^i)\}\right\}=0,\quad x\in\mathbb{R}^N,\ j=1,\dots,r;$$

(iv) $U^j(x)\le \min_{i\ne j}\{U^i(x)+k(a^j,a^i)\},\ x\in\mathbb{R}^N,\ j=1,\dots,r.$

Then $U^j(x)>0$ for $x\ne 0$.

Proof. Let $x\in\mathbb{R}^N$. By Theorem 3.9, U is a storage function for Σ_{sw} if we use $\alpha=\alpha_U$ given by (3.36) or, equivalently, (3.37). Thus

$$\begin{aligned}U^j(x)&\ge\int_{[0,T)} l(y_x(s),a^j,\alpha_{U,x}^j[0](s),0)\,ds+U^{j(T)}(y_x(T,\alpha_{U,x}^j[0],0))\\ &\ge\int_{[0,T)} l(y_x(s),a^j,\alpha_{U,x}^j[0](s),0)\,ds\ \text{ for all } T>0.\end{aligned}$$

Since k is nonnegative, we have

$$U^j(x) \geq \int_0^T h(y_x(s), \alpha_x^j[0](s), 0)\, ds \quad \text{for all } T \geq 0.$$

Thus, if $U^j(x) = 0$, then $h(y_x(s, \alpha_x^j[0], 0), \alpha_x^j[0](s), 0) = 0$ for all $s \geq 0$ because h is nonnegative by assumption (A6). Since Σ_{sw} is zero-state observable for initial control setting a^j, it follows that $y_x(s, \alpha_x^j[0], 0) = 0$ for all $s \geq 0$. Thus $x = y_x(0, \alpha_x[0], 0) = 0$. Since U^j is nonnegative, we conclude that, if $x \neq 0$, then $U^j(x) > 0$. ∎

PROPOSITION 4.3. *Assume the following:*

(i) (A1)–(A6) *hold;*

(ii) $U = (U^1, \dots, U^r)$ *is a nonnegative continuous viscosity supersolution of the SQVI* (3.18)

$$\max\left\{H^j(x, DU^j(x)), U^j(x) - \min_{i \neq j}\{U^i(x) + k(a^j, a^i)\}\right\} = 0, \quad x \in \mathbb{R}^N,\ j = 1, \dots, r;$$

(iii) $U^j(x) \leq \min_{i \neq j}\{U^i(x) + k(a^j, a^i)\},\ x \in \mathbb{R}^N,\ j = 1, \dots, r;$

(iv) *there is an* $i_0 \in \{1, \dots, r\}$ *such that* $U^{i_0}(0) = 0$ *and* $U^{i_0}(x) > 0$ *for* $x \neq 0$.

(v) Σ_{sw} *is zero-state detectable for all initial control settings* $a^j \in A$.

Then the strategy α_U *associated with* U *as in* (3.36) *or* (3.37) *is such that* $\alpha_{U,0}^{i_0}[0](s) = a^{i_0}$ *for all* s *and* 0 *is an equilibrium point for the system* $\dot{y} = f(y, a_0^i, 0)$. *Moreover,* 0 *is a globally asymptotically stable equilibrium point for the system* Σ_{sw} *in the sense that the solution* $y(t) = y_x^j(t, \alpha_{U,x}^j[0], 0)$ *of*

$$\dot{y} = f(y, \alpha_{U,x}^j[0], 0), \quad y(0) = x,$$

has the property that

$$\lim_{t \to \infty} y_x^j(t, \alpha_{U,x}^j[0], 0) = 0$$

for all $x \in \mathbb{R}^N$ *and all* $a^j \in A$.

Proof. Suppose that $U^{i_0}(0) = 0$ and $U^{i_0}(x) > 0$ for $x \neq 0$. Let $T \geq 0$ and $x \in \mathbb{R}^N$. Since U is a storage function for the closed-loop system formed from (1.1)–(1.2) with $\alpha = \alpha_U$, we have

$$U^{i_0}(x) \geq \int_0^T h(y_x(s), \alpha_x^{i_0}[0](s), 0)\, ds + \sum_{\tau < T} k(\alpha_{U,x}^{i_0}(\tau^-), \alpha_{U,x}^{i_0}(\tau)) + U^{j(T)}(y_x(T, \alpha_{U,x}^{i_0}[0], 0)). \tag{4.2}$$

Since $h,\ k,\ U$ are nonnegative and $U^{i_0}(0) = 0$ by our assumptions, substitution of $x = 0$ in (4.2) forces

$$\sum_{\tau < T} k(\alpha_{U,0}^{i_0}[0](\tau^-), \alpha_{U,0}^{i_0}[0](\tau)) = 0.$$

This implies that $\alpha_{U,0}^{i_0}[0](t) = a^{i_0}$ for all $0 \leq t \leq T$. Thus

$$0 \leq U^{j(T)}(y_0(T, \alpha_{U,0}^{i_0}[0], 0)) = U^{i_0}(y_0(T, \alpha_{U,0}^{i_0}[0], 0)) \leq U^{i_0}(0) = 0.$$

By the positive definite property of U^{i_0}, we have $y_0(T, \alpha^{i_0}_{U,0}[0], 0) = 0$. Since $T \geq 0$ is arbitrary, we conclude that 0 is an equilibrium point of the system $\dot{y} = f(y, a^{i_0}, 0)$.

Next we want to show that 0 is a globally asymptotically stable equilibrium point for the closed-loop switching system Σ_{sw} with $\alpha = \alpha_U$. Again, from the storage-function property of $U = (U^1, \ldots, U^r)$ for the system Σ_{sw} with $\alpha = \alpha_U$, we have

$$\int_0^T h(y_x(s), \alpha^j_{U,x}[0](s), 0)\, ds \leq U^j(x) < \infty \text{ for all } T > 0.$$

Thus $\lim_{t\to\infty} h(y_x(t), \alpha^j_{U,x}[0], 0) = 0$ by Lemma 4.1. By the detectability assumption (v), we have $\lim_{t\to\infty} y_x(t, \alpha^j_{U,x}[0], 0) = 0$ as required. □

5. An example. We consider in this section an optimal switching problem with one-dimensional state space ($x \in \mathbb{R}^1$) for which the value function and corresponding control are explicitly computable via a simple geometric construction. There will be two controls: $a \in \{1, 2\}$. The switching cost will be symmetric: $k(1, 2) = k(2, 1) = \beta > 0$. For each a value, we will use $a' = 3 - a$ to denote the other control value. The system dynamics will be given by

$$f(y, 1, b) = -y + b; \quad f(y, 2, b) = -\mu(y - 1) + b, \tag{5.1}$$

with output function h taken simply to be the squared state

$$h(y, a, b) = y^2. \tag{5.2}$$

We use the specific parameter values

$$\mu = 3, \quad \beta = .4, \quad \gamma = 2$$

throughout.

This example satisfies all of our hypotheses except that we take $B = \mathbb{R}$, which is not compact. Our purpose is to make the SQVI (2.5) more tangible in the context of the example, to show how the optimal strategy α_* is determined, and to show how one might establish its optimality. We note that, even apart from the fact that our B is not compact, Theorem 3.9 would not by itself imply that our solution V^a of the SQVI is that value of the game (i.e., that α_* is optimal); an additional argument to establish the minimality of V^a among nonnegative solutions is also necessary. Instead we will outline a direct proof of the optimality of α_*. Since we are not appealing to any of the theorems above, the fact that the compactness hypothesis on B is not satisfied does not pose a problem. (It would be possible to modify the example so that $B = [-M, M]$ could be used for some sufficiently large M. For instance, if the $f(y, a, b)$ were bounded (nonlinear) functions of y, this would be possible. However, the linear-affine dynamics of (5.1) are simpler to work with, so we have kept them.)

Our task is to construct the appropriate solution of the SQVI (2.5). With just two control values a, the SQVI reduces to the following: for each a,

$$V^a(x) \leq \beta + V^{a'}(x) \text{ for all } x, \tag{5.3}$$

$$H^a(x, D^+V^a(x)) \leq 0 \text{ for all } x, \tag{5.4}$$

$$H^a(x, D^-V^a(x)) \geq 0 \text{ for those } x \text{ with } V^a(x) < \beta + V^{a'}(x). \tag{5.5}$$

Here we have used the standard notation $D^+V^a(x)$ to refer to the set of all possible slopes $\varphi'(x)$ of smooth test functions φ for which $V^a - \varphi$ has a local maximum at x,

usually called the *superdifferential* of V^a at x. Similarly, the *subdifferential* $D^-V^a(x)$ denotes the set of all $\varphi'(x)$ for smooth test functions φ such that $V^a(x) - \varphi$ has a local minimum at x. (See [6, page 29].) At points x where both V^1 and V^2 are differentiable, (5.3)–(5.5) can be expressed more explicitly as

$$|V^1(x) - V^2(x)| \le \beta,$$

together with the following.

1. If $V^1(x) - V^2(x) = \beta$, then $(V^1)'(x) = (V^2)'(x) =: q(x)$ (since $V^1 - V^2$ has a maximum at x), and

$$H^1(x, q(x)) \le 0, \quad H^2(x, q(x)) = 0.$$

2. If $V^1(x) - V^2(x) = -\beta$, then, similarly, $(V^1)'(x) = (V^2)'(x) =: q(x)$ and

$$H^1(x, q(x)) = 0, \quad H^2(x, q(x)) \le 0.$$

3. If $|V^1(x) - V^2(x)| < \beta$, then

$$H^1(x, (V^1)'(x)) = 0, \quad H^2(x, (V^2)'(x)) = 0.$$

Where one or the other of V^a is not differentiable, we must revert to (5.3)–(5.5). However, one of the cases 1–3 above will apply at most x.

The two Hamiltonian functions are

$$H^1(x,p) = px - x^2 - \frac{1}{4\gamma^2}p^2,$$
$$H^2(x,p) = \mu p(1-x) - x^2 - \frac{1}{4\gamma^2}p^2.$$

These are both instances of the general formula

$$\begin{aligned} H(x,p) &= \inf_b\{-(g(x)+b)\cdot p - x^2 + \gamma^2 b^2\} \\ &= -pg(x) - x^2 - \frac{1}{4\gamma^2}p^2 \\ &= (\gamma g(x))^2 - x^2 - \left(\frac{1}{2\gamma}p + \gamma g(x)\right)^2, \end{aligned} \tag{5.6}$$

where $g(x) = -x$ for $a = 1$ and $g(x) = -\mu(x-1)$ for $a = 2$. We are interested in $V'(x) = p(x)$ solving $H(x, p(x)) = 0$. Provided $|x| < \gamma|g(x)|$, the equation $H(x,p) = 0$ has two distinct real solutions:

$$-2\gamma^2 g(x) \pm 2\gamma\sqrt{\gamma^2 g(x)^2 - x^2}.$$

For each a, we need to select an appropriate branch $p^a(x)$ of the solution to $H^a(x,p) = 0$. For $a = 1$, we take

$$p^1(x) = 2\rho x,$$

where $\rho = \gamma^2 - \gamma\sqrt{\gamma^2 - 1}$. Note that, for $x < 0$, $p^1(x)$ is the larger of the two solutions of $H^1(x,p) = 0$, and so

$$H^1(x,p) \le 0 \text{ for all } p \ge p^1(x),\ x < 0. \tag{5.7}$$

For $x > 0$, however, $p^1(x)$ is the smaller of the two solutions, and so

$$H^1(x,p) \le 0 \text{ for all } p \le p^1(x),\ x > 0. \tag{5.8}$$

For $p^2(x)$, we note that $H^2(x,p) = 0$ has no solution for $\frac{6}{7} < x < \frac{6}{5}$ because, with $g(x) = -\mu(x-1)$, $|x| > \gamma|g(x)|$ there. We take

$$p^2(x) = \begin{cases} 2\gamma^2\mu(x-1) - 2\gamma\sqrt{\gamma^2\mu^2(x-1)^2 - x^2} & \text{for } x \ge \frac{6}{5}, \\ 2\gamma^2\mu(x-1) + 2\gamma\sqrt{\gamma^2\mu^2(x-1)^2 - x^2} & \text{for } x \le \frac{6}{7}. \end{cases}$$

Let

$$W^1 = \rho x^2$$

and

$$W^2(x) = \int p^2(x)\,dx.$$

This determines W^2 up to two constants, one each for $(-\infty, 6/7)$ and $(6/5, \infty)$. Those constants are determined uniquely by

$$W^2(x) = \beta + W^1(x) \text{ for } x = 0, 3/2. \tag{5.9}$$

These are solutions of $H^a(x, DW^a(x)) = 0$. The desired solutions of the SQVI are given by

$$V^2(x) = \begin{cases} W^2(x) & \text{for } x < 0, \\ \beta + W^1(x) & \text{for } 0 \le x \le x_1, \\ W^2(x) & \text{for } x_1 < x \end{cases} \tag{5.10}$$

and

$$V^1(x) = \begin{cases} \beta + W^2(x) & \text{for } x \le x_2, \\ W^1(x) & \text{for } x_2 < x < x_3, \\ \beta + W^2(x) & \text{for } x_3 \le x, \end{cases} \tag{5.11}$$

where $x_2 = -1.3175\ldots$, $x_1 = 3/2$, $x_3 = 2.55389\ldots$, values whose significance will emerge below. Graphs are presented in Figure 5.1.

We now outline the verification that V^a as defined above do satisfy the SQVI, leaving many of the details to the interested reader. First, consider $0 < x < x_1$. Here $V^{1\,\prime}(x) = p^1(x)$ so that $H^1(x, V^{1\,\prime}(x)) = 0$. Since $V^2 = \beta + V^1$, we also have $V^{2\,\prime}(x) = p^1(x)$. Case 2 above requires $H^2(x, p^1(x)) \le 0$, which is true up to $x = 3/2 = x_1$ but not beyond. For $x_1 < x < x_3$, we have $H^a(x, V^{a\,\prime}(x)) = H^a(x, p^a(x)) = 0$ for both a, so all of the necessary derivative conditions are satisfied. Note that (5.9) ensures that V^2 is continuous at x_1. Moreover, $H^2(x, p^1(x)) = 0$ at $x = x_1$ because $DW^1(x_1) = p^1(x_1) = p^2(x_1) = DW^2(x_1)$ there. This means V^2 is C^1 at x_1. We have $V^1(x) - \beta < V^2(x) < \beta + V^1(x)$ for $x_1 < x < x_3$, but at x_3 we find $V^1(x_3) - \beta = V^2(x_3)$. (This determines the value of x_3.) Next consider $x > x_3$. Here $V^{2\,\prime} = p^2(x)$ so that $H^2(x, V^{2\,\prime}(x)) = 0$. Since $V^1(x) = \beta + V^2(x)$, $V^{1\,\prime}(x) = V^{2\,\prime}(x) = p^2(x)$ and otherwise case 1 requires only that $H^1(x, p^2(x)) \le 0$, which does hold. Note that the choice of x_3 makes V^1 continuous at x_3, but it is not differentiable there. One finds

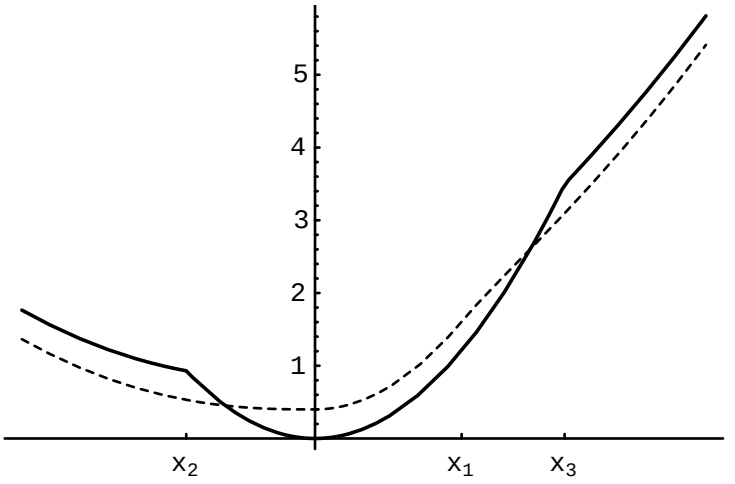


FIG. 5.1. V^1 *(solid) and* V^2 *(dashed).*

that $D^-V^1(x_3) = \emptyset$ and $D^+V^1(x_3)$ is the interval $[p^2(x_3), p^1(x_3)]$. By virtue of (5.8), the viscosity solution requirement $H^1(x_3, D^+V^1(x_3)) \le 0$ is satisfied.

Next consider $x_2 < x < 0$. Here again we have $V^{a\,\prime}(x) = p^a(x)$ so that, for both a,

$$H^a(x, p^a(x)) = 0.$$

We also have $V^1(x) - \beta < V^2(x) < \beta + V^1(x)$, but at x_2 we find $V^1(x_2) - \beta = V^2(x_2)$. This determines x_2 and makes V^1 continuous at x_2. For $x < x_2$, we have $V^{1\,\prime}(x) = V^{2\,\prime}(x) = p^2(x)$ so that case 1 above requires $H^2(x, p^2(x)) = 0$ and $H^1(x, p^2(x)) \le 0$, both of which are true for $x < x_2$. Note that V^1 is not differentiable at x_2. One finds that $D^-V^1(x_2) = \emptyset$ and $D^+V^1(x_2)$ is the interval $[p^1(x_2), p^2(x_1)]$. By virtue of (5.7), the viscosity solution requirement $H^1(x_2, D^+V^1(x_2)) \le 0$ is satisfied.

The strategy α^* associated with our solution (5.10), (5.11) is easy to describe in state-feedback terms. Define the *switching sets*

$$S_1 = \{x : \ V^2(x) = \beta + V^1(x)\} = [0, x_1],$$
$$S_2 = \{x : \ V^1(x) = \beta + V^2(x)\} = (-\infty, x_2] \cup [x_3, \infty).$$

The strategy α^* will instantly switch from $a = 1$ to $a = 2$ whenever $y(t) \in S_2$ and will instantly switch from $a = 2$ to $a = 1$ whenever $y(t) \in S_1$. We will prove directly that, in fact, $V^a_\gamma = V^a$ and that our strategy α^* is optimal. To be precise, we shall show that, for any j and any strategy $\alpha \in \Gamma$,

(5.12)
$$V^j(y(0)) \le \sup_{b \in \mathcal{B}} \sup_{T > 0} \left\{ \int_0^T [h(y_x(s), \alpha^j_x[b](s), b(s)) - \gamma^2 |b(s)|^2]\, ds + \sum_{\tau_i \le T} k(a_{i-i}, a_i) \right\}.$$

Moreover, for our strategy α^*, (5.12) will be an equality for all x, j. The key to this is the existence of a particular "worst case" disturbance, as claimed by the following proposition. (This proposition is intended only in the context of the particular example and the parameter values described above.)

PROPOSITION 5.1. *For any $x \in \mathbb{R}^N$, $j \in \{1,2\}$, and strategy $\alpha \in \Gamma$, there exists a disturbance $b^* = b^*_{\alpha_x^j} \in \mathcal{B}$ with the property that*

$$b^*(t) = \frac{1}{2\gamma^2}(V^{\alpha_x^j[b^*])(t)})'(y_x(t, \alpha_x^j[b], b))$$

holds for all but finitely many t in every interval $[0,T]$.

A proof can be based on the obvious construction. Given $\alpha \in \Gamma$, an initial control j, and an initial point $x \in \mathbb{R}^N$, consider the solution of

$$\dot{y} = f\left(y, j, \frac{1}{\gamma^2}(V^j)'(y)\right); \quad y(0) = x. \tag{5.13}$$

We simply take $b^*(t) = (V^j)'(y(t))$ up until the first time τ_1 that the policy $\alpha_x^j[b^*]$ calls for a switch from $a = j$ to $a = j'$. For $t > \tau_1$, we continue by solving

$$\dot{y} = f\left(y, j', \frac{1}{\gamma^2}(V^{j'})'(y)\right)$$

with initial value $y(\tau_1)$ as already determined. We take $b^*(t) = (V^{j'})'(y(t))$ for $\tau_1 < t$ up until the next time τ_2 that $\alpha_x^j[b^*]$ calls for a switch from $a = j'$ to $a = j$. We continue this construction iteratively.

A number of observations are needed to justify the construction. One is the existence of a unique solution to (5.13). For $j = 2$, the right side is C^1, so the solution is uniquely determined. For $j = 1$, the right side has discontinuities at x_2 and x_3, but, since $f(x, 1, \frac{1}{\gamma^2}(V^1)'(x))$ does not change sign across the discontinuities, the solution *is* uniquely determined. Graphs of $f(y, a, \frac{1}{\gamma^2}(V^a)'(y))$ are provided in Figures 5.2 and 5.3 below. (We comment that, although the graphs appear piecewise linear, they are not. Figure 5.2 is linear only for $0 < x < x_1$, and Figure 5.3 is linear only for $x_2 < x < x_3$, as inspection of the formulas shows.) One can check that $y\dot{y} < 0$ for sufficiently large $|y|$, which implies that solutions of (5.13) are defined for all $t \geq 0$. Observe also for $j = 1$ that, for any solution of (5.13), there is at most one value of t for which $y(t)$ is at one of the discontinuities of $(V^1)'$ and for which there is any ambiguity in the specification $b^*(t) = (V^j)'(y(t))$.

The other concern is that the sequence τ_i of switching times generated by our construction might have a finite accumulation point: $\lim \tau_i = s < \infty$. Our hypotheses on the strategy α disallow this, however, for the following reason. If it were the case that $\lim \tau_i = s < \infty$, then extend our definition of b^* in any way to $t \geq s$, say, $b^*(t) = 0$. By hypothesis, $\alpha_x^j[b^*]$ is an admissible control in $\mathcal{A}$, which means, in particular, that its switching times τ_i do not have a finite accumulation point. However, extension of b^* for $t > s$ does not alter the switching times $\tau_i < s$ by the nonanticipating property of α. This would mean that $\alpha[b^*]$ does have an infinite number of switching times $\tau_i < s$, which is a contradiction. Finally, by our comments above, on each interval $[\tau_i, \tau_{i+1}]$, there is at most a single t value at which $b^*(t) = (V^{\alpha_x^j[b^*]})'(y(t))$ is ambiguous. Thus there are at most a finite number of such t in any $[0,T]$.

Consider now any strategy $\alpha \in \Gamma$, initial position $x = y(0)$, and control setting j, and let $b^*(t)$ be the disturbance described in the proposition. We will let $a_i = \alpha_x^j[b^*](t)$ denote the control settings on the intervals $[\tau_i, \tau_{i+1}]$ between consecutive switching times. In particular, $a_0 = j$. On each interval $[\tau_i, \tau_{i+1}]$, (5.4) and the fact that $b^*(t)$ achieves the infimum in (5.6) for $x = y(t)$ and $p = (V^{a_i})'(x)$ imply that (for all but

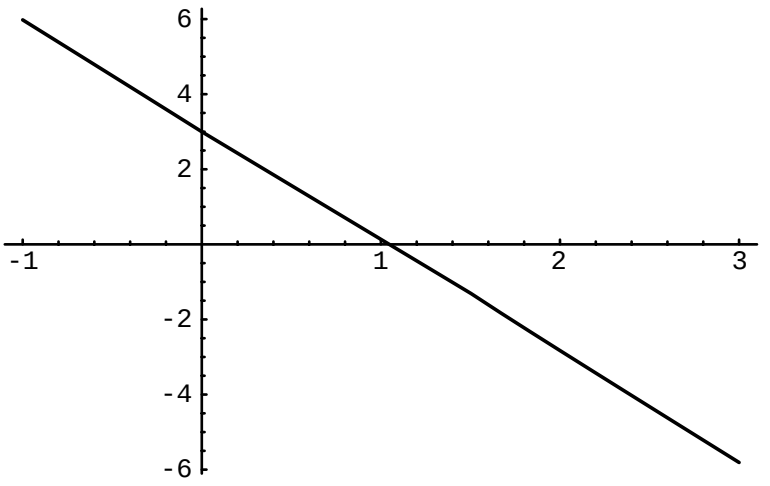


FIG. 5.2. *Plot of* $f(x, 2, \frac{1}{2\gamma^2} V^{2\,\prime}(x))$.

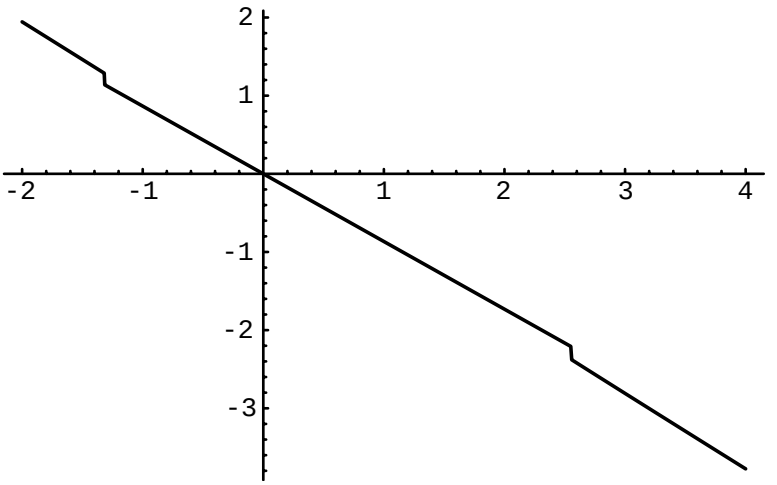


FIG. 5.3. *Plot of* $f(x, 1, \frac{1}{2\gamma^2} V^{1\,\prime}(x))$.

one t)

$$\frac{d}{dt} V^{a_i}(y(t)) \geq \gamma^2 b^*(t)^2 - h(y(t), a_i, b^*(t)).$$

Thus, for any $\tau_i < t \leq \tau_{i+1}$, we have

$$V^{a_i}(y(t)) - V^{a_i}(y(\tau_i)) \geq \int_{\tau_i}^{t} [\gamma^2 |b^*(s)|^2 - h(s)]\, ds.$$

Across a switching time τ_i, we have from (5.3)

$$V^{a_i}(y(\tau_i)) - V^{a_{i-1}}(y(\tau_i)) \geq -\beta = -k(a_{i-1}, a_i).$$

Adding these inequalities over $\tau_i \le T$, we see that

$$V^{\alpha[b^*](T)}(y(T)) - V^{\alpha[b^*](0)}(y(0)) \ge -\left\{ \int_0^T [h(s) - \gamma^2 |b^*(s)|^2]\, ds + \sum_{\tau_i \le T} k(a_{i-1}, a_i) \right\}.$$

A rearrangement of this gives

(5.14)
$$V^{\alpha[b^*](T)}(y(T)) + \left\{ \int_0^T [h(s) - \gamma^2 |b^*(s)|^2]\, ds + \sum_{\tau_i \le T} k(a_{i-1}, a_i) \right\} \ge V^{\alpha[b^*](0)}(y(0)).$$

When we consider α^* specifically, we recognize that

$$H^{a_i}(y(t), (V^{a_i})'(y(t))) = 0$$

for t between two τ_i's, and at τ_i we have

$$V^{a_{i+1}}(y(\tau_i)) - V^{a_i}(y(\tau_i)) = -\beta = -k(a_{i+1}, a_i).$$

This means that (5.14) is an equality for α^* specifically.

To finish our optimality argument, we will show that, for α, a general strategy, initial condition (x, a^j), and associated disturbance $b^* = b^*_{\alpha_x^j}$ as above, as $T \to \infty$ we must have either $y(T) \to 0$ and $\alpha[b^*](T) \to 1$, or else

$$\int_0^T [h(s) - \gamma^2 |b^*(s)|^2]\, ds + \sum_{\tau_i \le T} k(a_{i-1}, a_i) \to +\infty. \tag{5.15}$$

In the case of $\alpha = \alpha^*$ specifically, we will have the former possibility. Since $V^1(0) = 0$ and is continuous, these facts imply (5.12) as claimed. The verification of these asserted limiting properties for the case of general α depends on some particular inequalities for $(V^a)'(x)$ as determined by (5.11), (5.10). First, we assert that, for both a values,

$$h(y(t), a, b^*(t)) - \gamma^2 |b^*(t)|^2 = |y(t)|^2 - \frac{1}{4\gamma^2}[(V^a)'(y(t))]^2 > 0 \text{ for } x \ne 0. \tag{5.16}$$

Moreover, $|x|^2 - \frac{1}{4\gamma^2}[(V^a)'(x)^2$ has a positive lower bound on $\{x : |x| \ge \epsilon\}$ for each $\epsilon > 0$. Instead of what would be a very tedious algebraic demonstration of this, we simply offer the graphical demonstration in Figure 5.4. We have plotted $b^* = \frac{1}{2\gamma}(V^a)'(x)$ (solid lines) and $q = x$ (dashed lines) as functions of x. The validity of (5.16) is apparent.

The other fact we need is that, for $a = 2$ and the corresponding disturbance $b^*(t)$, the state-dynamics do not have an equilibrium at 0. This is easy to see because at $x = 0$ we have $b^* = \frac{1}{2\gamma^2}(V^2)'(0) = 0$, but $f(0, 2, b^*) = -\mu + b^*$. A graph of $f(x, 2, b^*) = -\mu(x - 1) + \frac{1}{2\gamma^2}(V^2)'(x)$ is provided in Figure 5.2, where we see the unique equilibrium just beyond $x = 1$.

In the case of $a = 1$, however, $\dot{x} = f(x, 1, \frac{1}{2\gamma^2}(V^1)'(x))$ has a unique globally asymptotically stable equilibrium at $x = 0$, as is evident in Figure 5.3.

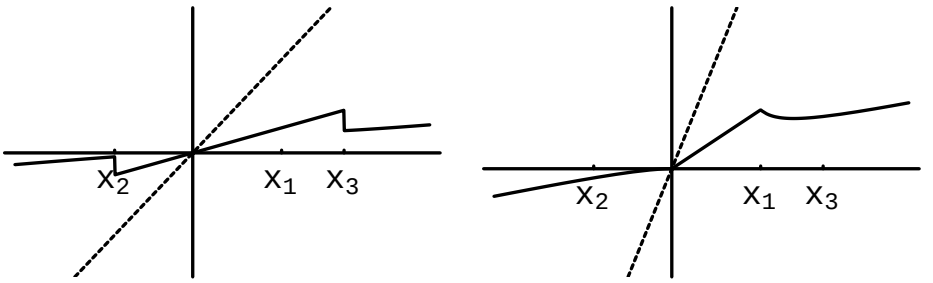


FIG. 5.4. *Graphical verification of* (5.16) *for* $V^{1\prime}$ *(left) and* $V^{2\prime}$ *(right).*

We turn then to the verification of the assertion of (5.15) or its alternative: assuming (5.15) to be false, we claim that $y(T) \to 0$ and $\alpha[b^*](T) \to 1$. By the nonnegativity from (5.16), we must have both

$$\sum_{\tau_i < \infty} k(a_{i-1}, a_i) < \infty \text{ and } \int_0^\infty [h(y(s)) - \gamma^2 |b^*(s)|^2]\, ds < \infty. \tag{5.17}$$

The first of these implies that there are only a finite number of switches; $\alpha[b^*](t) = a^{i^*}$ is constant from some time on. It is not possible that $i^* = 2$ because, in that case, $y(t)$ would be converging to the positive equilibrium of Figure 5.2, which implies by (5.16) that, as $t \to \infty$,

$$h(y(t), a^{i^*}, b^*(t)) - \gamma^* |b^*(t)|^2 \to C > 0.$$

This contradicts the second part of (5.17). Therefore, $i^* = 1$, which shows that $\alpha[b^*](T) \to 1$. However, since $\alpha[b^*](t) = 1$ from some point on, the stability illustrated in Figure 5.3 means that $y(t) \to 0$ as claimed. This completes our verification of the optimality of the strategy α^*.

REFERENCES

[1] J. A. BALL AND J. CHUDOUNG, *Comparison theorems for viscosity solutions of systems of quasivariational inequalities with applications to switching-cost control problems*, J. Math. Anal. Appl., 251 (2000), pp. 40–64.

[2] J. A. BALL, J. CHUDOUNG, AND M. V. DAY, *Robust optimal stopping-time control for nonlinear systems*, Appl. Math. Optim., 29 (2002).

[3] J. A. BALL, M. V. DAY, T. YU, AND P. KACHROO, *Robust L_2-gain control for nonlinear systems with projection dynamics and input constraints: An example from traffic control*, Automatica J. IFAC, 35 (1999), pp. 429–444.

[4] J. A. BALL, M. V. DAY, AND P. KACHROO, *Robust feedback control of a single server queueing system*, Math. Control Signals Systems, 12 (1999), pp. 307–345.

[5] J. A. BALL AND J. W. HELTON, *Viscosity solutions of Hamilton-Jacobi equations arising in nonlinear H_∞ control*, J. Math. Systems Estim. Control, 6 (1996), pp. 1–22.

[6] M. BARDI AND I. CAPPUZZO-DOLCETTA, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser Boston, Boston, 1997.

[7] E. N. BARRON, R. JENSEN, AND J. L. MENALDI, *Optimal control and differential games with measures*, Nonlinear Anal., 21 (1993), pp. 241–268.

[8] T. BAŞAR AND P. BERNHARD, *H^∞-Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*, 2nd ed., Birkhäuser Boston, Boston, 1995.

[9] A. BENSOUSSAN AND J. L. LIONS, *Applications of Variational Inequalities to Stochastic Control*, North–Holland, New York, 1982.

[10] M. S. BRANICKY, V. S. BORKAR, AND S. K. MITTER, *A unified framework for hybrid control: Model and optimal control theory*, IEEE Trans. Automat. Control, 43 (1998), pp. 31–45.

[11] M.S. Branicky, *Multiple Lyapunov functions and other analysis tools for switched and hybrid systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 475–482.
[12] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.
[13] H. Frankowska and M. Quincampoix, *Dissipative control systems and disturbance attenuation for nonlinear H^∞ problems*, Appl. Math. Optim., 40 (1999), pp. 163–181.
[14] J. W. Helton and M. R. James, *Extending H_∞ Control to Nonlinear Systems: Control of Nonlinear Systems to Achieve Performance Objectives*, SIAM, Philadelphia, 1999.
[15] M. R. James, *A partial differential inequality for dissipative nonlinear systems*, Systems Control Lett., 21 (1993), pp. 315–320.
[16] D. Liberzon and A. S. Morse, *Basic problems in stability and design of switched systems*, Control Systems, 19 (1999), pp. 59–70.
[17] M. Motta and F. Rampazzo, *Space-time trajectories of nonlinear systems driven by ordinary and impulsive control*, Differential Integral Equations, 8 (1995), pp. 269–288.
[18] M. Motta and F. Rampazzo, *Dynamic programming for nonlinear systems driven by ordinary and impulsive controls*, SIAM J. Control Optim., 34 (1996), pp. 199–225.
[19] S. Sastry and M. Bodson, *Adaptive Systems: Stability, Convergence and Robustness*, Prentice–Hall, Englewood Cliffs, NJ, 1989.
[20] A. van der Schaft, *L_2-gain and Passivity Techniques in Nonlinear Control*, Springer-Verlag, New York, 1996.
[21] P. Soravia, *H_∞ control of nonlinear systems: Differential games and viscosity solutions*, SIAM J. Control Optim., 34 (1996), pp. 1071–1097.
[22] P. Soravia, *Equivalence between nonlinear H_∞ control problems and existence of viscosity solutions of Hamilton-Jacobi-Isaacs equations*, Appl. Math. Optim., 39 (1999), pp. 17–32.
[23] J. Yong, *A zero-sum differential game in a finite duration with switching strategies*, SIAM J. Control Optim., 28 (1990), pp. 1234–1250.
[24] J. Yong, *Differential games with switching strategies*, J. Math. Anal. Appl., 145 (1990), pp. 455–469.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 3, pp. 932–953

NONLINEAR OBSERVER DESIGN IN THE SIEGEL DOMAIN*

ARTHUR J. KRENER† AND MINGQING XIAO‡

Abstract. We extend the method of Kazantzis and Kravaris [*Systems Control Lett.*, 34 (1998), pp. 241–247] for the design of an observer to a larger class of nonlinear systems. The extended method is applicable to any real analytic observable nonlinear system. It is based on the solution of a first-order, singular, nonlinear PDE. This solution yields a change of state coordinates which linearizes the error dynamics. Under very general conditions, the existence and uniqueness of the solution is proved. Lyapunov's auxiliary theorem and Siegel's theorem are obtained as corollaries. The technique is constructive and yields a method for constructing approximate solutions.

Key words. nonlinear systems, nonlinear observers, linearizable error dynamics, output injection, Siegel domain, Lyapunov's auxiliary theorem, Siegel's theorem

AMS subject classifications. 93, 35, 32

PII. S0363012900375330

1. Introduction.

We consider the problem of estimating the current state $x(t)$ of a nonlinear dynamical system, described by a system of first-order differential equations,

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

from the past observations $y(s), s \le t$. The vector fields $f : \mathbf{R}^n \to \mathbf{R}^n$ and $h : \mathbf{R}^n \to \mathbf{R}^p$ are assumed to be real analytic functions with $f(0) = 0,\ h(0) = 0$. One technique of constructing an observer is to find a nonlinear change of state and output coordinates which transforms the system (1.1) into a system with linear output map and linear dynamics driven by nonlinear output injection. The design of an observer for such systems is relatively easy [8], [6], [2], and the error dynamics is linear in the transformed coordinates. Recently Kazantzis and Kravaris proposed a simpler method [5]. One seeks a change of state coordinates $z = \theta(x)$ such that the dynamics of (1.1) is linear driven by nonlinear output injection

$$\dot{z} = Az - \beta(y), \tag{1.2}$$

where A is an $n \times n$ matrix and $\beta : \mathbf{R}^p \to \mathbf{R}^n$ is a real analytic vector field. One does not have to linearize the output map.

Such a θ must satisfy the following first-order PDE:

$$\frac{\partial \theta}{\partial x}(x) f(x) = A\theta(x) - \beta(h(x)). \tag{1.3}$$

Using a particular form of the Lyapunov auxiliary theorem [10], Kazantzis and Kravaris showed that (1.3) has a unique solution under certain assumptions.

*Received by the editors July 14, 2000; accepted for publication (in revised form) February 22, 2002; published electronically September 19, 2002. A preliminary version of this paper appeared in the Proceedings of the 2001 IEEE Conference on Decision and Control.

http://www.siam.org/journals/sicon/41-3/37533.html

†Department of Mathematics, University of California, Davis, CA 95616-8633 (ajkrener@ucdavis.edu). Research for this author was supported in part by NSF 9970998.

‡Department of Mathematics, Southern Illinois University, Carbondale, IL 62901-4408 (mxiao@math.siu.edu).

THEOREM [10]. *Assume that* $f : \mathbf{R}^n \to \mathbf{R}^n, h : \mathbf{R}^n \to \mathbf{R}^p$, *and* $\beta : \mathbf{R}^p \to \mathbf{R}^n$ *are analytic vector fields with* $f(0) = 0$, $h(0) = 0$, $\beta(0) = 0$ *and* $F = \frac{\partial f}{\partial x}(0)$, $H = \frac{\partial h}{\partial x}(0)$, $B = \frac{\partial \beta}{\partial x}(0)$. *Let the eigenvalues of* F *be* $(\lambda_1, \ldots, \lambda_n)$ *and the eigenvalues of* A *be* $(\mu_1, \ldots, \mu_n)$. *If*

1. 0 *does not lie in the convex hull of* $(\lambda_1, \ldots, \lambda_n)$,
2. *there do not exist nonnegative integers* $m_1, m_2, \ldots, m_n$ *not all zero such that* $\sum_{i=1}^{n} m_i \lambda_i = \mu_j$,

then the first-order PDE (1.3), *with initial condition* $\theta(0) = 0$, *admits a unique analytic solution* θ *in a neighborhood of* $x = 0$.

Based on the above theorem, Kazantzis and Kravaris proposed a nonlinear observer design method [5], where the state observer is constructed using the coordinate transformation $z = \theta(x)$ and the output injection $\beta(y)$.

KAZANTZIS AND KRAVARIS THEOREM [5]. *Assume that* f, h, θ, β *are as in the above theorem and additionally that*

3. θ *is a local diffeomorphism,*
4. A *is Hurwitz.*

Then the local state observer for (1.1) *given by*

$$\dot{\hat{x}} = f(\hat{x}) - \left[\frac{\partial \theta}{\partial \hat{x}}(\hat{x})\right]^{-1} (\beta(y) - \beta(h(\hat{x}))) \tag{1.4}$$

has locally asymptotically stable error dynamics. In z *coordinates, the system is given by* (1.2), *the observer is*

$$\dot{\hat{z}} = A\hat{z} - \beta(y), \tag{1.5}$$

and the error $\tilde{z} = z - \hat{z}$ *dynamics is*

$$\dot{\tilde{z}} = A\tilde{z}. \tag{1.6}$$

One can show that if the conditions of this theorem hold, then (H, F) is an observable pair and (A, B) is a controllable pair. On the other hand, if (H, F) is an observable pair, then one can choose an invertible T and B so that $A = (TF + BH)T^{-1}$ satisfies 2, 3, and if the solution of (1.3) exists for some β such that $\beta(0) = 0$, $\frac{\partial \beta}{\partial x}(0) = B$, then θ is a local diffeomorphism. The size of the neighborhood of 0 on which θ is a diffeomorphism varies with the higher derivatives of β, hence the advantage of allowing them to be different from zero.

The approach of Kazantzis and Kravaris has an advantage over that of Krener and Respondek [8] and similar attempts to transform the dynamics and output map into observer form. The former uses the Lyapunov auxiliary theorem, which depends on a nonresonance condition, assumption 2 above, while the latter depends on integrability conditions. The nonresonance condition is generically satisfied while the integrability conditions are generically not satisfied. However, assumption 1 of Kazantzis and Kravaris is quite restrictive, as it requires the system to be locally asymptotically stable to the origin in either forward or reverse time. If the system is stable in forward time, then an observer is not needed, as we know where it is going. If the system is stable in reverse time, then it is unstable in forward time, so what good is a local observer?

Assumption 1 requires that the eigenvalues of the linear part of $f(x)$ at the origin lie in the *Poincaré domain*, whose definition follows.

DEFINITION 1. *An n-tuple $\lambda = (\lambda_1, \ldots, \lambda_n)$ of complex numbers belongs to the Poincaré domain if the convex hull of $(\lambda_1, \ldots, \lambda_n)$ does not contain zero. An n-tuple of complex numbers belongs to the Siegel domain if zero lies in the convex hull of $(\lambda_1, \ldots, \lambda_n)$.*

Clearly, requiring the spectrum of F to be in the Poincaré domain rules out many interesting problems, including critical ones where there are eigenvalues on the imaginary axis [9]. In this paper we extend the observer design method of Kazantzis and Kravaris to the Siegel domain [1]. We start with a definition.

DEFINITION 2. *Given an $n \times n$ matrix F with spectrum $\sigma(F) = \lambda = (\lambda_1, \ldots, \lambda_n)$ and constants $C > 0$, $\nu > 0$, we say a complex number μ is of type (C, ν) with respect to $\sigma(F)$ if for any vector $m = (m_1, m_2, \ldots, m_n)$ of nonnegative integers, $|m| = \sum m_i > 0$, we have*

$$|\mu - m \cdot \lambda| \geq \frac{C}{|m|^\nu}. \tag{1.7}$$

Now we are ready to state the main result of this paper.

MAIN THEOREM. *Assume that $f : \mathbf{R}^n \to \mathbf{R}^n$, $h : \mathbf{R}^n \to \mathbf{R}^p$, and $\beta : \mathbf{R}^p \to \mathbf{R}^n$ are analytic vector fields with $f(0) = 0$, $h(0) = 0$, $\beta(0) = 0$ and $F = \frac{\partial f}{\partial x}(0)$, $H = \frac{\partial h}{\partial x}(0)$, $B = \frac{\partial \beta}{\partial y}(0)$. Suppose there exists*

1. *an invertible $n \times n$ matrix T so that $TF = AT - BH$;*
2. *a $C > 0$, $\nu > 0$ such that all the eigenvalues of A are of type (C, ν) with respect to $\sigma(F)$.*

Then there exists a unique analytic solution $z = \theta(x)$ to the PDE (1.3) locally around $x = 0$ with $\frac{\partial \theta}{\partial x}(0) = T$, so θ is a local diffeomorphism.

Notes. We have stated this theorem for real analytic functions because we are applying it to a real analytic system. However, it is true for complex analytic functions, as can be seen from the proof. Assumption 2 implies that the eigenvalues of A are distinct from those of F. We shall show the following. Assumptions 1 and 2 imply that (H, F) is an observable pair. On the other hand, if (H, F) is an observable pair, then one can let $T = I$ and set the spectrum of A arbitrarily by choice of B. Almost all complex numbers are of type (C, ν) with respect to $\sigma(F)$, so assumption 2 is hardly a restriction on A when (H, F) is an observable pair. If A is chosen to be Hurwitz, then the state estimator is given by (1.4) and the error dynamics is locally asymptotically stable as before. We defer the proof of the main theorem to the next section.

CONVERSE TO THE MAIN THEOREM. *Consider the class of nonlinear systems described by the following equation:*

$$\begin{aligned} \dot{z} &= g(z), \\ y &= h(z), \end{aligned} \tag{1.8}$$

where $z \in \mathbf{R}^n$, $y \in \mathbf{R}^p$, and g, h are continuous vector fields on $\mathbf{R}^n$, $\mathbf{R}^p$, respectively, with $g(0) = 0$ and $h(0) = 0$. If there exists a nonlinear observer

$$\dot{\hat{z}} = \hat{g}(\hat{z}, y) \tag{1.9}$$

such that the error $\tilde{z} = z - \hat{z}$ dynamics is linear,

$$\dot{\tilde{z}} = A\tilde{z}, \tag{1.10}$$

where A is an $n \times n$ matrix, then there exists a continuous vector field $\beta : \mathbf{R}^p \to \mathbf{R}^n$ such that

$$g(z) = Az - \beta(h(z)), \tag{1.11}$$

$$\hat{g}(\hat{z}, y) = A\hat{z} - \beta(y). \tag{1.12}$$

Proof. The error dynamics is

$$\dot{\tilde{z}} = A\tilde{z} = g(z) - \hat{g}(\hat{z}, y).$$

Assume $z = 0$. Then

$$A\hat{z} = \hat{g}(\hat{z}, 0).$$

Assume $\tilde{z} = 0$. Then

$$g(z) = \hat{g}(z, h(z)).$$

Define

$$\beta(\hat{z}, y) = \hat{g}(\hat{z}, 0) - \hat{g}(\hat{z}, y).$$

Then

$$\begin{aligned} A\tilde{z} &= g(z) - \hat{g}(\hat{z}, y) \\ &= \hat{g}(z, h(z)) - \hat{g}(\hat{z}, h(z)) \\ &= \hat{g}(z, 0) - \beta(z, h(z)) - \hat{g}(\hat{z}, 0) + \beta(\hat{z}, h(z)) \\ &= Az - \beta(z, h(z)) - A\hat{z} + \beta(\hat{z}, h(z)). \end{aligned}$$

So

$$\beta(z, h(z)) = \beta(\hat{z}, h(z)).$$

But the left side does not depend on $\hat{z}$, so neither does the right, and thus

$$\beta(\hat{z}, h(z)) = \beta(h(z)).$$

Therefore

$$\begin{aligned} \hat{g}(\hat{z}, y) &= A\hat{z} - \beta(y), \\ g(z) &= Az - \beta(h(z)). \end{aligned}$$ □

Note. This converse shows that if a system (1.1) admits an observer with linear error dynamics after a smooth change of coordinates, it is because the PDE (1.3) is solvable for some smooth θ and continuous β.

The rest of the paper is organized as follows. Section 2.1 discusses the relationship between the linear part of the nonlinear system (1.1) and the terms of degree 1 of the solution (1.3). A unique formal solution of (1.3) is given in section 2.2 and this is shown to be convergent in section 2.3. We also show in section 2.1 that (1.3) has a unique solution for any choice of the eigenvalues of A except for a set of zero measure in $\mathbf{C}^n$. Several examples are treated in section 3. Section 4 applies the main result to the case when the system has inputs.

2. Solution of the PDE.

2.1. Terms of degree 1. If we focus on the terms of degree 1 in (1.3), we obtain the equation

$$TF = AT - BH. \tag{2.1}$$

We view this as a linear equation for T in terms of given F, H, A, B.

LEMMA 1. *Equation* (2.1) *admits a unique solution* T *if and only if the eigenvalues of* F *and* A *are distinct, that is,* $\sigma(F) \cap \sigma(A) = \emptyset$.

Proof. We give the proof when F admits a basis of right eigenvectors, $\{\mathbf{v}^j, j = 1, \ldots, n\}$, and A admits a basis of left eigenvectors, $\{\mathbf{w}_i, i = 1, \ldots, n\}$. The general case is similar using bases of generalized eigenvectors. Define an operator $\mathcal{F} : T \longmapsto TF - AT$ on the space of $n \times n$ matrices $\{T\}$. Let λ_i be the eigenvalue of F corresponding to the right eigenvector $\mathbf{w}_i$, and let μ_j be the eigenvalue of F corresponding to left eigenvector $\mathbf{v}^j$. Now $\{\mathbf{v}^j\mathbf{w}_i, i, j = 1, \ldots, n\}$ is a basis for $\{T\}$ and

$$\begin{aligned}\mathcal{F}(\mathbf{v}^j\mathbf{w}_i) &= (\mathbf{v}^j\mathbf{w}_i)F - A(\mathbf{v}^j\mathbf{w}_i)\\ &= (\lambda_i - \mu_j)\mathbf{v}^j\mathbf{w}_i.\end{aligned}$$

Thus $\mathcal{F}$ is invertible if and only if $\lambda_i - \mu_j \neq 0$ for all possible i and j. Therefore $\mathcal{F}T = -BH$ admits a unique solution if and only if $\sigma(F) \cap \sigma(A) = \emptyset$. □

LEMMA 2. *Suppose* $\sigma(F) \cap \sigma(A) = \emptyset$. *If* T *is invertible, then* (H, F) *is observable and* (A, B) *is controllable.*

Proof. Suppose (H, F) is not observable. Then there exist $\lambda_i \in \sigma(F)$ and a vector $x \in \mathbf{R}^{n\times 1}$, $x \neq 0$, such that $Hx = 0$ and $Fx = \lambda_i x$. Multiply (2.1) by x to obtain

$$\lambda_i Tx = TFx = ATx + BHx = ATx.$$

Since $Tx \neq 0$, this implies that $\lambda_i \in \sigma(A)$, a contradiction.

Similarly, suppose (A, B) is not controllable. Then there is $\mu_j \in \sigma(A)$ and a vector $\xi \in \mathbf{R}^{1\times n}$ such that $\xi A = \mu_j \xi$ and $\xi B = 0$. Multiply (2.1) by ξ to obtain

$$\xi TF = \xi AT + \xi BH = \mu_j \xi T.$$

Since $\xi T \neq 0$, this implies that $\mu_j \in \sigma(F)$, a contradiction. □

LEMMA 3. *If* T *is an invertible solution to* (2.1)*, then* A *is conjugate to* F *modified by output injection.*

Proof. Since T satisfies equation

$$TF + BH = AT$$

and T is invertible, we thus have

$$T(F + T^{-1}BH)T^{-1} = A. \qquad \square$$

LEMMA 4. *If* $\sigma(F) \cap \sigma(A) = \emptyset$ *and* A *is conjugate to* F *modified by output injection, then there exists* B *such that the unique solution to* (2.1) *is invertible.*

Proof. Since A is conjugate to F modified by output injection, there exist an $n \times n$ invertible matrix S and an $n \times p$ matrix G such that

$$S(F + GH)S^{-1} = A.$$

Hence we have $SF = AS - SGH$. Let $B = SG$. Then $SF = AS - BH$, so $T = S$ according to Lemma 1. Therefore T is invertible. □

Loosely speaking, a complex number μ is of type (C, ν) with respect to $\sigma(F) = \lambda$ if $|\mu - m \cdot \lambda|$ is never zero and does not approach zero too fast as $|m| \to \infty$. If ν is large enough, then the set of μ's which are of type (C, ν) for some $C > 0$ is dense in the complex plane.

LEMMA 5. *If $C > 0$ and $\nu > \frac{n}{2}$, then*

$$\text{meas}\left\{\mu : \mu \text{ is not of type } (C,\nu)\right\} \le k(n,\nu)C^2, \tag{2.2}$$

where $k(n, \nu)$ is a constant which depends only on n and ν.

If $\nu > \frac{n}{2}$, then the set of points which are not of type (C, ν) for any $C > 0$ is a set of zero measure.

Proof. Clearly, the set $\{\mu : \mu \text{ is not of type } (C, \nu)\}$ is

$$\bigcup_{|m|\ge 1} \text{Ball}\left(m \cdot \lambda,\ \frac{C}{|m|^\nu}\right),$$

where $\text{Ball}(p, r)$ stands for an open ball in $\mathbf{C}$ centered at $p \in \mathbf{C}$ with radius r. The measure of the $\text{Ball}(m \cdot \lambda, \frac{C}{|m|^\nu})$ is $\frac{\pi C^2}{|m|^{2\nu}}$. There are no more than $(d+1)^{n-1}$ choices of $m = (m_1, m_2, \ldots, m_n)$ such that $|m| = d$. To see this note that each of $m_1, \ldots, m_{n-1}$ must lie between 0 and d, and then $m_n = d - m_1 - \cdots - m_{n-1}$. Since $(d + 1) \le 2d$, we have

$$\text{meas} \bigcup_{|m|=d} \text{Ball}\left(m \cdot \lambda,\ \frac{C}{|m|^\nu}\right) \le \pi C^2 (2d)^{n-1-2\nu}.$$

Therefore, if $n - 1 - 2\nu < -1$, then

$$\text{meas} \bigcup_{|m|>0} \text{Ball}\left(m \cdot \lambda,\ \frac{C}{|m|^\nu}\right) \le \pi C^2 \left(\sum_{d=1}^{\infty} (2d)^{n-1-2\nu}\right),$$

so (2.2) follows. □

2.2. The formal solution of the PDE. Assume the hypothesis of the main theorem holds. We show that there is a unique solution to the PDE (1.3) within the class of formal power series. It is convenient to assume that F and A are diagonal; the proof in the general case is similar but much messier. We expand the terms in power series

$$\begin{aligned} f(x) &= Fx + f^{[2]}(x) + f^{[3]}(x) + \cdots, \\ \beta(h(x)) &= BHx + \beta^{[2]}(x) + \beta^{[3]}(x) + \cdots, \\ \theta(x) &= Tx + \theta^{[2]}(x) + \theta^{[3]}(x) + \cdots, \end{aligned}$$

where $f^{[d]}$, $\beta^{[d]}$, and $\theta^{[d]}$ are homogeneous polynomial vector fields of degree d in x. The knowns are f, h, β, T and the unknowns are the higher degree terms $\theta^{[2]}, \theta^{[3]}, \ldots$. The linear terms satisfy (2.1) by the above assumption.

The degree d part of (1.3) is

$$\frac{\partial \theta^{[d]}}{\partial x}(x)Fx - A\theta^{[d]}(x) = -\tilde{\beta}^{[d]}(x), \tag{2.3}$$

where

$$\tilde{\beta}^{[d]}(x) = \beta^{[d]}(x) + Tf^{[d]}(x) + \sum_{j=2}^{d-1} \frac{\partial \theta^{[j]}}{\partial x}(x) f^{[d+1-j]}(x). \tag{2.4}$$

Let $\mathbf{e}^k$ denote the kth unit vector in z space and let $x^m = x_1^{m_1} \cdots x_n^{m_n}$. Then the above terms can be expanded as

$$\tilde{\beta}^{[d]}(x) = \sum_{k=1}^{n} \sum_{|m|=d} \tilde{\beta}_{k,m} \mathbf{e}^k x^m,$$

$$\theta^{[d]}(x) = \sum_{k=1}^{n} \sum_{|m|=d} \theta_{k,m} \mathbf{e}^k x^m,$$

and we obtain the equations

$$(\mu_k - m \cdot \lambda)\theta_{k,m} = \tilde{\beta}_{k,m}. \tag{2.5}$$

These equations have unique solutions because $m \cdot \lambda - \mu_k \neq 0$. □

The formal approach yields a method for constructing an observer with approximately linear error dynamics. Start by choosing a T, A, B satisfying the linear equation (2.1). Then successively solve (2.3) up to some degree d. At each step $\beta^{[j]}$ can be chosen to make $\theta^{[j]}$ smaller and thereby try to keep $\theta(x)$ close to its globally invertible linear part Tx. The approximate solution

$$\theta(x) = Tx + \theta^{[2]}(x) + \theta^{[3]}(x) + \cdots + \theta^{[d]}(x),$$
$$\beta(y) = By + \beta^{[2]}(y) + \beta^{[3]}(y) + \cdots + \beta^{[d]}(y)$$

transforms the system (1.1) into

$$\dot{z} = Az - \beta(y) + O(x)^{d+1},$$

so the observer (1.4) has approximately linearizable error dynamics. The error is $O(x, \hat{x})^{d+1}$. When implementing the method, the matrices F, A need not be diagonal, but this makes solving (2.3) very easy.

2.3. Convergence of the formal solution. Let $|x| = \max\{|x_1|, \ldots, |x_n|\}$. We write

$$f(x) = Fx + \bar{f}(x),$$
$$\beta(y) = BHx + \bar{\beta}(x),$$

where $AT - TF = BH$. We first show that the sequence of PDEs

$$A\theta_2(x) - \frac{\partial}{\partial x}\theta_2(x)Fx = T\bar{f}(x) + \bar{\beta}(x),$$
$$A\theta_k(x) - \frac{\partial}{\partial x}\theta_k(x)Fx = \frac{\partial}{\partial x}\theta_{k-1}(x)\bar{f}(x)$$

admits a sequence of analytical solutions $\theta_2(x), \theta_3(x), \ldots$ in some neighborhood of the origin. Then we show that the sum

$$Tx + \theta_2(x) + \theta_3(x) + \cdots$$

converges to an analytic function which solves (1.3).

We define a positive real function $b_k : [0,1) \to [0,\infty)$ to be

$$b_k(q) := \max_{d\in\mathbf{Z}_+, d\geq k} \left[C^{-1} d^{\nu} q^{\frac{d}{2}} \right],$$

where $C > 0$ and $\nu > 0$ are given. We start with an important theorem.

THEOREM 1. *Let $P(x)$ be a real analytic function in $|x| < r$ with $P(0) = 0$ and $\frac{\partial P}{\partial x}(0) = 0$. Suppose all of the eigenvalues of A are of type (C,ν) with respect to $\sigma(F)$. Then the first-order PDE*

$$A\theta(x) - \frac{\partial \theta(x)}{\partial x} Fx = P(x) \tag{2.6}$$

admits a unique analytic solution $\theta(x)$ in $|x| < r$ with $\theta(0) = 0$.

Proof. The analyticity of $P(x)$ implies that $P(x)$ can be expanded into a Taylor series

$$P(x) = P^{[k]}(x) + P^{[k+1]}(x) + \cdots \qquad \text{for} \quad |x| < r$$

with

$$P^{[d]}(x) = \sum_{j=k}^{n} \sum_{|m|=d} p_{j,m} \mathbf{e}^j x^m,$$

where $k \geq 2$ is the lowest degree of $P(x)$. We assume a series solution

$$\theta(x) = \theta^{[k]}(x) + \theta^{[k+1]}(x) + \cdots + \theta^{[d]}(x) + \cdots \tag{2.7}$$

with

$$\theta^{[d]}(x) = \sum_{j=1}^{n} \sum_{|m|=d} \theta_{j,m} \mathbf{e}^j x^m.$$

If we plug (2.7) into (2.6), then we have

$$\theta_{j,m} = \frac{p_{j,m}}{\mu_j - m\cdot\lambda} \qquad \text{for } |m| \geq k, \quad 1 \leq j \leq n.$$

Since the eigenvalues of A are of type (C,ν) with respect to $\sigma(F)$, we have

$$|\theta_{j,m}| = \left|\frac{p_{j,m}}{\mu_j - m\cdot\lambda}\right| \leq \frac{|m|^{\nu}|p_{j,m}|}{C}.$$

We shall show that (2.7) converges on the closed polydisk $|x| \leq qr$ for any $0 < q < 1$. Hence (2.7) converges on $|x| < r$.

Consider a new series

$$\hat{P}(x) = \hat{P}^{[k]}(x) + \hat{P}^{[k+1]}(x) + \cdots \tag{2.8}$$

with

$$\hat{P}^{[d]}(x) = \sum_{j=k}^{n} \sum_{|m|=d} \frac{|m|^{\nu}|p_{j,m}|}{C} \mathbf{e}^j x^m, \qquad d \geq k.$$

We next claim that (2.8) converges in $|x| \le qr$. Let $\xi := (qr, qr, \dots, qr)$. Then

$$\begin{aligned}|\hat{P}^{[d]}(x)| &\le \max_{1\le j\le n} \sum_{|m|=d} \frac{|m|^\nu |p_{j,m}|}{C} |x|^m \le \max_{1\le j\le n} \sum_{|m|=d} \frac{|m|^\nu |p_{j,m}|}{C} |\xi|^m \\ &\le \max_{1\le j\le n} \sum_{|m|=d} |m|^\nu C^{-1} q^{\frac{|m|}{2}} |p_{j,m}| (\sqrt{q}r)^{|m|} \\ &\le b_k(q) \max_{1\le j\le n} \sum_{|m|=d} |p_{j,m}| (\sqrt{q}r)^{|m|}.\end{aligned}$$

Notice that $P(x)$ is an analytic function for $|x| < r$, so its Taylor series converges there absolutely, which yields

$$|\hat{P}(x)| \le b_k(q) \max_{1\le j\le n} \sum_{d=k}^{\infty} \left(\sum_{|m|=d} |p_{j,m}| (\sqrt{q}r)^{|m|} \right) < +\infty.$$

Thus (2.7) defines an analytic function $\theta(x)$ for $|x| < r$, which solves (2.6). ∎

From Theorem 1, we immediately have the following corollary.

COROLLARY 1. *Suppose all of the eigenvalues of A are of type (C, ν) with respect to $\sigma(F)$. The PDEs*

$$A\theta_2(x) - \frac{\partial \theta_2}{\partial x}(x) Fx = T\bar{f}(x) + \bar{\beta}(x), \qquad \theta_2(0) = 0, \tag{2.9}$$

$$A\theta_k(x) - \frac{\partial \theta_k}{\partial x}(x) Fx = \frac{\partial \theta_{k-1}}{\partial x}(x) \bar{f}(x), \qquad \theta_k(0) = 0, \quad k \ge 3, \tag{2.10}$$

admit analytic solutions in $|x| < r$.

The next step is to prove that

$$\theta_2(x) + \theta_3(x) + \cdots + \theta_k(x) + \cdots$$

converges near the origin and solves the PDE (1.3).

Since $\bar{f}(x) = O(|x|^2)$ is an analytic function in the polydisk $|x| \le r$, it can be expanded into a Taylor series:

$$\bar{f}(x) = f^{[2]}(x) + f^{[3]}(x) + \cdots, \qquad |x| \le r,$$

where $f^{[d]}(x) = \sum_{j=1}^n \sum_{|m|=d} f_{j,m} \mathbf{e}^j x^m$. Thus the following series converges:

$$\sum_{|m|=2} |f_{j,m}| r^2 + \sum_{|m|=3} |f_{j,m}| r^3 + \cdots := M_j$$

for $j = 1, 2, \dots, n$. We define

$$\bar{M}_f := \max \left\{ \frac{M_1}{r^2}, \dots, \frac{M_n}{r^2} \right\}$$

and

$$\|P(x)\| := \max_{1\le i\le n} \sum_m |p_{i,m} x^m|$$

if $P(x)$ is analytic in $|x| < r$ with

$$P(x) = \left(\sum_m p_{1,m} x^m, \sum_m p_{2,m} x^m, \ldots, \sum_m p_{n,m} x^m \right).$$

LEMMA 6. *There exists $0 < r_1 < r$ such that if $P(x)$ is analytic in $|x| < r_1$, where $\|P(x)\| \le N$, then*

$$\left\| \frac{\partial P}{\partial x}(x) \bar{f}(x) \right\| \le N \qquad in \quad |x| < r_1.$$

Proof. First it is easy to see that for any $r_1 < r$ we have

$$|\bar{f}(x)| \le r_1^2 \bar{M}_f \qquad \text{for} \quad |x| \le r_1,$$

since for $j = 1, 2, \ldots, n$

$$\begin{aligned} &\sum_{|m|=2} |f_{j,m}| r_1^2 + \sum_{|m|=3} |f_{j,m}| r_1^3 + \cdots \\ &\quad = r_1^2 \left(\sum_{|m|=2} |f_{j,m}| + \sum_{|m|=3} |f_{j,m}| r_1^1 + \cdots \right) \\ &\quad \le r_1^2 \frac{M_j}{r^2} \le r_1^2 \bar{M}_f. \end{aligned} \tag{2.11}$$

Next let

$$P(x) = (P_1(x), P_2(x), \ldots, P_n(x)),$$

with $P_i(x) = \sum_m p_{i,m} x^m$ and

$$N(r) := \max_{|x| \le r} \|P(x)\|.$$

The analyticity of $P(x)$ implies that

$$\frac{\partial P_i}{\partial x_j}(x) = \sum_m \frac{\partial}{\partial x_j} (p_{i,m} x^m) = \sum_m p_{i,m} m_j x_1^{m_1} \cdots x_j^{m_j - 1} \cdots x_n^{m_n}, \qquad |x| < r_1,$$

and for any given $\varepsilon > 0$ there exists $K > 0$ such that when $|m| > K$

$$\sum_{m, |m| \ge K} \left| p_{i,m} m_j x_1^{m_1} \cdots x_j^{m_j - 1} \cdots x_n^{m_n} \right| < \varepsilon$$

for $|x| < r_1$. Thus

$$\sum_{m, |m| \le K} |p_{i,m} m_j x_1^{m_1} \cdots x_j^{m_j - 1} \cdots x_n^{m_n}| \|\bar{f}_j(x)\| \le \sum_{m, |m| \le K} |p_{i,m}| m_j r_1^{|m|} r_1 \bar{M}_f.$$

Let r_1 be small enough such that

$$\sum_{m, |m| \le K} |p_{i,m}| m_j r_1^{|m|} r_1 \bar{M}_f \le \frac{N(r_1)}{n}.$$

Then for $|x| < r_1$

$$\sum_m \left| \frac{\partial}{\partial x_j} (p_{i,m} x^m) \right| \left\| \bar{f}_j(x) \right\| < \frac{N(r_1)}{n} + \varepsilon r_1^2 \bar{M}_f.$$

Thus we have

$$\left\| \frac{\partial P_i}{\partial x_j}(x) \bar{f}_j(x) \right\| \leq \sum_m \left| \left(\frac{\partial}{\partial x_j} (p_{i,m} x^m) \right) \right| \left\| \bar{f}_j(x) \right\| \leq \frac{N(r_1)}{n}.$$

Therefore

$$\left\| \frac{\partial P}{\partial x}(x) \bar{f}(x) \right\| \leq \max_{1 \leq i \leq n} \sum_{j=1}^{n} \left\| \frac{\partial P_i}{\partial x_j}(x) \bar{f}_j(x) \right\| \leq N(r_1). \qquad \square$$

In the definition of type (C, ν), without lose of generality we can assume that ν is a positive integer since if ν is not, we can replace it by a larger integer.

LEMMA 7. *Let $r_2 := r_1/n$, where r_1 is given in Lemma 6. Let $\theta_k(x)$ be the solution of*

$$A\theta_k(x) - \frac{\partial \theta_k}{\partial x}(x) Fx = \frac{\partial \theta_{k-1}}{\partial x}(x) \bar{f}(x).$$

Then if $\|\theta_{k-1}(x)\| \leq N$ for $|x| < r_2$, we have

$$\|\theta_k(x)\| \leq \frac{NP(|x_1| + |x_2| + \cdots + |x_n|)}{C\left(r_1 - (|x_1| + |x_2| + \cdots + |x_n|)\right)^{\nu+1}}$$

for $|x| < r_2$, where P is a polynomial of degree ν with coefficients depending only on r_1.

Proof. We first let $g(x) := \frac{\partial \theta_{k-1}}{\partial x}(x) \bar{f}(x)$ and

$$\phi(x) := \frac{N r_1}{r_1 - (x_1 + \cdots + x_n)}.$$

Clearly for $|x| < r_2$,

$$\begin{aligned} \phi(x) &= \frac{N}{1 - (x_1 + \cdots + x_n)/r_1} = N \sum_{d=0}^{\infty} \left(\frac{x_1 + \cdots + x_n}{r_1} \right)^d \\ &= N \sum_{d=0}^{\infty} \frac{1}{r_1^d} \sum_{|m|=d} \frac{|m|!}{m!} x^m \end{aligned}$$

and

$$D^m \phi(0) = N |m|! r_1^{-|m|}.$$

By the previous lemma, $|g(x)| \leq N$ for $|x| < r_1$, so the Cauchy estimate yields

$$|D^m g(0)| \leq N |m|! r_1^{-|m|},$$

where D^m is a partial differential operator of order m defined to be

$$D^m = \frac{\partial^m}{\partial x_1^{m_1} \cdots \partial x_n^{m_n}}.$$

Let

$$g(x) = g^{[k]}(x) + g^{[k+1]}(x) + \cdots + g^{[d]}(x) + \cdots$$

with $g^{[d]}(x) = \sum_{j=1}^{n} \sum_{|m|=d} g_{j,m} \mathbf{e}^j x^m$, where

$$|g_{j,m}| = \left| \frac{1}{m!} D^m g(0) \right| \le N \frac{|m|!}{m!} r_1^{-|m|}$$

and

$$\theta_k(x) = \theta_k^{[k]}(x) + \theta_k^{[k+1]}(x) + \cdots + \theta_k^{[d]}(x) + \cdots$$

with $\theta_k^{[d]}(x) = \sum_{j=1}^{n} \sum_{|m|=d} \theta_{j,m} \mathbf{e}^j x^m$. Then (2.12) implies that

$$\theta_{j,m} = \frac{g_{j,m}}{\mu_j - \lambda \cdot m}.$$

Since the eigenvalues of A are of type (C, ν) with respect to $\sigma(F)$, it follows that

$$|\theta_{j,m}| = \left| \frac{g_{j,m}}{\mu_j - \lambda \cdot m} \right| \le \frac{|m|^\nu}{C} |g_{j,m}| \le \frac{|m|^\nu |m|!}{Cm!} r_1^{-|m|}.$$

Next we claim that

$$N \sum_{d=0}^{\infty} \frac{1}{r_1^d} \sum_{|m|=d} \frac{|m|^\nu |m|!}{m! C} x^m = \frac{NP(x_1 + x_2 + \cdots + x_n)}{C(r_1 - x_1 - x_2 - \cdots - x_n)^{\nu+1}}.$$

For convenience, we denote $\hat{x} = x_1 + \cdots + x_n$. Notice that for $|x| < r_2$,

$$\frac{r_1}{r_1 - (x_1 + \cdots + x_n)} = \sum_{d=0}^{\infty} \frac{1}{r_1^d} \sum_{|m|=d} \frac{|m|!}{m!} x^m.$$

We differentiate above both sides with respect to $\hat{x}$ and then multiply both sides by $\hat{x}$,

$$\frac{r_1 \hat{x}}{(r_1 - \hat{x})^2} = \sum_{d=0}^{\infty} \frac{1}{r_1^d} \sum_{|m|=d} \frac{|m||m|!}{m!} x^m.$$

We repeat this procedure ν times and obtain

$$\frac{P(\hat{x})}{(r_1 - \hat{x})^{\nu+1}} = \sum_{d=0}^{\infty} \frac{1}{r_1^d} \sum_{|m|=d} \frac{|m|^\nu |m|!}{m!} x^m,$$

where $P(\hat{x})$ is a polynomial of degree ν with coefficients depending only on r_1. Hence

$$\sum_{d=0}^{\infty} \frac{1}{r_1^d} \sum_{|m|=d} \frac{|m|^\nu |m|!}{m!} |x^m| = \frac{P(|x_1| + \cdots + |x_n|)}{(r_1 - (|x_1| + \cdots + |x_n|))^{\nu+1}},$$

which yields the conclusion. □

Let $r_3 := r_2/2$ and

$$\hat{N} := \max_{|x| \le r_3} \frac{P(|x_1| + \cdots + |x_n|)}{C(r_1 - (|x_1| + \cdots + |x_n|))^{\nu+1}}$$

and

$$M := \max_{|x|\le r} \sum_{d=2}^{\infty} \left(|\beta^{[d]}(x)| + |Tf^{[d]}(x)| \right).$$

THEOREM 2. *Let $\theta_k(x)$ be the solution of*

$$A\theta_k(x) - \frac{\partial \theta_k}{\partial x}(x)Fx = \frac{\partial \theta_{k-1}}{\partial x}(x)\bar{f}(x), \qquad \theta_k(0) = 0.$$

Then for any $|x| \le qr_3$ with $0 < q < 1$ we have

$$\|\theta_k(x)\| \le b_k(q)\hat{N}^{k-2}M.$$

Proof. According to the previous lemma, we know that

$$\|\theta_2(x)\| \le M\hat{N} \quad \text{for} \quad |x| \le r_3.$$

Applying the lemma in a recursive way yields

$$\|\theta_k(x)\| \le M\hat{N}^{k-1} \quad \text{for} \quad |x| \le r_3 \quad \text{and} \quad k = 3, 4, \ldots.$$

Let $g(x) = \frac{\partial \theta_{k-1}}{\partial x}(x)\bar{f}(x)$. Then $g(x)$ can be expanded into a Taylor series in $|x| \le r_3$:

$$g(x) = g^{[k]}(x) + g^{[k+1]}(x) + \cdots$$

with $g^{[d]}(x) = \sum_{j=1}^{n} \sum_{|m|=d} g_{j,m}\mathbf{e}^j x^m$. Similar to the proof given in Theorem 1,

$$\|\theta_k(x)\| \le b_k(q) \sum_{d=k}^{\infty} \left(\sum_{|m|=d} |g_{j,m}|(\sqrt{q}r_3)^{|m|} \right) \le b_k(q)\hat{N}^{k-2}M,$$

and the proof is complete. ∎

COROLLARY 2. *When q is small enough, the series*

$$\theta_2(x) + \theta_3(x) + \cdots + \theta_k(x) + \cdots$$

converges in $|x| \le qr_3$, where $\theta_d(x)$ for $d = 2, 3, \ldots$ is the solution of (2.10).

Proof. Let $q \le \frac{1}{2^{\nu+1}\hat{N}}$. It is sufficient to show that

$$\theta_k(x) + \theta_{k+1}(x) + \cdots$$

converges for some fixed k in $|x| \le qr_3$. According to the definition of $b_k(q)$, we know that when $k \ge 2\nu/\ln\frac{1}{q}$, the following holds:

$$b_k(q) > b_{k+1}(q) > \cdots > b_d(q) > \cdots \to 0 \qquad \text{as} \quad d \to \infty.$$

Choose $k \ge 2\nu/\ln\frac{1}{q}$ and notice that

$$b_k(q) = k^\nu q^k, \quad b_{k+1} = (k+1)^\nu q^{k+1}, \ldots, \quad b_d(q) = d^\nu q^d, \ldots.$$

According to Theorem 2, we have

$$\|\theta_k(x)\| + \|\theta_{k+1}(x)\| + \cdots \le b_k(q)\hat{N}^{k-2}M + b_{k+1}(q)\hat{N}^{k-1}M + \cdots.$$

Since

$$\frac{b_{d+1}(q)\hat{N}^{d-1}M}{b_d(q)\hat{N}^{d-2}M} = \left(1+\frac{1}{d}\right)^{\nu} q\hat{N} < 2^{\nu} q\hat{N} \le \frac{1}{2}, \qquad d \ge k,$$

we thus complete the proof. □

From Corollary 2, we know that series

$$\theta_2(x) + \theta_3(x) + \cdots + \theta_d(x) + \cdots \tag{2.12}$$

defines an analytic function in $|x| \le qr_3$. Now we are ready to prove the main result of this paper.

Proof of the main theorem. We first define two functions in $|x| \le qr_3$:

$$\theta(x) := Tx + \theta_2(x) + \theta_3(x) + \cdots + \theta_d(x) + \cdots$$

and

$$\theta^L(x) := Tx + \theta_2(x) + \theta_3(x) + \cdots + \theta_L(x),$$

where $\theta_2(x), \theta_3(x), \ldots$ are the solutions of (2.9), (2.10). We next show that $\theta(x)$ solves (1.3). Now

$$\begin{aligned} &A\theta^L(x) - \frac{\partial\theta^L(x)}{\partial x} f(x) - \beta(h(x)) \\ &\quad = A\theta^L(x) - \frac{\partial\theta^L(x)}{\partial x}(Fx + \bar{f}(x)) - (BHx + \bar{\beta}(x)) \\ &\quad = \frac{\partial\theta_L(x)}{\partial x}\bar{f}(x). \end{aligned}$$

If $|x| \le qr_3$, then $\|\theta_L(x)\| \le b_L(q)\hat{N}^{L-2}M$ and

$$\left\| \frac{\partial\theta_L(x)}{\partial x}\bar{f}(x) \right\| \le b_L(q)\hat{N}^{L-2}M \to 0 \qquad \text{as} \quad L \to \infty$$

since series

$$b_k(q)\hat{N}^{k-2}M + b_{k+1}(q)\hat{N}^{k-1}M + \cdots$$

converges. Therefore $\theta(x)$ is an analytic solution of (1.3). Uniqueness follows from the uniqueness of the formal power series. □

A slight modification of the proof of the main theorem yields the following.

COROLLARY 3 (Lyapunov's auxiliary theorem). *Assume that $f : \mathbf{R}^n \to \mathbf{R}^n$ and $\gamma : \mathbf{R}^n \to \mathbf{R}^n$ are analytic vector fields with $f(0) = 0$, $\frac{\partial f}{\partial x}(0) = F$, and $\gamma(0) = 0$. Suppose that the eigenvalues $\lambda_1, \ldots, \lambda_n$ of F lie wholly in the open left half plane or lie wholly in the open right half plane. Let A be an $n \times n$ matrix with eigenvalues $\mu_1, \ldots, \mu_n$ such that there do not exist nonnegative integers $m_1, m_2, \ldots, m_n$ not all zero such that $\sum_{i=1}^n m_i\lambda_i = \mu_j$. Then there is a unique analytic solution in some neighborhood of the origin of the first-order PDE:*

$$\frac{\partial\theta}{\partial x}(x) f(x) - A\theta(x) + \gamma(x) = 0$$

with initial condition $\theta(0) = 0$.

Proof. Let $h(x) = x$ and $\beta(h(x) = \gamma(x)$. The main theorem cannot be applied directly because the Lyapunov auxiliary theorem does not require $\theta(x)$ to be a local diffeomorphism. But the proof stills holds provided we can show that the spectrum of A is of class (C, ν) with respect to the spectrum of F. Suppose the spectrum of F lies wholly in the open right half plane. Then there is a constant $c > 0$ such that $c \leq Re\ \lambda_i$, $i = 1, \ldots, n$. Suppose $M \geq Re\ \mu_j$, $j = 1, \ldots, n$. Then

$$|m \cdot \lambda - \mu_j| \geq 1$$

whenever $|m| \geq \frac{M+1}{c}$. Let $\nu = 1$ and choose $0 < C \leq 1$ satisfying

$$|m \cdot \lambda - \mu_j| \geq C$$

whenever $|m| < \frac{M+1}{c}$. This is possible because the left side is never zero. We have shown that the spectrum of A is of class (C, ν) with respect to the spectrum of F. □

COROLLARY 4 (Siegel's theorem). *Assume that $f : \mathbf{R}^n \to \mathbf{R}^n$ is an analytic vector field with $f(0) = 0$, $\frac{\partial f}{\partial x}(0) = F$. Suppose, for some $C > 0, \nu > 0$, the eigenvalues of F are of type (C, ν) with respect to $\sigma(F)$. Then there is an analytic solution in some neighborhood of the origin of the first-order PDE:*

$$\frac{\partial \theta}{\partial x}(x) f(x) = F\theta(x)$$

with initial condition $\theta(0) = 0$. Moreover $z = \theta(x)$ is a local analytic diffeomorphism around $x = 0$ which transforms the differential equation

$$\dot{x} = f(x)$$

into its linear part

$$\dot{z} = Fz.$$

Proof. Apply the main theorem with $\beta = 0$, $A = F$, and $T = I$. □

Note. Lyapunov's auxiliary theorem and Siegel's theorem are usually stated for complex analytic vector fields. We have stated them for real analytic vector fields since we stated our main theorem that way. But the proof of the main theorem holds for complex vector fields too.

3. Examples. As discussed in the introduction, there are distinct advantages to considering *nonlinear output injection* $\beta(y)$. It is desirable that θ be a diffeomorphism over as large a range as possible, for this is the domain of convergence of the observer. Nonlinear output injection can make θ a global diffeomorphism.

To illustrate this, we consider a Duffing oscillator

$$\begin{aligned} \ddot{x} &= x - x^3, \\ y &= x, \end{aligned}$$

which is equivalent to the planar system

$$\begin{aligned} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ -x_1^3 \end{bmatrix}, \\ y &= x_1. \end{aligned}$$

This system is trivially transformed into a linear system with output injection (1.2)

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ -2 & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} - \begin{bmatrix} -2y \\ -3y + y^3 \end{bmatrix}$$

by

$$\begin{aligned} \theta(x) &= x, \\ \beta(y) &= \begin{bmatrix} -2y \\ -3y + y^3 \end{bmatrix}. \end{aligned}$$

Notice that β is nonlinear and θ is trivially a global diffeomorphism. The observer (1.4) is

$$\begin{bmatrix} \dot{\hat{x}}_1 \\ \dot{\hat{x}}_2 \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ -2 & 0 \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} - \begin{bmatrix} -2y \\ -3y + y^3 \end{bmatrix},$$

and the error dynamics

$$\begin{bmatrix} \dot{\tilde{x}}_1 \\ \dot{\tilde{x}}_2 \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ -2 & 0 \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix}$$

is linear and exponentially stable with poles at $-1 \pm i$.

The example is trivial but illustrates two important facts. The first is the advantage of allowing nonlinear β. We could take it to be linear,

$$\beta(y) = \begin{bmatrix} -2 \\ -3 \end{bmatrix} y,$$

and still solve the PDE (1.3) for θ. But the solution might be hard to find, it could have an infinite power series expansion, and it might not be a global diffeomorphism.

The second point is that the Duffing oscillator is truly nonlinear; it has three equilibria and two homoclinic orbits, and the rest of the trajectories are limit cycles. Yet it is possible to build a globally convergent error with linear error dynamics.

Next we consider a Van der Pol oscillator,

$$\begin{aligned} \ddot{x} &= -(x^2 - 1)\dot{x} - x, \\ y &= x, \end{aligned}$$

which is equivalent to the planar system

$$\begin{aligned} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} &= \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 0 \\ x_1^2 x_2 \end{bmatrix}, \\ y &= \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \end{aligned}$$

Now we have

$$\begin{aligned} f(x) &= \begin{bmatrix} x_2 \\ -x_1 + x_2 - x_1^2 x_2 \end{bmatrix}, \qquad h(x) = x_1, \\ F &= \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}, \qquad H = \begin{bmatrix} 1 & 0 \end{bmatrix}. \end{aligned}$$

We look for a nonlinear coordinate transformation $z = \theta(x)$ such that in the new coordinates z, the system can be described in the form

$$\dot{z} = Az - \beta(y).$$

Let us choose A and β to be

$$A = \begin{bmatrix} b_1 & 1 \\ b_2 - 1 & 1 \end{bmatrix}, \qquad \beta(y) = \begin{bmatrix} b_1 y + \frac{y^3}{3} \\ b_2 y + \frac{y^3}{3} \end{bmatrix},$$

where b_1, b_2 are constants such that $1 + b_1 < 0$, $b_1 - b_2 + 1 > 0$. Clearly, A is stable since $\operatorname{trace}(A) = 1 + b_1 < 0$ and $\det(A) = b_1 - b_2 + 1 > 0$. Moreover $A = F + BH$ with $B = [b_1, b_2]'$. The solution of (1.3) in this case is given by

$$\theta(x) = \begin{bmatrix} x_1 \\ x_2 + \frac{x_1^3}{3} \end{bmatrix}.$$

Note that θ is polynomial and *globally invertible* on $\mathbf{R}^2$. This is because we chose a nonlinear β. The resulting observer is again globally convergent with exponentially stable linear error dynamics in $\tilde{z}$ coordinates despite the nonlinearities of the Van der Pol oscillator. See Figure 1.

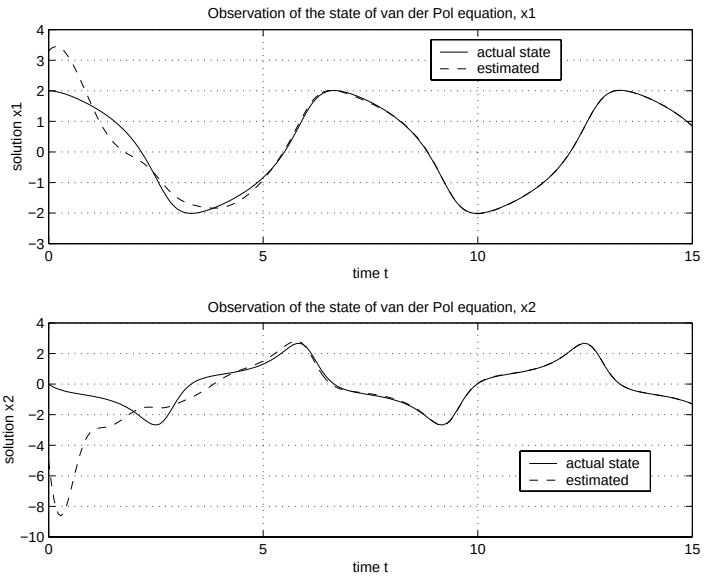


FIG. 1. *Observation of Van der Pol oscillator.*

Both these examples could be treated by the method of Krener and Respondek [8]. In particular, they showed that any observable two-dimensional system of the form

$$\begin{aligned} y &= x_1, \\ \dot{x}_1 &= x_2, \\ \dot{x}_2 &= f_2(x) = a(x_1) + b(x_1)x_2 + c(x_1)x_2^2, \end{aligned}$$

where $a(x_1), b(x_1), c(x_1)$ are smooth functions, admits a local observer with linear error dynamics in transformed coordinates. But their method is not applicable to more general f_2. The above method is applicable to any observable system with arbitrary f_2. The conditions of Krener and Respondek become more restrictive as the dimension of the system is increased, while there are no additional conditions for the above method.

The next example cannot be treated by the method of Krener and Respondek:

$$\begin{aligned} y &= x_1, \\ \dot{x}_1 &= 2x_2, \\ \dot{x}_2 &= 2x_1 - 3x_1^2 - x_2(x_1^3 - x_1^2 + x_2^2). \end{aligned}$$

There is a saddle at $(0,0)$ and an unstable source at $(2/3,0)$. The stable and unstable manifolds of the saddle form a homoclinic orbit given by $x_1^3 - x_1^2 + x_2^2 = 0$ which wraps around the unstable source.

The system is linearly observable around $x = 0$ with

$$F = \begin{bmatrix} 0 & 2 \\ 2 & 0 \end{bmatrix}, \quad H = \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

The spectrum of F is $\lambda = (2, -2)$. We choose a linear output injection based on a long time Kalman filter for the linear part of the system corrupted by standard white noises, and this leads to

$$A = \begin{bmatrix} -\sqrt{17} & 2 \\ -2 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} -\sqrt{17} \\ -4 \end{bmatrix}.$$

The spectrum of A is

$$\frac{-\sqrt{17} \pm 1}{2},$$

and clearly these are not resonant with the spectrum of F because they are not even integers.

First we compute θ for up to degree 3 with $\beta^{[2]} = 0$, $\beta^{[3]} = 0$:

$$\begin{aligned} \theta^{[1]}(x) &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \\ \theta^{[2]}(x) &= \begin{bmatrix} 1.2188 & -0.7731 & -0.2812 \\ 1.7394 & 0.2812 & -1.3529 \end{bmatrix} \begin{bmatrix} x_1^2 \\ x_1 x_2 \\ x_2^2 \end{bmatrix}, \\ \theta^{[3]}(x) &= \begin{bmatrix} -20.4026 & 19.1878 & 20.8159 & -20.1972 \\ -21.7136 & 20.8245 & 20.6972 & -20.8216 \end{bmatrix} \begin{bmatrix} x_1^3 \\ x_1^2 x_2 \\ x_1 x_2^2 \\ x_2^3 \end{bmatrix}. \end{aligned}$$

Figure 2 shows the system starting at $x_1 = 0.5, x_2 = 0$ and the observer starting at $\hat{x}_1 = 0, \hat{x}_2 = 0$. Clearly this observer does not converge; in particular, the observer seems to stall around $(0.3, 0.4)$. The problem appears to be caused by the large sizes of $\theta^{[2]}$ and $\theta^{[3]}$.

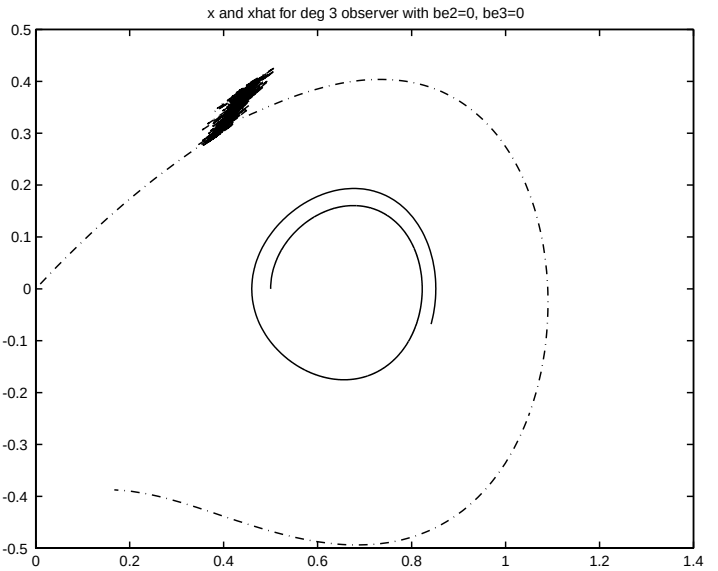


FIG. 2. *Solid line: state trajectory. Dashed line: observer trajectory.*

Next we choose $\beta^{[2]}$ to minimize the Euclidean norm of the coefficients of $\theta^{[2]}$, and then we choose $\beta^{[3]}$ to minimize the Euclidean norm of the coefficients of $\theta^{[3]}$. The result is

$$\theta^{[1]}(x) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

$$\theta^{[2]}(x) = \begin{bmatrix} 0.0000 & -0.0000 & 0.0000 \\ 0.0000 & -0.0000 & 0.0000 \end{bmatrix} \begin{bmatrix} x_1^2 \\ x_1x_2 \\ x_2^2 \end{bmatrix},$$

$$\theta^{[3]}(x) = \begin{bmatrix} 0.0330 & 0.0938 & -0.2219 & 0.1925 \\ -0.4030 & -0.1514 & 0.3075 & 0.1749 \end{bmatrix} \begin{bmatrix} x_1^3 \\ x_1^2x_2 \\ x_1x_2^2 \\ x_2^3 \end{bmatrix},$$

$$\beta(y) = \begin{bmatrix} -4.1231 & 0.0000 & -1.1296 \\ -4.0000 & 3.0000 & 0.2368 \end{bmatrix} \begin{bmatrix} y \\ y^2 \\ y^3 \end{bmatrix}.$$

Notice how much smaller $\theta^{[2]}$ and $\theta^{[3]}$ are. The resulting observer performs much better, as can be seen from Figure 3.

4. Nonlinear observer design with inputs. We now consider a nonlinear system with inputs:

$$\dot{x} = f(x,u), \tag{4.1}$$

$$y = h(x,u), \tag{4.2}$$

where $f : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^n$ and $h : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^p$ are continuous. We assume here that

$$f(x,u) = f_0(x) + f_1(x,u), \qquad h(x,u) = h_0(x) + h_1(x,u)$$

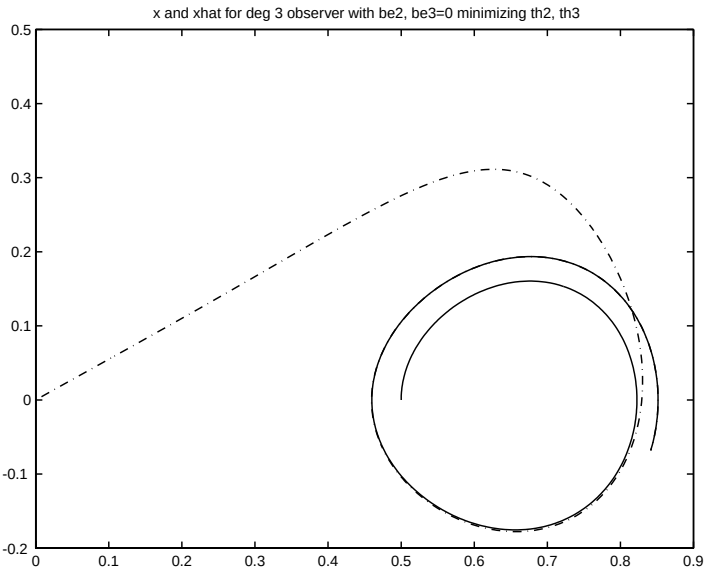


FIG. 3. *Solid line: state trajectory. Dashed line: observer trajectory.*

with $f_1(x,0) \equiv 0, h_1(x,0) \equiv 0$, and $f_0 : \mathbf{R}^n \to \mathbf{R}^n$ and $h_0 : \mathbf{R}^n \to \mathbf{R}^p$ are real analytic functions with $f_0(0) = 0, h_0(0) = 0$. Let $\beta : \mathbf{R}^p \to \mathbf{R}^n$ be a real analytic function and $F = \frac{\partial f_0}{\partial x}(0)$, $H = \frac{\partial h_0}{\partial x}(0)$, and $B = \frac{\partial \beta}{\partial x}(0)$. We further assume that

1. for a given $n \times n$ matrix A, there exists an invertible $n \times n$ matrix T so that $TFT^{-1} = A - BH$;
2. there exists a $C > 0, \nu > 0$ such that all the eigenvalues of A are of type (C, ν) with respect to $\sigma(F)$.

Then according to the main result of this paper, we know that the first-order PDE

$$\frac{\partial \phi}{\partial x}(x) f_0(x) = A\phi(x) - \beta(h_0(x)) \tag{4.3}$$

has a unique analytic solution $z = \phi$, which is a diffeomorphism in some neighborhood U of the origin with $\frac{\partial \phi}{\partial x}(0) = T$.

Now we let the estimate of the true state obey the equation

$$\dot{\hat{x}} = f(\hat{x}, u) - \left[\frac{\partial \phi}{\partial \hat{x}}\right]^{-1} (\beta(y) - \beta(h(\hat{x}, u))). \tag{4.4}$$

Let e denote

$$e = \phi(\hat{x}) - \phi(x).$$

Then e satisfies the differential equation

$$\begin{aligned} \dot{e} &= \frac{\partial \phi}{\partial \hat{x}} f(\hat{x}, u) - (\beta(y) - \beta(h(\hat{x}, u))) - \frac{\partial \phi}{\partial x} f(x, u) \\ &= \frac{\partial \phi}{\partial \hat{x}} (f_0(\hat{x}) + f_1)(\hat{x}, u) - (\beta(y) - \beta(h(\hat{x}, u))) - \frac{\partial \phi}{\partial x} (f_0(x) + f_1(x, u)). \end{aligned}$$

Since

$$\frac{\partial \phi}{\partial \hat{x}} f_0(\hat{x}) = A\phi(\hat{x}) - \beta(h_0(\hat{x})),$$
$$\frac{\partial \phi}{\partial x} f_0(x) = A\phi(x) - \beta(h_0(x)),$$

this yields

$$\dot{e} = Ae + N(\hat{x}, u) - N(x, u), \tag{4.5}$$

where the nonlinear function N is defined to be

$$N(x, u) := \frac{\partial \phi}{\partial x}(x) f_1(x, u) + \beta(h(x, u)) - \beta(h_0(x)). \tag{4.6}$$

We further assume that $f_1(\cdot, u)$ is locally Lipschitz about the origin; then there exists a positive constant $L(u)$ such that

$$\|N(x_1, u) - N(x_2, u)\| \leq L(u)\|x_1 - x_2\|$$

for all x_1, x_2 in some open neighborhood U containing the origin. If we choose A to be Hurwitz, then for any given positive-definite $Q \in \mathbf{R}^{n\times n}$ there exists a unique positive-definite $P \in \mathbf{R}^{n\times n}$ such that

$$A^T P + PA = -2Q.$$

Now we consider the Lyapunov function

$$V(e) = e^T P e.$$

The derivative of $V(e)$ evaluated along the solution of the error dynamics is given by

$$\dot{V}(e) = \dot{e}^T P e + e^T P \dot{e} = -2e^T Q e + 2e^T P[N(x + e, u) - N(x, u)].$$

Therefore we have

$$\begin{aligned} \dot{V}(e) &\leq -2e^T Q e + 2L(u)\|Pe\|\|e\| \\ &\leq (-2\lambda_{\min}(Q) + 2L(u)\lambda_{\max}(P))\|e\|, \end{aligned}$$

where $\lambda_{\min}(Q)$ is the minimum eigenvalue of Q and $\lambda_{\max}(P)$ is the maximum eigenvalue of P. Hence if

$$\lambda_{\min}(Q)/\lambda_{\max}(P) > L(u),$$

then $e = 0$ is local asymptotically stable.

REFERENCES

[1] V. I. Arnol'd, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer-Verlag, Berlin, 1988.

[2] D. Bestle and M. Zeitz, *Canonical form observer design for nonlinear time-variable systems*, Internat. J. Control, 38 (1983), pp. 419–431.

[3] L. C. Evans, *Partial Differential Equations*, AMS, Providence, RI, 1998.

[4] P. Glendinning, *Stability, Instability and Chaos: An Introduction to the Theory of Nonlinear Differential Equations*, Cambridge University Press, Cambridge, UK, 1994.

[5] N. Kazantzis and C. Kravaris, *Nonlinear observer design using Lyapunov's auxiliary theorem*, Systems Control Lett., 34 (1998), pp. 241–247.
[6] A. J. Krener and A. Isidori, *Linearization by output injection and nonlinear observers*, Systems Control Lett., 3 (1983), pp. 47–52.
[7] A. J. Krener, *Approximate linearization by state feedback and coordinate change*, Systems Control Lett., 5 (1984), pp. 181–185.
[8] A. J. Krener and W. Respondek, *Nonlinear observers with linearizable error dynamics*, SIAM J. Control Optim., 23 (1985), pp. 197–216.
[9] A. J. Krener, *Nonlinear stabilizability and detectability*, in Systems and Networks: Mathematical Theory and Applications, U. Helmke, R. Mennicken, and J. Saurer, eds., Akademie Verlag, Berlin, 1994, pp. 231–250.
[10] A. M. Liapunov, *Stability of Motion*, Academic Press, New York, London, 1966.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 3, pp. 954–974

MULTISCALE SINGULARLY PERTURBED CONTROL SYSTEMS: LIMIT OCCUPATIONAL MEASURES SETS AND AVERAGING*

VLADIMIR GAITSGORY† AND MINH-TUAN NGUYEN†‡

Abstract. An averaging technique for nonlinear multiscale singularly perturbed control systems is developed. Issues concerning the existence and structure of limit occupational measures sets generated by such systems are discussed. General results are illustrated with special cases.

Key words. multiscale singularly perturbed control systems, occupational measures, averaging method, limit occupational measures sets, nonlinear control, approximation of slow motions

AMS subject classifications. 34E15, 34C29, 34A60, 93C70, 34A4

PII. S0363012901393055

1. Introduction.

In this paper we consider a singularly perturbed control system containing several small parameters $\epsilon_1, \dots, \epsilon_m$ ($m \geq 1$). The parameters are introduced in such a way that the state variables of the system are decomposed into a group of "slow" variables which change their values with the rates of the order $O(1)$ and m groups of "fast" variables which change their values with the rates of the orders $O(\epsilon_1^{-1}), O(\epsilon_1^{-1}\epsilon_2^{-1}), \dots, O(\epsilon_1^{-1}\epsilon_2^{-1}\dots\epsilon_m^{-1})$, respectively.

The main contribution of the paper is the description of the structure of the limit control system, the solutions of which allow us to approximate the slow variables when the parameters ϵ_i, $i = 1, \dots, m$, tend to zero. The role of controls in the limit system is played by probability measures defined on the product of the original control set and a subset of the state space containing all the fast trajectories (both are assumed to be compact). These probability measures are chosen from a limit set of occupational measures generated by the admissible controls and trajectories of an associated system which describes the dynamics of the fast variables if the slow ones are "frozen" (see exact definitions below). The existence of such a set (called limit occupational measures set (LOMS)) and its structure are the central issues discussed in the paper.

Singularly perturbed control systems (SPCS) with one small parameter ($m = 1$) have been intensively studied in the literature, the most common approaches being related either to Tikhonov-type theorems justifying the equating of the small parameter to zero with further application of the boundary layer method (see [24], [30]) to asymptotically describe the fast dynamics (see, e.g., [13], [21], [22], [25], [28], [31]) or to different types of averaging techniques (see [1], [2], [3], [4], [5], [8], [11], [14], [15], [16], [17], [18], [19], [20], [27], [32]) which allow us to deal with the situation when the equating of the parameter to zero may not lead to a right approximation.

The literature on multiscale SPCS ($m > 1$) is much less intensive. Most available references concern linear control systems (see, e.g., [12], [26], and references therein).

*Received by the editors July 30, 2001; accepted for publication (in revised form) February 25, 2002; published electronically September 19, 2002. This work was supported by Australian Research Council grant A49906132.

http://www.siam.org/journals/sicon/41-3/39305.html

†University of South Australia, School of Mathematics, The Mawson Lakes Campus, Mawson Lakes SA 5095, Australia (v.gaitsgory@unisa.edu.au).

‡Present address: Joint Systems Branch, DSTO, P.O. Box 1500, Edinburgh SA 5111, Australia (minh@linus.levels.unisa.edu.au).

A technique of averaging type applicable to nonlinear control systems having a triangular structure (weakly coupled) was proposed in [20].

In [18] an averaging technique allowing us to deal with a general form of SPCS containing two small parameters ($m = 2$) was developed. The extension of the technique to the case $m > 2$ is, however, hardly possible since it involves a multiple averaging over time and leads to really complex expressions which are difficult to comprehend. In this paper, an averaging over time is replaced by averaging over measures from the LOMS. It resembles approaches used in dealing with stochastic SPCS (see, e.g., [9], [23], [34]) and makes the transition from the case $m = k$ to the case $m = k + 1$ ($\forall k = 1, 2, \ldots$) very natural.

Different issues related to averaging over occupational measures in SPCS with one small parameter were discussed in [2], [3], [4], [5], [17], [32]. In [17], in particular, LOMS for control systems without small parameters were considered. In this paper, we introduce and study such sets for singularly perturbed control systems (as is the associated system if the original system is multiscale).

The paper is organized as follows. Section 1 is this introduction. In section 2 statements about approximation of the slow motions by the solutions of the averaged system are formulated under the assumption that the LOMS of the associated system exists. An application of these results to problems of optimal control is demonstrated and a special case concerning systems linear in fast variables and controls is considered. In section 3 issues of existence and structure of the LOMS are addressed and a multistage averaging procedure for the construction of the LOMS is presented. The procedure is then illustrated with a special case of control systems which have a triangular structure (similar to those studied in [20]). Proofs of most of the statements are provided in section 4.

2. Averaging of multiscale SPCS.

2.1. Preliminaries. Given a compact metric space W, $\mathcal{B}(W)$ will stand for the σ-algebra of its Borel subsets and $\mathcal{P}(W)$ will denote the set of probability measures defined on $\mathcal{B}(W)$. The set $\mathcal{P}(W)$ will always be treated as a compact metric space with a metric ρ, which is consistent with its weak convergence topology. That is, a sequence $\gamma^k \in \mathcal{P}(W)$, $k = 1, 2, \ldots$, converges to $\gamma \in \mathcal{P}(W)$ in this metric if and only if

$$\lim_{k\to\infty} \int_W \phi(w)\gamma^k(dw) \;=\; \int_W \phi(w)\gamma(dw)$$

for any continuous $\phi(w) : W \to \mathbb{R}^1$.

Using the metric ρ, one can define the Hausdorff metric ρ_{H} on the set of subsets of $\mathcal{P}(W)$:

$$\rho_{\text{H}}(\Gamma_1, \Gamma_2) \stackrel{\text{def}}{=} \max\Big\{ \sup_{\gamma\in\Gamma_1} \rho(\gamma, \Gamma_2), \sup_{\gamma\in\Gamma_2} \rho(\gamma, \Gamma_1)\Big\} \quad \forall \Gamma_1, \Gamma_2 \in \mathcal{P}(W), \tag{2.1}$$

$$\text{where} \qquad \rho(\gamma, \Gamma_i) \stackrel{\text{def}}{=} \inf_{\gamma'\in\Gamma_i} \rho(\gamma, \gamma'), \quad i = 1, 2.$$

We will deal with the convergence in the Hausdorff metric of sets in $\mathcal{P}(W)$ defined as unions of occupational measures. Given a measurable function $w(t) : [0, T] \to W$, the occupational measure $p^{w(\cdot)} \in \mathcal{P}(W)$ generated by this function is defined by taking

$$p^{w(\cdot)}(Q) \stackrel{\text{def}}{=} \frac{1}{T}\text{meas}\Big\{ t \mid w(t) \in Q\Big\} \quad \forall Q \in \mathcal{B}(W), \tag{2.2}$$

where meas $\{\cdot\}$ stands for the Lebesgue measure on $[0, T]$.

2.2. Setting. Consider the SPCS

$$\begin{aligned}
\epsilon_1\epsilon_2\dots\epsilon_{m-1}\epsilon_m\dot{y}_1(t) &= f_1\big(u(t),y_1(t),\dots,y_m(t),z(t)\big),\\
&\vdots\ \vdots\ \vdots \qquad \vdots\ \vdots\ \vdots\\
\epsilon_{m-1}\epsilon_m\dot{y}_{m-1}(t) &= f_{m-1}\big(u(t),y_1(t),\dots,y_m(t),z(t)\big),\\
\epsilon_m\dot{y}_m(t) &= f_m\big(u(t),y_1(t),\dots,y_m(t),z(t)\big),\\
\dot{z}(t) &= g\big(u(t),y_1(t),\dots,y_m(t),z(t)\big),
\end{aligned} \tag{2.3}$$

where $\epsilon \stackrel{\text{def}}{=} (\epsilon_1,\epsilon_2,\dots,\epsilon_m)$ is a vector of small positive parameters, $t\in[0,T]$, and the functions $f_i : U\times\mathbb{R}^{M_1}\times\dots\times\mathbb{R}^{M_m}\times\mathbb{R}^N\to\mathbb{R}^{M_i}$, $i=1,\dots,m$, and $g : U\times\mathbb{R}^{M_1}\times\dots\times\mathbb{R}^{M_m}\times\mathbb{R}^N\to\mathbb{R}^N$ are continuous and satisfy Lipschitz conditions in $(y_1,\dots,y_m,z)$. Admissible controls are Lebesgue measurable functions $u(t):[0,T]\to U$, where U is a compact metric space.

Consider also the system

$$\begin{aligned}
\epsilon_1\epsilon_2\dots\epsilon_{m-1}\dot{y}_1(\tau) &= f_1\big(u(\tau),y_1(\tau),\dots,y_m(\tau),z\big),\\
&\vdots\ \vdots\ \vdots \qquad \vdots\ \vdots\ \vdots\\
\epsilon_{m-1}\dot{y}_{m-1}(\tau) &= f_{m-1}\big(u(\tau),y_1(\tau),\dots,y_m(\tau),z\big),\\
\dot{y}_m(\tau) &= f_m\big(u(\tau),y_1(\tau),\dots,y_m(\tau),z\big),\\
z &= \text{constant},
\end{aligned} \tag{2.4}$$

in which z is fixed and $\tau\in[0,S]$. This system will be referred to as an *associated system* with respect to SPCS (2.3). It is formally obtained from the "fast" subsystem of (2.3) via the replacement of the time scale $\tau = t\epsilon_m^{-1}$. Admissible controls for the associated system (2.4) are Lebesgue measurable functions $u(\tau):[0,S]\to U$. The solutions of (2.3) and (2.4) which are obtained with admissible controls are called admissible trajectories.

Assumption 2.1. (i) *There exist compact sets $Y_i''\subseteq Y_i'\subset\mathbb{R}^{M_i}$, $i=1,\dots,m$, and $Z''\subseteq Z'\subset\mathbb{R}^N$ such that the admissible trajectories of SPCS* (2.3) *which satisfy the initial conditions*

$$\big(y_1(0),\dots,y_m(0),z(0)\big)\in Y_1''\times\dots\times Y_m''\times Z'' \tag{2.5}$$

do not leave the set $Y_1'\times\dots\times Y_m'\times Z'$ on the interval $[0,T]$.

(ii) *There exist compact sets Y_i ($Y_i'\subseteq Y_i$), $i=1,\dots,m$, and Z ($Z'\in\operatorname{int}Z$) such that for any z from Z, the admissible trajectories of system* (2.4) *which satisfy the initial conditions*

$$\big(y_1(0),\dots,y_m(0)\big)\in Y_1'\times\dots\times Y_m' \tag{2.6}$$

do not leave the set $Y_1\times\dots\times Y_m$ on the interval $[0,\infty)$.

Note that to verify this assumption, one can use results from viability theory (see Chapter 5 in [6] and also [29] for further references).

Let us introduce the following notation:

$$y(\tau)\stackrel{\text{def}}{=}\big(y_1(\tau),\dots,y_m(\tau)\big),\quad Y\stackrel{\text{def}}{=}Y_1\times\dots\times Y_m,$$

and also $Y'\stackrel{\text{def}}{=}Y_1'\times\dots\times Y_m'$, $Y''\stackrel{\text{def}}{=}Y_1''\times\dots\times Y_m''$.

Let $u(\tau)$ be an admissible control defined on the interval $[0,S]$ and let $y(\tau)$ be the solution of the associated system (2.4) obtained with this control and the initial conditions (2.6). Let $p^{(u(\cdot),y(\cdot))} \in \mathcal{P}(U\times Y)$ be the occupational measure generated by the pair $\big(u(\tau),y(\tau)\big):[0,S]\to U\times Y$ and let

$$\Gamma\big(z,\epsilon_1,\dots,\epsilon_{m-1},S,y(0)\big) \stackrel{\text{def}}{=} \bigcup_{(u(\cdot),y(\cdot))} \Big\{p^{(u(\cdot),y(\cdot))}\Big\}, \tag{2.7}$$

where the union is taken over all admissible controls and the corresponding solutions of (2.4). Notice that the dependence on $(z,\epsilon_1,\dots,\epsilon_{m-1})$ in (2.7) is due to the dependence of the solutions of (2.4) on these parameters.

Assumption 2.2. *For any $z\in Z$, there exists a convex and compact set $\Gamma(z)\subset \mathcal{P}(U\times Y)$ such that*

$$\rho_{\text{H}}\Big(\Gamma\big(z,\epsilon_1,\dots,\epsilon_{m-1},S,y(0)\big),\Gamma(z)\Big) \le \nu(\epsilon_1,\dots,\epsilon_{m-1},S) \quad \forall y(0)\in Y', \tag{2.8}$$

where $\lim_{(\epsilon_1,\dots,\epsilon_{m-1},S^{-1})\to 0}\nu(\epsilon_1,\dots,\epsilon_{m-1},S)=0$.

The set $\Gamma(z)$ introduced in Assumption 2.2 will be referred to as the *limit occupational measures set* (LOMS). Some sufficient conditions for the existence of the LOMS are considered in section 3.

Assumption 2.3. *For any $S>0$, any absolutely continuous function $\tilde z(\tau):[0,S]\to Z$, and any admissible control $u(\tau)\colon[0,S]\to U$,*

$$\max_{\tau\in[0,S]}\big\|y^z(\tau)-\tilde y(\tau)\big\| \le c\max_{\tau\in[0,S]}\big\|z-\tilde z(\tau)\big\| + \kappa(\epsilon_1,\dots,\epsilon_{m-1}), \quad c=\text{const}, \tag{2.9}$$

where $y^z(\tau)$ is the solution of (2.4) obtained with a given $z\in Z$ and $\tilde y(\tau)$ is the solution of the same system obtained with the replacement of z by the function $\tilde z(\tau)$. Initial conditions for $y^z(\tau)$ and $\tilde y(\tau)$ are the same: $y^z(0)=\tilde y(0)\in Y'$ and the function $\kappa(\epsilon_1,\dots,\epsilon_{m-1})$ is either zero (for $m=1$) or tends to zero as $(\epsilon_1,\dots,\epsilon_{m-1})$ tends to zero (for $m>1$).

Lemma 2.4. *Let Assumptions* 2.1–2.3 *be satisfied. Then for any vector function $h(u,y,z):U\times Y\times Z\to\mathbb{R}^j$, $j=1,2,\dots$, which is continuous in (u,y,z) and satisfies Lipschitz conditions in (y,z), there exists a constant c_h such that*

$$\mathrm{d}_{\text{H}}\big(V_h(z'),V_h(z'')\big)\le c_h\|z'-z''\| \quad \forall z',z''\in Z, \tag{2.10}$$

where

$$V_h(z)\stackrel{\text{def}}{=}\bigcup_{p\in\Gamma(z)}\int_{U\times Y} h(u,y,z)p\,(du,dy). \tag{2.11}$$

Note that $\mathrm{d}_{\text{H}}(\cdot,\cdot)$ in (2.10) stands for the Hausdorff metric in a finite-dimensional space. That is, for arbitrary bounded subsets V_1, V_2 of $\mathbb{R}^j$ $(j=1,2,\dots)$,

$$\mathrm{d}_{\text{H}}(V_1,V_2)\stackrel{\text{def}}{=}\max\Big\{\sup_{v\in V_1}\mathrm{d}(v,V_2),\ \sup_{v\in V_2}\mathrm{d}(v,V_1)\Big\}, \quad \mathrm{d}(v,V_i)\stackrel{\text{def}}{=}\inf_{v'\in V_i}\|v-v'\|, \tag{2.12}$$

where $\|\cdot\|$ is a norm in $\mathbb{R}^j$.

The proof of Lemma 2.4 is in section 4.1.

Note that Assumption 2.3 is satisfied automatically if the functions $f_1,\dots,f_{m-1}$ defining the right-hand side of the associated systems (2.4) do not depend on z. In a

general case, Assumption 2.3 can be verified to be valid if the associated system (2.4) satisfies stability conditions similar to that introduced in [16] (see [16, Assumption 4.1, Lemma 4.1]), the latter being implied by the existence of a Lyapunov-like function (as in [17, p. 467]). For the case $m = 1$ (one singular perturbation parameter), Assumption 2.3 can be replaced by the assumption that the statement of Lemma 2.4 is valid (see [17]). A slightly different assumption which can replace Assumption 2.3 for $m > 1$ is discussed in Remark 4.1.

2.3. Approximation of the slow trajectories. Let the function $\tilde{g}(\gamma, z) : \mathcal{P}(U \times Y) \times \mathbb{R}^N \to \mathbb{R}^N$ be defined as follows:

$$\tilde{g}(\gamma, z) \stackrel{\text{def}}{=} \int_{U\times Y} g(u, y, z)\, \gamma(du, dy). \tag{2.13}$$

We will assume that the metric ρ of $\mathcal{P}(U\times Y)$ is chosen in such a way that the function $\tilde{g}(\gamma, z)$ satisfies the Lipschitz conditions:

$$\|\tilde{g}(\gamma', z') - \tilde{g}(\gamma'', z'')\| \le b(\rho(\gamma', \gamma'') + \|z' - z''\|) \quad \forall z', z'',\ \forall \gamma', \gamma'', \tag{2.14}$$

where b is a positive constant. Let us consider the system

$$\dot{z}(t) = \tilde{g}\big(\gamma(t), z(t)\big), \tag{2.15}$$

which will be referred to as the *averaged system*. The role of controls in the averaged system is played by functions $\gamma(t)$ satisfying the inclusion

$$\gamma(t) \in \Gamma\big(z(t)\big). \tag{2.16}$$

Note that the fact that the functions $\gamma(t)$ are measure valued underlines the similarity of our description with classical relaxed control setting (see [33]).

DEFINITION 2.5. *A pair* $\big(\gamma(t), z(t)\big) : [0, T] \to \mathcal{P}(U\times Y)\times\mathbb{R}^N$ *is called admissible for the averaged system if* $\gamma(t)$ *is Lebesgue measurable,* $z(t)$ *is absolutely continuous, and* (2.15)–(2.16) *are satisfied for almost all* $t \in [0, T]$.

THEOREM 2.6. *Let Assumptions* 2.1–2.3 *be satisfied and let* $h(u, y, z) : U \times Y \times Z \to \mathbb{R}^j$, $j = 1, 2, \ldots$, *be an arbitrary Lipschitz continuous vector function. There exist* $\mu(\epsilon, T)$ *and* $\mu_h(\epsilon, T)$,

$$\lim_{\epsilon\to 0} \mu(\epsilon, T) = 0, \qquad \lim_{\epsilon\to 0} \mu_h(\epsilon, T) = 0, \tag{2.17}$$

such that the following two statements are valid:

(i) *Let* $u(t)$ *be an admissible control and let* $\big(y(t), z(t)\big)$ *be the corresponding trajectory of SPCS* (2.3) *which satisfies initial condition* (2.5). *There exists an admissible pair* $\big(\gamma^a(t), z^a(t)\big)$ *of the averaged system* (2.15) *with the initial conditions* $z^a(0) = z(0)$ *such that*

$$\max_{t\in[0,T]} \big\|z(t) - z^a(t)\big\| \le \mu(\epsilon, T), \tag{2.18}$$

and also

$$\left\| \int_0^T h\big(u(t), y(t), z(t)\big) dt \ - \ \int_0^T \tilde{h}\big(\gamma^a(t), z^a(t)\big) dt \right\| \ \le \ \mu_h(\epsilon, T), \tag{2.19}$$

where

$$\tilde{h}(\gamma, z) \stackrel{\text{def}}{=} \int_{U\times Y} h(u, y, z)\, \gamma(du, dy). \tag{2.20}$$

(ii) *Conversely, let* $\big(\gamma^a(t), z^a(t)\big)$ *be an admissible pair of the averaged system* (2.15)*, which satisfies initial conditions* $z^a(0) \in Z''$*. One can construct an admissible control* $u(t)$ *such that the trajectory* $\big(y(t), z(t)\big)$ *of SPCS* (2.3) *obtained with this control and initial conditions* (2.5) $\big(z(0) = z^a(0)\big)$ *will satisfy* (2.18)–(2.19)*.*

The proof of the theorem is in section 4.1. Estimates (2.18)–(2.19) of Theorem 2.6 are not uniform with respect to the length T of the time interval. Additional assumptions are needed to make them uniform. The assumption we use in this paper is as follows.

ASSUMPTION 2.7. *There exist positive definite matrices* C*,* D *and a constant* a *such that corresponding to any* z', z'' *from* Z *and any* $\gamma' \in \Gamma(z')$ *there exists* $\gamma'' \in \Gamma(z'')$ *such that*

$$\big(\tilde{g}(\gamma', z') - \tilde{g}(\gamma'', z'')\big)^T\ C\ (z' - z'') \le -\|z' - z''\|_D^2 \tag{2.21}$$

and

$$\rho(\gamma', \gamma'') \le a\|z' - z''\|, \tag{2.22}$$

where $\|x\|_D^2$ *in* (2.21) *(and in what follows) stands for* $x^T Dx$*.*

Note that Assumption 2.7 is satisfied if the inequality (2.21) is valid for any $\gamma' = \gamma''$ and the LOMS $\Gamma(z)$ is independent of z (that is, the associated system does not depend on z).

THEOREM 2.8. *Let Assumptions* 2.1–2.3 *and* 2.7 *be satisfied. Assume also that all the admissible trajectories of averaged system* (2.15) *which start in* Z'' *do not leave* Z' *and those which start in* Z' *do not leave* $\mathrm{int} Z$ *on the infinite time horizon. Then there exist* $\mu(\epsilon)$ *and* $\mu_h(\epsilon)$*,*

$$\lim_{\epsilon\to 0} \mu(\epsilon) = 0, \qquad \lim_{\epsilon\to 0} \mu_h(\epsilon) = 0,$$

such that statements (i) *and* (ii) *of Theorem* 2.6 *remain valid with*

$$\sup_{t>0} \big\|z(t) - z^a(t)\big\| \le \mu(\epsilon) \tag{2.23}$$

replacing (2.18) *and*

$$\sup_{T>T_0} \left\| T^{-1} \int_0^T h\big(u(t), y(t), z(t)\big) dt \ - \ T^{-1} \int_0^T \tilde{h}\big(\gamma^a(t), z^a(t)\big) dt \right\| \le \mu_h(\epsilon), \quad T_0 = \text{const} \tag{2.24}$$

replacing (2.19) *for any Lipschitz continuous vector function* $h(u, y, z) : U \times Y \times Z \to \mathbb{R}^j$*,* $j = 1, 2, \ldots$*, such that the corresponding* $\tilde{h}(\gamma, z)$ *defined by* (2.20) *satisfies the Lipschitz condition*

$$\|\tilde{h}(\gamma', z') - \tilde{h}(\gamma'', z'')\| \le a_h(\rho(\gamma', \gamma'') + \|z' - z''\|) \quad \forall z', z'',\ \forall \gamma', \gamma'', \tag{2.25}$$

where a_h *is some positive constant.*

The proof of the theorem is in section 4.1.

2.4. Application to optimal control. Let $h(u,y,z) : U \times Y \times Z \to \mathbb{R}^1$ be continuous and satisfy the Lipschitz conditions in (y,z). Consider the optimal control problem

$$\inf_{(u(\cdot),y(\cdot),z(\cdot))} \left\{ \int_0^T h(u(t),y(t),z(t))dt \right\}, \tag{2.26}$$

where inf is sought over all admissible controls and trajectories of (2.3). Under the assumptions of Theorem 2.6, the optimal value of this problem converges to the optimal value of the problem

$$\inf_{(\gamma(\cdot),z(\cdot))} \left\{ \int_0^T \tilde{h}(\gamma(t),z(t))dt \right\}, \tag{2.27}$$

where $\tilde{h}(\gamma,z)$ is defined according to (2.20) and inf is over the admissible pairs of the averaged system (2.15). Near optimal controls of (2.26) can also be constructed on the basis of the solution of (2.27). These will be the controls which provide the validity of (2.18)–(2.19) for the admissible pair $(\gamma^a(t), z^a(t))$ which delivers the optimal (or near optimal) value to (2.27) (see statement (ii) of Theorem 2.6). If the assumptions of Theorem 2.8 are satisfied, then a similar approximation of a problem on the infinite time horizon with a time average criterion is possible.

In some cases the "limit" problem (2.27) can be significantly simplified with the help of the following proposition.

PROPOSITION 2.9. *Let $\phi(y_i) : Y_i \to \mathbb{R}^1$ be continuously differentiable. Then*

$$\int_{U\times Y} (\phi'(y_i))^T f_i(u,y,z)\gamma(du,dy) \;=\; 0 \quad \forall \gamma \in \Gamma(z), \tag{2.28}$$

and, in particular,

$$\int_{U\times Y} f_i(u,y,z)\gamma(du,dy) \;=\; 0 \quad \forall \gamma \in \Gamma(z), \tag{2.29}$$

where $f_i(u,y,z)$, $i = 1,\dots,m$, are the functions defining the right-hand side of (2.4).

The proof of the proposition is in section 4.1. To illustrate how this proposition can be applied let us consider the following special case. Assume that the set U is convex and the functions $f_i(u,y,z)$, $g(z,y,u)$ are linear in fast variables and controls. That is,

$$f_i(u,y,z) = \sum_{j=1}^m A_{i,j}(z)y_j + A_{i,m+1}(z)u + A_{i,m+2}(z), \quad i = 1,\dots,m, \tag{2.30}$$

$$g(u,y,z) = \sum_{j=1}^m A_{0,j}(z)y_j + A_{0,m+1}(z)u + A_{0,m+2}(z), \tag{2.31}$$

where $A_{i,j}$ are matrix functions of the corresponding dimensions. By (2.31), the averaged system is equivalent to

$$\dot{z}(t) = g(\bar{u}(t),\bar{y}(t),z(t)), \quad (\bar{u}(t),\bar{y}(t)) \in \Omega(z(t)), \tag{2.32}$$

where $\Omega(z)$ is the set of the first moments corresponding to the probability measures from the LOMS $\Gamma(z)$:

$$\Omega(z) \stackrel{\text{def}}{=} \left\{(\bar{u},\bar{y}) \mid (\bar{u},\bar{y}) = \int_{Y\times U}(u,y)\gamma(du,dy), \quad \gamma\in\Gamma(z)\right\}.$$

By (2.29) and (2.30), this set allows the representation

$$\Omega(z) = \{(\bar{u},\bar{y}) \mid f_i(\bar{u},\bar{y},z) = 0, \quad i = 1,\dots,m, \quad \bar{u}\in U\}, \tag{2.33}$$

and thus (2.32) is equivalent to the control system

$$\dot{z}(t) = g(\bar{u}(t),\psi(\bar{u}(t),z(t)),z(t)), \quad \bar{u}(t)\in U, \tag{2.34}$$

where $\bar{y} = \psi(\bar{u},z)$ is the root of the system of equations $f_i(\bar{u},\bar{y},z) = 0$, $i = 1,\dots,m$. This is a so-called reduced system and can be obtained from (2.3) via formally equating ϵ to zero. If, in addition, the function $h(u,y,z)$ used in (2.26) is convex in (u,y), then limit problem (2.27) becomes equivalent to

$$\inf_{(\bar{u}(\cdot),z(\cdot))}\left\{\int_0^T h(\bar{u}(t),\psi(\bar{u}(t),z(t)),z(t))dt\right\}, \tag{2.35}$$

where inf is over the admissible controls and corresponding trajectories of (2.34). Notice that the reasoning above is valid if Assumptions 2.1–2.3 are satisfied. It can be shown (although it is quite technical and we do not prove it in this paper) that these assumptions are satisfied if the eigenvalues of the matrices $A_{l,l}^{(l-1)}(z)$, $l = 1,\dots,m$, defined below have negative real parts for all z from a sufficiently large domain. The matrices are defined recursively for $l = 1,\dots,m$ by the equations

$$A_{i,j}^{(l)}(z) = A_{i,j}^{(l-1)}(z) - A_{i,l}^{(l-1)}(z)(A_{l,l}^{(l-1)}(z))^{-1}A_{l,j}^{(l-1)}(z) \tag{2.36}$$

($i = l+1,\dots,m$, $j = l+1,\dots,m+2$), with $A_{i,j}^{(0)}(z) \stackrel{\text{def}}{=} A_{i,j}(z)$ ($i = 1,\dots,m$, $j = 1,\dots,m+2$). Note that the condition that the matrices (2.36) have negative real parts is similar to that used in [12] to asymptotically describe the reachability set of a multiscale linear SPCS.

3. Existence of LOMS.

3.1. Approximation of the occupational measures set. Let $u(t)$ be an admissible control and let $\big(y(t),z(t)\big)$ be the corresponding admissible trajectory of SPCS (2.3) which satisfies initial conditions (2.5). Let $p^{(u(\cdot),y(\cdot),z(\cdot))}\in\mathcal{P}(U\times Y\times Z)$ be the occupational measure generated by the vector function $\big(u(\cdot),y(\cdot),z(\cdot)\big)\colon [0,T]\to U\times Y'\times Z'\subset U\times Y\times Z$ and let

$$\Gamma\big(\epsilon,T,y(0),z(0)\big) \stackrel{\text{def}}{=} \bigcup_{(u(\cdot),y(\cdot),z(\cdot))}\left\{p^{(u(\cdot),y(\cdot),z(\cdot))}\right\}, \tag{3.1}$$

where the union is taken over all admissible controls and the corresponding trajectories of SPCS (2.3). In this section, we will describe the asymptotics of this set as the vector of small parameters $\epsilon = (\epsilon_1,\dots,\epsilon_{m-1},\epsilon_m)$ tends to zero.

Let $(\gamma(t),z(t)) : [0,T]\to\mathcal{P}(U\times Y)\times Z$ be an admissible pair of the averaged system (2.15) with the initial condition

$$z(0)\in Z''. \tag{3.2}$$

Let $p^{(\gamma(t),z(t))} \in \mathcal{P}\big(\mathcal{P}(U\times Y)\times Z\big)$ be the occupational measure generated by this pair and let $\tilde{\Gamma}(T,z(0))$ be the union of the occupational measures generated by all such pairs

$$\tilde{\Gamma}\big(T,z(0)\big) \stackrel{\text{def}}{=} \bigcup_{(\gamma(\cdot),z(\cdot))} \Big\{p^{(\gamma(\cdot),z(\cdot))}\Big\}. \tag{3.3}$$

We will use $\tilde{\Gamma}(T,z(0))$ to specify the limit of (3.1) as ϵ tends to zero. To do that let us define a map $\psi(p) : p \in \mathcal{P}\big(\mathcal{P}(U\times Y)\times Z\big) \to \mathcal{P}(U\times Y\times Z)$ in such a way that for any $Q\in\mathcal{B}(U\times Y)$ and any $F\in\mathcal{B}(Z)$,

$$\psi(p)(Q\times F) = \int_{\mathcal{P}(U\times Y)\times Z} \gamma(Q)\chi_F(z)p(d\gamma,dz), \tag{3.4}$$

where $\chi_F(\cdot)$ is the indicator function of F. The integration in (3.4) is legitimate since the function

$$\gamma(Q)\chi_F(z):\ (\gamma,z)\in\mathcal{P}(U\times Y)\times Z \to [0,1] \tag{3.5}$$

is measurable with respect to $\mathcal{B}(\mathcal{P}(U\times Y)\times Z)$ (see [10, Proposition 7.25, p. 133]). Notice that for any $p\in\mathcal{P}\big(\mathcal{P}(U\times Y)\times Z\big)$ and any continuous function $h(u,y,z) : U\times Y\times Z\to\mathbb{R}^j$, $j=1,2,\dots,$

$$\int_{U\times Y\times Z} h(u,y,z)\psi(p)(du,dy,dz) = \int_{\mathcal{P}(U\times Y)\times Z} \tilde{h}(\gamma,z)p(d\gamma,dz), \tag{3.6}$$

where $\tilde{h}(\gamma,z)$ is defined by (2.20). For $p=p^{(\gamma(\cdot),z(\cdot))}$ (that is, for p being the occupational measure generated by an admissible pair $\big(\gamma(\cdot),z(\cdot)\big)$ of (2.15))

$$\int_{U\times Y\times Z} h(u,y,z)\psi\Big(p^{(\gamma(\cdot),z(\cdot))}\Big)(du,dy,dz) = \frac{1}{T}\int_0^T \tilde{h}\big(\gamma(t),z(t)\big)dt. \tag{3.7}$$

Let us now define the set $\Gamma(T,z(0))\subset\mathcal{P}(U\times Y\times Z)$ as follows:

$$\Gamma(T,z(0)) \stackrel{\text{def}}{=} \bigcup_{p\in\tilde{\Gamma}(T,z(0))} \Big\{\psi(p)\Big\} = \bigcup_{(\gamma(\cdot),z(\cdot))} \Big\{\psi\Big(p^{(\gamma(\cdot),z(\cdot))}\Big)\Big\}, \tag{3.8}$$

where the second union is taken over all admissible pairs of (2.15) satisfying initial conditions (3.2). (The second equality follows from the definition (3.3) of the set $\tilde{\Gamma}\big(T,z(0)\big)$.)

THEOREM 3.1. (i) *Let the assumptions of Theorem* 2.6 *be satisfied. Then there exists* $\nu(\epsilon,T)$, $\lim_{\epsilon\to 0}\nu(\epsilon,T)=0$, *such that*

$$\rho_{\text{H}}\Big(\Gamma\big(\epsilon,T,y(0),z(0)\big),\Gamma\big(T,z(0)\big)\Big)\le\nu(\epsilon,T)\quad\forall\big(y(0),z(0)\big)\in Y''\times Z''. \tag{3.9}$$

(ii) *Let the assumptions of Theorem* 2.8 *be satisfied and let there be a sequence* $q_k(u,y,z) : U\times Y\times Z\to\mathbb{R}^1$, $k=1,2,\dots,$ *of Lipschitz continuous functions such that it is dense in* $C(U\times Y\times Z)$ *and for any*

$$h(z,y,u) \stackrel{\text{def}}{=} (q_1(u,y,z),\dots,q_j(u,y,z)),\quad j=1,2,\dots, \tag{3.10}$$

the corresponding $\tilde{h}(\gamma, z)$ *defined by* (2.20) *satisfies Lipschitz condition* (2.25). *Then estimate* (3.9) *becomes uniform with respect to* $T \geq T_0$. *That is, there exists* $\nu(\epsilon)$, $\lim_{\epsilon \to 0} \nu(\epsilon) = 0$, *such that* $\forall T \geq T_0$,

$$\rho_{\text{H}}\Big(\Gamma\big(\epsilon, T, y(0), z(0)\big), \Gamma\big(T, z(0)\big)\Big) \leq \nu(\epsilon) \quad \forall \big(y(0), z(0)\big) \in Y'' \times Z''. \tag{3.11}$$

The proof of the theorem is in section 3.4.

3.2. LOMS of the averaged system and LOMS of the multiscale SPCS.

PROPOSITION 3.2. *Let the uniform estimate* (3.11) *be valid and let the LOMS of the averaged system* (2.15) *exist. That is, there exists the convex and compact set* $\tilde{\Gamma} \subset \mathcal{P}\big(\mathcal{P}(U \times Y) \times Z\big)$ *such that*

$$\rho_{\text{H}}\Big(\tilde{\Gamma}\big(T, z(0)\big), \tilde{\Gamma}\Big) \leq \tilde{\mu}(T) \quad \forall z(0) \in Z'', \tag{3.12}$$

where $\lim_{T \to \infty} \tilde{\mu}(T) = 0$. *Then the set*

$$\Gamma \stackrel{\text{def}}{=} \bigcup_{p \in \tilde{\Gamma}} \Big\{\psi(p)\Big\} \subset \mathcal{P}\big(U \times Y \times Z\big) \tag{3.13}$$

is convex and compact, and the following estimate is valid:

$$\rho_{\text{H}}\Big(\Gamma\big(T, z(0)\big), \Gamma\Big) \leq \mu(T) \quad \forall z(0) \in Z'', \tag{3.14}$$

where $\lim_{T \to \infty} \mu(T) = 0$. *Also,*

$$\rho_{\text{H}}\Big(\Gamma\big(\epsilon, T, y(0), z(0)\big), \Gamma\Big) \leq \mu(T) + \nu(\epsilon) \quad \forall \big(y(0), z(0)\big) \in Y'' \times Z'', \tag{3.15}$$

where $\mu(T)$ *and* $\nu(\epsilon)$ *are as in* (3.14) *and* (3.11), *respectively. Thus,* Γ *is the LOMS of SPCS* (2.3).

Proof. The validity of (3.14) is implied by (3.12) and by the fact that the map $\psi(p)$ defined by (3.4) is continuous (see Lemma 4.3 in section 4.2). This continuity implies also the fact that the set Γ is compact. The convexity of Γ follows from the linearity of $\psi(p)$. Estimate (3.15) follows from (3.14), (3.11), and the triangle inequality. ∎

THEOREM 3.3. *Let the assumptions of Theorem* 3.1(ii) *be satisfied. Then*

(i) *the LOMS* $\tilde{\Gamma}$ *of the averaged system* (2.15) *exists and the estimate* (3.12) *is valid;*

(ii) *the LOMS* Γ *of the SPCS system* (2.3) *exists and the estimate* (3.15) *is valid;* Γ *is presented in the form* (3.13).

Proof. The statements included in (ii) follow from Theorem 3.1(ii), Proposition 3.2, and Theorem 3.3(i). The proof of Theorem 3.3(i) is in section 4.2. ∎

3.3. LOMS via multistage averaging. System (2.4), which was introduced as associated with respect to (2.3), is singularly perturbed itself. One can thus consider a system which would be associated with respect to (2.4):

$$\begin{aligned}
\epsilon_1 \epsilon_2 \dots \epsilon_{m-2} \dot{y}_1(\tau) &= f_1\big(u(\tau), y_1(\tau), \dots, y_{m-1}(\tau), y_m, z\big), \\
&\vdots \quad \vdots \quad \vdots \qquad \vdots \quad \vdots \quad \vdots \\
\epsilon_{m-2} \dot{y}_{m-2}(\tau) &= f_{m-2}\big(u(\tau), y_1(\tau), \dots, y_{m-1}(\tau), y_m, z\big), \\
\dot{y}_{m-1}(\tau) &= f_{m-1}\big(u(\tau), y_1(\tau), \dots, y_{m-1}(\tau), y_m, z\big), \\
(y_m, z) &= \text{constant},
\end{aligned} \tag{3.16}$$

in which both y_m and z are fixed. For the sake of convenience, in this section we will refer to (2.4) and (3.16) as to y_m- and y_{m-1}-associated systems, respectively (by the name of the group of variables changing their values with rates of the order $O(1)$). One can also consider y_{m-2}-, $\dots$, y_2- and y_1-associated systems, the latter two being of the form

$$\begin{aligned} \epsilon_1 \dot{y}_1(\tau) &= f_1\big(u(\tau), y_1(\tau), y_2(\tau), y_3, \dots, y_m, z\big), \\ \dot{y}_2(\tau) &= f_2\big(u(\tau), y_1(\tau), y_2(\tau), y_3, \dots, y_m, z\big), \\ (y_3, \dots, y_m, z) &= \text{constant} \end{aligned} \tag{3.17}$$

and

$$\begin{aligned} \dot{y}_1(\tau) &= f_1\big(u(\tau), y_1(\tau), y_2, y_3, \dots, y_m, z\big), \\ (y_2, y_3, \dots, y_m, z) &= \text{constant}. \end{aligned} \tag{3.18}$$

Assume that the LOMS $\Gamma_1(y_2, y_3, \dots, y_m, z) \subset \mathcal{P}\big(U \times Y_1\big)$ of system (3.18) exists (sufficient conditions for the existence of LOMS of systems which, like (3.18), do not involve small parameters were discussed in [17]) and that Theorem 2.6 is applicable to system (3.17). Then y_2-components of the trajectories of this system are approximated by the trajectories of the averaged system

$$\dot{y}_2(\tau) = \tilde{f}_2\big(\gamma_1(\tau), y_2(\tau), y_3, \dots, y_m, z\big), \quad \gamma_1(\tau) \in \Gamma_1(y_2(\tau), y_3, \dots, y_m, z), \tag{3.19}$$

where $(y_3, \dots, y_m, z)$ are fixed and

$$\tilde{f}_2\big(\gamma_1, y_2, y_3, \dots, y_m, z\big) \stackrel{\text{def}}{=} \int_{U \times Y_1} f_2\big(u, y_1, y_2, y_3, \dots, y_m, z\big)\gamma_1(du, dy_1). \tag{3.20}$$

Suppose that the LOMS $\tilde{\Gamma}_2(y_3, \dots, y_m, z) \subset \mathcal{P}\big(\mathcal{P}(U \times Y_1) \times Y_2\big)$ of system (3.19) exists and that the other assumptions of Proposition 3.2 or Theorem 3.3 are satisfied. One then can come to the conclusion that the LOMS $\Gamma_2(y_3, \dots, y_m, z) \subset \mathcal{P}\big(U \times Y_1 \times Y_2\big)$ of system (3.17) exists and is presented in the form

$$\Gamma_2(y_3, \dots, y_m, z) \;=\; \bigcup_{p \in \tilde{\Gamma}_2(y_3, \dots, y_m, z)} \Big\{\psi_1(p)\Big\}, \tag{3.21}$$

where the map $\psi_1(p) : p \in \mathcal{P}\big(\mathcal{P}(U \times Y_1) \times Y_2\big) \to \mathcal{P}(U \times Y_1 \times Y_2)$ is such (compare with (3.4) above) that for any $Q \in \mathcal{B}(U \times Y_1)$ and any $F \in \mathcal{B}(Y_2)$,

$$\psi_1(p)(Q \times F) = \int_{\mathcal{P}(U \times Y_1) \times Y_2} \gamma_1(Q)\chi_F(y_2)p(d\gamma_1, dy_2), \tag{3.22}$$

$\chi_F(\cdot)$ being the indicator function of F. Assuming further that Proposition 3.2 or Theorem 3.3 can be applied step by step to y_3-, $\dots$, y_m-associated systems, one can establish the existence of the LOMS $\Gamma(z) \stackrel{\text{def}}{=} \Gamma_m(z)$ of system (2.4), which is presented in the form

$$\Gamma_m(z) \;=\; \bigcup_{p \in \tilde{\Gamma}_m(z)} \Big\{\psi_{m-1}(p)\Big\}, \tag{3.23}$$

with the corresponding definition of $\psi_{m-1}(p)$ and $\tilde{\Gamma}_m(z)$ being the LOMS of the averaged system

$$\dot{y}_m(\tau) = \tilde{f}_m\big(\gamma_{m-1}(\tau), y_m(\tau), z\big), \quad \gamma_{m-1}(\tau) \in \Gamma_{m-1}(y_m(\tau), z), \tag{3.24}$$

where $z = \text{const}$, $\Gamma_{m-1}(y_m, z)$ is the LOMS of the y_{m-1}-associated system, and

$$\tilde{f}_m\big(\gamma_{m-1}, y_m, z\big) \stackrel{\text{def}}{=} \int_{U\times Y_1\times\cdots\times Y_{m-1}} f_m\big(u, y_1, \dots, y_{m-1}, y_m, z\big)\gamma_{m-1}(du, dy_1, \dots, dy_{m-1}).$$

The applicability of Theorem 3.3 to each of the above systems is easy to verify, for example, if

$$f_i(u, y, z) \stackrel{\text{def}}{=} f_i(u, y_1, \dots, y_i), \quad i = 1, \dots, m. \tag{3.25}$$

That is, the dynamics of y_i-components in (2.3) is not influenced by the dynamics of y_{i+1}-, $\dots$, y_m- and z-components. Assuming that this is the case, let us also introduce the following assumption about the functions $f_i(\cdot)$.

ASSUMPTION 3.4. *There exist positive definite matrices C_i, D_i $(i = 1, \dots, m)$ such that for any $u \in U$ and any $y_1, \dots, y_{i-1}, y_i', y_i''$,*

$$\big(f_i(u, y_1, \dots, y_{i-1}, y_i') - f_i(u, y_1, \dots, y_{i-1}, y_i'')\big)^T\, C_i\, (y_i' - y_i'') \le -\|y_i' - y_i''\|^2_{D_i}. \tag{3.26}$$

By (3.25), the y_1-associated system (3.18) does not depend on $(y_2, \dots, y_m, z)$ and, by (3.26) with $i = 1$, the LOMS Γ_1 of this system exists (see Proposition 3.3 in [17]). Again, by (3.25), the dependence on $(y_3, \dots, y_m, z)$ in the function (3.20) defining the right-hand side of (3.19) disappears and, by (3.26) with $i = 2$, this function satisfies the inequality

$$\big(\tilde{f}_2(\gamma_1, y_2') - \tilde{f}_2(\gamma_1, y_2'')\big)^T\, C_2\, (y_2' - y_2'') \le -\|y_2' - y_2''\|^2_{D_2} \quad \forall \gamma_1 \in \mathcal{P}(U \times Y_1),$$

$\forall y_2', y_2'' \in \mathbb{R}^{M_2}$ and $\forall \gamma_1 \in \mathcal{P}(U \times Y_1)$. This implies the applicability of Theorem 3.3 according to which the LOMS $\tilde{\Gamma}_2$ of averaged system (3.19) and the LOMS Γ_2 of the y_2-associated system both exist and the representation (3.21) is valid. Continuing in a similar way, one can finally verify that the LOMS $\tilde{\Gamma}_m$ of averaged system (3.24) and the LOMS Γ_m of y_m-associated system (2.4) exist and that the representation (3.23) is valid. The applicability of Theorem 3.3 at this final stage can be verified by using the fact that the function $\tilde{f}_m\big(\gamma_{m-1}, y_m\big)$ defining the right-hand side of the averaged system (3.24) (which, by (3.25), does not involve the dependence on z) satisfies the inequality

$$\big(\tilde{f}_m(\gamma_{m-1}, y_m') - \tilde{f}_m(\gamma_{m-1}, y_m'')\big)^T\, C_m\, (y_m' - y_m'') \le -\|y_m' - y_m''\|^2_{D_m}$$

$\forall y_m', y_m'' \in \mathbb{R}^{M_m}$ and $\forall \gamma_{m-1} \in \mathcal{P}(U \times Y_1 \times \cdots \times Y_{m-1})$.

Note that a different multistage averaging procedure for SPCS with $f_i(\cdot)$ having the form (3.25) and satisfying an assumption similar to Assumption 3.4 (with C_i, D_i being identity matrices) was suggested in [20].

3.4. Basic lemma and the proof of Theorem 3.1. The proofs of Theorems 3.1 and 3.3 are based on the lemma and its corollaries presented below. Let W be a compact metric space and $q_k(w) : W \to \mathbb{R}^1$, $k = 1, 2, \dots$, be a sequence of Lipschitz continuous functions which is dense in $C(W)$.

LEMMA 3.5. *Let $\Gamma^i(\alpha, \beta) \subset \mathcal{P}(W)$, $i = 1, 2$, where α and β take values in some metric spaces $\mathcal{A}$ and $\mathcal{B}$. Assume that corresponding to any vector function*

$$h(w) = \big(q_1(w), \dots, q_j(w)\big), \quad j = 1, 2, \dots, \tag{3.27}$$

there exists a function

$$\nu_h(\alpha) : \mathcal{A} \to \mathbb{R}^1, \quad \lim_{\alpha \to \alpha_0} \nu_h(\alpha) = 0, \tag{3.28}$$

such that

$$\sup_{v \in V_h^1(\alpha,\beta)} \mathrm{d}\big(v, V_h^2(\alpha,\beta)\big) \le \nu_h(\alpha), \tag{3.29}$$

where

$$V_h^i(\alpha,\beta) \stackrel{\text{def}}{=} \bigcup_{\gamma \in \Gamma^i(\alpha,\beta)} \left\{ \int_W h(w)\gamma(dw) \right\}, \quad i = 1, 2, \dots. \tag{3.30}$$

Then there also exists another function

$$\nu(\alpha) : \mathcal{A} \to \mathbb{R}^1, \quad \lim_{\alpha \to \alpha_0} \nu(\alpha) = 0, \tag{3.31}$$

such that

$$\sup_{\gamma \in \Gamma^1(\alpha,\beta)} \rho\big(\gamma, \Gamma^2(\alpha,\beta)\big) \le \nu(\alpha). \tag{3.32}$$

Corollary 3.6. *If for any $h(w) : W \to \mathbb{R}^j$ as in* (3.27) *there exists a function* (3.28) *such that*

$$\mathrm{d}_{\mathrm{H}}\big(V_h^1(\alpha,\beta), V_h^2(\alpha,\beta)\big) \le \nu_h(\alpha), \tag{3.33}$$

then there also exists a function (3.31) *such that*

$$\rho_{\mathrm{H}}\big(\Gamma^1(\alpha,\beta), \Gamma^2(\alpha,\beta)\big) \le \nu(\alpha). \tag{3.34}$$

Corollary 3.7. *Let $\Gamma(\alpha,\beta) \subset \mathcal{P}(W)$ for $(\alpha,\beta) \in \mathcal{A} \times \mathcal{B}$, and for any $h(w) : W \to \mathbb{R}^j$ as in* (3.27) *there exists a convex and compact set $V_h \subset \mathbb{R}^j$ and a function* (3.28) *such that*

$$\mathrm{d}_{\mathrm{H}}\big(V_h(\alpha,\beta), V_h\big) \le \nu_h(\alpha), \tag{3.35}$$

where

$$V_h(\alpha,\beta) = \bigcup_{\gamma \in \Gamma(\alpha,\beta)} \left\{ \int_W h(w)\gamma(dw) \right\}. \tag{3.36}$$

Then there exists a function (3.31) *such that*

$$\rho_{\mathrm{H}}\big(\Gamma(\alpha,\beta), \Gamma\big) \le \nu(\alpha), \tag{3.37}$$

where Γ is a convex and compact subset of $\mathcal{P}(W)$ defined by

$$\Gamma \stackrel{\text{def}}{=} \left\{ \gamma \,\middle|\, \gamma \in \mathcal{P}(W), \int_W h(w)\gamma(dw) \in V_h \ \forall h(w)\colon W \to \mathbb{R}^j \text{ as in } (3.27) \right\}. \tag{3.38}$$

The proof of Lemma 3.5 is in section 4.2. Corollary 3.6 is implied by Lemma 3.5 in an obvious way. The proof of Corollary 3.7 is similar to the proof of Theorem 3.1(i) in [17].

Proof of Theorem 3.1. Let $h(u,y,z): U\times Y\times Z\to\mathbb{R}^j$, $j=1,2,\dots$, be an arbitrary Lipschitz continuous vector function. Let $u(t)$ be an admissible control and let $\big(y(t),z(t)\big)$ be the corresponding admissible trajectory of SPCS (2.3) which satisfies initial conditions (2.5). Let $V_h\big(\epsilon,T,y(0),z(0)\big)$ be the set of time averages

$$V_h\big(\epsilon,T,y(0),z(0)\big)\stackrel{\text{def}}{=}\bigcup_{(u(\cdot),y(\cdot),z(\cdot))}\left\{\frac{1}{T}\int_0^T h\big(u(t),y(t),z(t)\big)dt\right\},\tag{3.39}$$

where the union is taken over all admissible controls and the corresponding trajectories of (2.3). Notice that by definition (3.1) of $\Gamma\big(\epsilon,T,y(0),z(0)\big)$, the set (3.39) also allows the representation

$$V_h\big(\epsilon,T,y(0),z(0)\big)=\bigcup_{\gamma\in\Gamma(\epsilon,T,y(0),z(0))}\left\{\int h(u,y,z)\gamma(du,dy,dz)\right\}.\tag{3.40}$$

Let the set $\tilde V_h\big(T,z(0)\big)$ be defined as follows:

$$\begin{aligned}\tilde V_h\big(T,z(0)\big)&\stackrel{\text{def}}{=}\bigcup_{\gamma\in\Gamma(T,z(0))}\left\{\int h(u,y,z)\gamma(du,dy,dz)\right\}\\&=\bigcup_{(\gamma(\cdot),z(\cdot))}\left\{\int h(u,y,z)\psi\big(p^{(\gamma(\cdot),z(\cdot))}\big)(du,dy,dz)\right\},\end{aligned}\tag{3.41}$$

where, as in (3.8), the second union is taken over all admissible pairs of (2.15) which satisfy the initial conditions (3.2).

By (3.7), the set $\tilde V_h\big(T,z(0)\big)$ can also be represented in the form

$$\tilde V_h\big(T,z(0)\big)=\bigcup_{(\gamma(\cdot),z(\cdot))}\left\{\frac{1}{T}\int_0^T\tilde h\big(\gamma(\cdot),z(\cdot)\big)\right\}.\tag{3.42}$$

Using estimate (2.19) from Theorem 2.6 and comparing (3.39) and (3.42), one obtains

$$\mathrm{d_H}\Big(V_h\big(\epsilon,T,y(0),z(0)\big),\tilde V_h\big(T,z(0)\big)\Big)\le\frac{1}{T}\mu_h(\epsilon,T)\tag{3.43}$$

$\forall\big(y(0),z(0)\big)\in Y''\times Z''$. Having in mind representations (3.40), (3.41) and applying Corollary 3.6, one proves (3.9). Under the conditions of Theorem 2.8, estimate (3.43) can be rewritten in the uniform with respect to the $T\ge T_0$ form

$$\mathrm{d_H}\Big(V_h\big(\epsilon,T,y(0),z(0)\big),\tilde V_h\big(T,z(0)\big)\Big)\le\mu_h(\epsilon)\quad\forall T\ge T_0,$$

where $h(\cdot)$ is as in (3.10). This, by Corollary 3.6, proves (3.11). □

4. Proofs and auxiliary results.

4.1. Proofs for section 2.

Proof of Lemma 2.4. Consider the set of the time averages

$$V_h\big(z,\bar\epsilon,S,y(0)\big)\stackrel{\text{def}}{=}\bigcup_{(u(\cdot),y(\cdot))}\left\{\frac{1}{S}\int_0^S h\big(u(\tau),y(\tau),z\big)\right\},\tag{4.1}$$

where $\bar\epsilon \stackrel{\text{def}}{=} (\epsilon_1, \dots, \epsilon_{m-1})$ and the union is taken over all admissible controls and the corresponding trajectories of (2.4). By Assumption 2.3,

$$\max_{\tau\in[0,S]} \big\|y^{z'}(\tau) - y^{z''}(\tau)\big\| \le c\|z' - z''\| + \kappa(\bar\epsilon) \quad \forall z', z'' \in Z, \tag{4.2}$$

where $y^{z'}(\tau)$ and $y^{z''}(\tau)$ are solutions of (2.4) obtained with the same control and initial conditions and with $z = z'$ and $z = z''$, respectively. Hence,

$$\mathrm{d_H}\Big(V_h\big(z', \bar\epsilon, S, y(0)\big), V_h\big(z'', \bar\epsilon, S, y(0)\big)\Big) \le c_h\|z' - z''\| + c_h\kappa(\bar\epsilon) \quad \forall z', z'' \in Z, \tag{4.3}$$

where c_h is a constant which is expressed via the Lipschitz constant of $h(\cdot)$ and c from (4.2) in an obvious way.

By definition (2.7) of $\Gamma\big(z, \bar\epsilon, S, y(0)\big)$, the set $V_h\big(z, \bar\epsilon, S, y(0)\big)$ defined in (4.1) allows also the representation

$$V_h\big(z, \bar\epsilon, S, y(0)\big) = \bigcup_{p\in\Gamma(z,\bar\epsilon,S,y(0))} \Big\{\int_{U\times Y} h(u,y)p(du,dy)\Big\}. \tag{4.4}$$

It follows from Assumption 2.2 that there exists a function $\nu_h(\bar\epsilon, S)$ such that

$$\lim_{(\bar\epsilon, S^{-1})\to 0} \nu_h(\bar\epsilon, S) = 0$$

and

$$\mathrm{d_H}\Big(V_h\big(z, \bar\epsilon, S, y(0)\big), V_h(z)\Big) \le \nu_h(\bar\epsilon, S) \quad \forall z \in Z,\ \forall y(0) \in Y'. \tag{4.5}$$

Passing to the limit as $(\bar\epsilon, S^{-1})$ tends to zero in (4.3), one obtains (2.10). □

Proof of Theorem 2.6. Let $\bar g(u,y,z) \stackrel{\text{def}}{=} \big(g(u,y,z), h(u,y,z)\big)$. Consider the set of time averages

$$V\big(z, \bar\epsilon, S, y(0)\big) = \bigcup_{(u(\cdot),y(\cdot))} \Big\{\frac{1}{S}\int_0^T \bar g\big(u(\tau), y(\tau), z\big)d\tau\Big\} \subset \mathbb{R}^{N+j},$$

where, as in (4.1), the union is taken over all admissible controls and corresponding trajectories of (2.4). From Assumption 2.2 it follows (similarly to (4.5)) that there exists $\bar\nu(\bar\epsilon, S)$, $\lim_{(\bar\epsilon,S^{-1})\to 0} \bar\nu(\bar\epsilon, S) = 0$ such that

$$\mathrm{d_H}\Big(V\big(z, \bar\epsilon, S, y(0)\big), V(z)\Big) \le \bar\nu(\bar\epsilon, S) \quad \forall z \in Z,\ \forall y(0) \in Y', \tag{4.6}$$

where

$$V(z) \stackrel{\text{def}}{=} \Big\{(v,w) \mid (v,w) = \big(\tilde g(\gamma, z), \tilde h(\gamma, z)\big),\ \gamma \in \Gamma(z)\Big\} \subset \mathbb{R}^{N+j}, \tag{4.7}$$

with $\tilde g$ and $\tilde h$ being defined by (2.13) and (2.20), respectively.

Let us augment the averaged system (2.15) with the equation

$$\dot\theta(t) = \tilde h\big(\gamma(t), z(t)\big), \quad \theta(0) = 0. \tag{4.8}$$

The map $V(z)\colon Z \to 2^{\mathbb{R}^{N+j}}$ defined by (4.7) is convex and compact valued. It also satisfies the Lipschitz conditions (Lemma 2.4)

$$\mathrm{d_H}\big(V(z'), V(z'')\big) \le \bar c\|z' - z''\| \quad \forall z', z'' \in Z, \quad \bar c = \text{const}. \tag{4.9}$$

By the Filippov theorem (see, e.g., [7, Theorem 8.2.10, p. 316]), the set of admissible trajectories $\big(z(t),\theta(t)\big) \stackrel{\text{def}}{=} \bar z(t)$ of systems (2.15) and (4.8) coincides with the set of solutions of the differential inclusion

$$\dot{\bar z}(t) \in V\big(z(t)\big). \tag{4.10}$$

Let us augment system (2.3) with the equation

$$\dot\theta(t) = h\big(u(t), y_1(t), \dots, y_m(t), z(t)\big), \quad \theta(0) = 0, \tag{4.11}$$

and again denote $\bar z(\tau) \stackrel{\text{def}}{=} \big(z(\tau),\theta(\tau)\big)$. To prove the theorem it is enough to show that, corresponding to any admissible trajectory $\big(y(t),\bar z(t)\big)$ of (2.3) and (4.11), there exists a solution $\bar z^a(t)$ of (4.10) satisfying the inequality

$$\max_{t\in[0,T]} \big\|\bar z(t) - \bar z^a(t)\big\| \le \bar\mu(\epsilon,T), \quad \lim_{\epsilon\to 0}\bar\mu(\epsilon,T) = 0, \tag{4.12}$$

and, conversely, for any solution $\bar z^a(t)$ of (4.10), there exists an admissible trajectory $\big(y(\tau),\bar z(\tau)\big)$ of (2.3) and (4.11) which satisfies (4.12).

The proof of these statements is similar to Lemma 2.1 in [16] or Theorem 3.1 in [19]. □

REMARK 4.1. *Note that an important step of the proof is an introduction of the new time scale* $\tau \stackrel{\text{def}}{=} t\epsilon_m^{-1}$ *and a partition of the interval* $[0, T\epsilon_m^{-1}]$ *by the points* $\tau_l = lS(\epsilon_m)$, $l = 0, 1, \dots$, *where* $S(\epsilon_m) > 0$ *is a function of* ϵ_m *such that* $\lim_{\epsilon_m\to 0} S(\epsilon_m) = \infty$ *and* $\lim_{\epsilon_m\to 0} \epsilon_m S(\epsilon_m) = 0$. *At the cost of making the proof slightly more involved, one can replace Assumption* 2.3 *by the assumption that the statement of Lemma* 2.4 *is valid and that the* y_{m-1}*-associated system* (3.16) *has a property similar to* (2.9), *with* (y_m, z) *playing the role of* z.

Proof of Theorem 2.8. The proof is based on the following result.

PROPOSITION 4.2. *Given a solution* $\big(z^1(t),\theta^1(t)\big)$ *of the differential inclusion* (4.10) *satisfying the initial condition* $\big(z^1(0),\theta^1(0)\big) = (z^1,\theta^1) \in Z' \times \mathbb{R}^j$ *and a vector* $(z^2,\theta^2) \in Z' \times \mathbb{R}^j$, *there exists a solution* $\big(z^2(t),\theta^2(t)\big)$ *of* (4.10) *which satisfies the initial condition* $\big(z^2(0),\theta^2(0)\big) = (z^2,\theta^2)$, *and the following inequalities hold:*

$$\big\|z^1(t) - z^2(t)\big\| \le b_1 e^{-\beta t}\|z^1 - z^2\|, \tag{4.13}$$

$$\big\|\theta^1(t) - \theta^2(t)\big\| \le \|\theta^1 - \theta^2\| + b_2\|z^1 - z^2\|, \tag{4.14}$$

where b_1, b_2, β *are some positive constants.*

Proof of Proposition 4.2. As mentioned above, the map $V(z)$ defined in (4.7) is convex and compact valued and satisfies Lipschitz conditions. Also, from Assumption 2.7 (see (2.21)–(2.22)) it follows that it has the following property: for any $z' \in Z$, $(v', w') \in V(z')$ and any $z'' \in Z$, there exists $(v'', w'') \in V(z'')$ such that

$$(v' - v'')^T C(z' - z'') \le -\|z' - z''\|_D^2, \tag{4.15}$$

$$\|w' - w''\| \le b_h\|z' - z''\|. \tag{4.16}$$

The claim of the proposition follows now from Lemma A.2 in [18]. □

To prove Theorem 2.8 let us choose T_0 in such a way that

$$b_1 e^{-\beta T_0} \stackrel{\text{def}}{=} \delta < 1, \tag{4.17}$$

and let $\big(y(t), z(t), \theta(t)\big)$ be an admissible trajectory of the systems (2.3) and (4.11) which satisfies the initial conditions $(y(0), z(0)) \in Y'' \times Z''$, $\theta(0) = 0$. By Theorem 2.6, there exists a solution $\big(z^a(t), \theta^a(t)\big)$ of the differential inclusion (4.10) satisfying the initial condition $\big(z^a(0), \theta^a(0)\big) = \big(z(0), 0\big)$ such that

$$\big\|z(t) - z^a(t)\big\| \le \mu(\epsilon, T_0), \ \big\|\theta(t) - \theta^a(t)\big\| \le \mu_h(\epsilon, T_0) \quad \forall t \in [0, T_0]. \tag{4.18}$$

When Theorem 2.6 is applied again, one can establish that there exists a solution $\big(\tilde{z}^a(t), \tilde{\theta}^a(t)\big)$ of (4.10) on the interval $[T_0, 2T_0]$ such that it satisfies the initial conditions $\big(\tilde{z}^a(T_0), \tilde{\theta}^a(T_0)\big) = \big(z(T_0), \theta(T_0)\big)$ and, also, such that the following estimates are valid:

$$\big\|z(t) - \tilde{z}^a(t)\big\| \le \mu(\epsilon, T_0), \quad \big\|\theta(t) - \tilde{\theta}^a(t)\big\| \le \mu_h(\epsilon, T_0) \quad \forall t \in [T_0, 2T_0]. \tag{4.19}$$

It follows from Proposition 4.2 that the solution $\big(z^a(t), \theta^a(t)\big)$ used in (4.18) can be extended to the interval $[T_0, 2T_0]$ in such a way that for any $t \in [T_0, 2T_0]$,

$$\begin{aligned}
&\big\|\tilde{z}^a(t) - z^a(t)\big\| \le b_1 e^{-\beta(t-T_0)}\big\|z(T_0) - z^a(T_0)\big\|,\\
&\big\|\tilde{\theta}^a(t) - \theta^a(t)\big\| \le \big\|\theta(T_0) - \theta^a(T_0)\big\| + b_2\big\|z(T_0) - z^a(T_0)\big\|.
\end{aligned}$$

These along with (4.19) allow us to establish that for any $t \in [T_0, 2T_0]$,

$$\begin{aligned}
&\big\|z(t) - z^a(t)\big\| \le \mu(\epsilon, T_0) + b_1 e^{-\beta(t-T_0)}\big\|z(T_0) - z^a(T_0)\big\|,\\
&\big\|\theta(t) - \theta^a(t)\big\| \le \mu_h(\epsilon, T_0) + \big\|\theta(T_0) - \theta^a(T_0)\big\| + b_2\big\|z(T_0) - z^a(T_0)\big\|.
\end{aligned}$$

Continuing in a similar fashion, one can construct a solution of (4.10) such that the following inequalities are satisfied $\forall t \in [lT_0, (l+1)T_0]$, $l = 1, 2, \ldots$:

$$\big\|z(t) - z^a(t)\big\| \le \mu(\epsilon, T_0) + b_1 e^{-\beta(t-lT_0)}\big\|z(lT_0) - z^a(lT_0)\big\|, \tag{4.20}$$

$$\big\|\theta(t) - \theta^a(t)\big\| \le \mu_h(\epsilon, T_0) + \big\|\theta(lT_0) - \theta^a(lT_0)\big\| + b_2\big\|z(lT_0) - z^a(lT_0)\big\|. \tag{4.21}$$

It follows now from (4.17) and (4.20)–(4.21) that

$$\begin{aligned}
&\Big\|z\big((l+1)T_0\big) - z^a\big((l+1)T_0\big)\Big\| \le \mu(\epsilon, T_0) + \delta\big\|z(lT_0) - z^a(lT_0)\big\|,\\
&\Big\|\theta\big((l+1)T_0\big) - \theta^a\big((l+1)T_0\big)\Big\| \le \mu_h(\epsilon, T_0) + \big\|\theta(lT_0) - \theta^a(lT_0)\big\| + b_2\big\|z(lT_0) - z^a(lT_0)\big\|,
\end{aligned}$$

which imply that

$$\begin{aligned}
&\big\|z(lT_0) - z^a(lT_0)\big\| \le \frac{\mu(\epsilon, T_0)}{1-\delta}, \quad l = 1, 2, \ldots,\\
&\big\|\theta(lT_0) - \theta^a(lT_0)\big\| \le l\bigg(\mu_h(\epsilon, T_0) + \frac{b_2}{1-\delta}\mu(\epsilon, T_0)\bigg), \quad l = 1, 2, \ldots.
\end{aligned}$$

These and (4.20)–(4.21) lead to statement (i) of the theorem (see also the proof of Lemma 3.2 in [18]). The proof of (ii) is similar. □

Proof of Proposition 2.9. Let $\gamma \in \Gamma(z)$. By (2.8), there exist sequences $\bar{\epsilon}^k$, S^k, and $\gamma^k \in \Gamma(z, \bar{\epsilon}^k, S^k, y(0))$ such that $(\bar{\epsilon}^k, (S^k)^{-1}) \to 0$ and $\gamma^k \to \gamma$ as k tends to infinity. The latter convergence is in the metric consistent with the weak convergence topology of $\mathcal{P}(U \times Y)$ and, hence, it implies in particular that

$$\lim_{k\to\infty} \int_{U\times Y} (\phi\prime(y_i))^T f_i(u, y, z)\gamma^k(du, dy) \ = \ \int_{U\times Y} (\phi\prime(y_i))^T f_i(u, y, z)\gamma(du, dy). \tag{4.22}$$

According to the definition of the set $\Gamma(z, \bar{\epsilon}^k, S^k, y(0))$ (see (2.7)) and the fact that $\gamma^k \in \Gamma(z, \bar{\epsilon}^k, S^k, y(0))$, there exists an admissible control $u^k(\tau)$ and the corresponding trajectory $y^k(\tau)$ of system (2.4) such that

$$\int_{U\times Y} (\phi\prime(y_i))^T f_i(u,y,z)\gamma^k(du,dy) \;=\; \frac{1}{S^k}\int_0^{S^k} (\phi\prime(y_i^k(\tau)))^T f_i(u^k(\tau), y^k(\tau), z)d\tau.$$

The second integral is apparently equal to $\frac{\phi(y_i^k(S^k))-\phi(y_i^k(0))}{S^k}$, which tends to zero as S^k tends to infinity (since, by Assumption 2.1, the solutions of (2.4) stay in the bounded area). This and (4.22) imply the validity of the proposition. □

4.2. Proofs for section 3.

LEMMA 4.3. *The map $\psi(p)$ defined by* (3.4) *is continuous. That is, $\psi(p_l)$ converges to $\psi(p)$ in the weak convergence topology of $\mathcal{P}(U\times Y\times Z)$ if p_l converges to p in the weak convergence topology of $\mathcal{P}\big(\mathcal{P}(U\times Y)\times Z\big)$.*

Proof of Lemma 4.3. Let $h(u,y,z)\colon U\times Y\times Z\to\mathbb{R}^1$ be a continuous function. Then

$$\lim_{p_l\to p}\int h(u,y,z)\psi(p_l)(du,dy,dz) = \lim_{p_l\to p}\int\left(\int h(u,y,z)\gamma(du,dy)\right)p_l(d\gamma,dz)$$
$$= \int\left(\int h(u,y,z)\gamma(du,dy)\right)p(d\gamma,dz) = \int h(u,y,z)\psi(p)(du,dy,dz),$$

where it is taken into account that, because $h(u,y,z)$ is continuous, it follows that the function $\tilde{h}(\gamma,z)$ defined by (2.20) is continuous as well. Since the last equalities are valid for any continuous $h(\cdot)$, it follows that $\lim_{p_l\to p}\psi(p_l)=\psi(p)$. □

Proof of Lemma 3.5. Let the metric ρ of $\mathcal{P}(W)$ be defined as follows:

$$\rho(\gamma',\gamma'') = \sum_{k=0}^{\infty}\frac{1}{2^k}\,\frac{|\langle\gamma',q_k\rangle-\langle\gamma'',q_k\rangle|}{1+|\langle\gamma',q_k\rangle-\langle\gamma'',q_k\rangle|}\quad\forall\gamma',\gamma''\in\mathcal{P}(W) \tag{4.23}$$

where $q_k: W\to\mathbb{R}^1$, $k=0,1,\dots$, is a sequence of Lipschitz continuous functions which is dense in the space of continuous functions $C(W)$ and $\langle\gamma,q_k\rangle=\int_W q_k(w)\gamma(dw)$. Note that this metric is consistent with the weak convergence topology of $\mathcal{P}(W)$. Define

$$\nu(\alpha)\stackrel{\text{def}}{=}\sup_{\beta\in\mathcal{B}}\sup_{\gamma\in\Gamma^1(\alpha,\beta)}\rho\big(\gamma,\Gamma^2(\alpha,\beta)\big) \tag{4.24}$$

and show that $\nu(\alpha)$ tends to zero as α tends to α_0. Assume that it does not. Then there exists a number $\delta>0$ and sequences $(\alpha_l,\beta_l)\in\mathcal{A}\times\mathcal{B}$, $\gamma^l\in\Gamma^1(\alpha_l,\beta_l)$, $l=1,2,\dots$, such that $\lim_{l\to\infty}\alpha_l=\alpha_0$ and $\rho(\gamma^l,\gamma)\ge\delta\;\forall\,\gamma\in\Gamma^2(\alpha,\beta)$. That is,

$$\sum_{k=0}^{\infty}\frac{1}{2^k}\,\frac{|\langle\gamma^l,q_k\rangle-\langle\gamma,q_k\rangle|}{1+|\langle\gamma_l,q_k\rangle-\langle\gamma,q_k\rangle|}\ge\delta\quad\forall\,\gamma\in\Gamma^2(\alpha,\beta). \tag{4.25}$$

Hence, for some integer K,

$$\sum_{k=1}^{K}|\langle\gamma^l,q_k\rangle-\langle\gamma,q_k\rangle|\ge\frac{\delta}{2}\quad\forall\gamma\in\Gamma^2(\alpha,\beta). \tag{4.26}$$

Let $h(w) \stackrel{\text{def}}{=} \big(q_1(w), \dots, q_k(w)\big) \colon W \to \mathbb{R}^K$. Assume that the norm of a vector in (2.12) is defined as the sum of the absolute values of its components. Then, by (3.30), one can rewrite (4.26) in the form

$$\mathrm{d}(v^l, v) \geq \frac{\delta}{2} \quad \forall v \in V_h^2(\alpha_l, \beta_l), \text{ where } v^l \stackrel{\text{def}}{=} \int_W h(w)\gamma^l(dw) \in V_h^1(\alpha_l, \beta_l).$$

Hence, $\mathrm{d}\big(v^l, V_h^2(\alpha_l, \beta_l)\big) \geq \frac{\delta}{2}$, $l = 1, 2, \dots$, which contradicts (3.29) and thus proves the lemma. □

Proof of Theorem 3.3(i). Let $\tilde{h}(\gamma, z) \colon \mathcal{P}(U \times Y) \times Z \to \mathbb{R}^j$, $j = 1, 2, \dots$, be an arbitrary Lipschitz continuous vector function. That is,

$$\|\tilde{h}(\gamma', z') - \tilde{h}(\gamma'', z'')\| \leq c_{\tilde{h}}(\|z' - z''\| + \rho(\gamma', \gamma'')), \quad c_{\tilde{h}} = \text{const}. \tag{4.27}$$

Consider a set-valued map $V(z)$ defined by (4.7) with $\tilde{h}(\gamma, z)$ as above. Note that this map is not necessarily convex valued since $\tilde{h}(\gamma, z)$ may not be represented as the integral (2.20). By (2.14), (4.27), and (2.22) (see Assumption 2.7), it satisfies Lipschitz conditions (4.9). Hence, by the relaxation theorem (see, e.g., [7, Theorem 10.4.4, p. 402]), the set of solutions of the differential inclusion (4.10) is dense in the set of solutions of the differential inclusion

$$\dot{\bar{z}}(t) \in \mathrm{co}V\big(z(t)\big), \tag{4.28}$$

where $\mathrm{co}V(z)$ is the convex hull of $V(z)$.

By Corollary 3.7, to establish the existence of a convex and compact set $\tilde{\Gamma} \subset \mathcal{P}(\mathcal{P}(U \times Y) \times Z)$ satisfying (3.12) it is enough to show that for any Lipschitz continuous $\tilde{h}(\gamma, z) \colon \mathcal{P}(U \times Y) \times Z \to \mathbb{R}^j$, $j = 1, 2, \dots$, there exist a convex and compact set $V_{\tilde{h}} \subset \mathbb{R}^j$ and a function $\mu_{\tilde{h}}(T)$ such that

$$\mathrm{d_H}\Big(V_{\tilde{h}}\big(T, z(0)\big), V_{\tilde{h}}\Big) \leq \mu_{\tilde{h}}(T) \quad \forall z(0) \in Z'', \quad \lim_{T \to \infty} \mu_{\tilde{h}}(T) = 0, \tag{4.29}$$

where

$$\begin{aligned} V_{\tilde{h}}\big(T, z(0)\big) &= \bigcup_{p \in \tilde{\Gamma}(T, z(0))} \left\{ \int_{\mathcal{P}(U \times Y) \times Z} \tilde{h}(\gamma, z) p(d\gamma, dz) \right\} \\ &= \bigcup_{(\gamma(\cdot), z(\cdot))} \left\{ \frac{1}{T} \int_0^T \tilde{h}\big(\gamma(t), z(t)\big) dt \right\}, \end{aligned} \tag{4.30}$$

with the second union being taken over all admissible pairs of averaged system (2.15). The closure of the set (4.30), $\mathrm{cl}V_{\tilde{h}}\big(T, z(0)\big)$, allows also the representations

$$\mathrm{cl}V_{\tilde{h}}\big(T, z(0)\big) = \mathrm{cl} \bigcup_{\bar{z}(\cdot)} \left\{ \frac{\theta(T)}{T} \right\} = \bigcup_{\bar{z}(\cdot)} \left\{ \frac{\theta(T)}{T} \right\}, \tag{4.31}$$

where the first union is taken over the solutions of (4.10) and the second over the solutions of (4.28), which satisfy the initial conditions $\bar{z}(0) = (z(0), 0)$.

As in the proof of Proposition 4.2, from Assumption 2.7 it follows that for any $z' \in Z$, $(v', w') \in V(z')$, and $z'' \in Z$, there exists $(v'', w'') \in V(z'')$ such that (4.15)–(4.16) are satisfied. It can be verified that the map $\mathrm{co}V(z)$ has a similar property. That is, for any $z' \in Z$, $(v', w') \in \mathrm{co}V(z')$, and $z'' \in Z$, there exists

$(v'', w'') \in \mathrm{co}V(z'')$ such that (4.15)–(4.16) are satisfied. As with Proposition 4.2, this allows us to establish that, given a solution $\big(z^1(t), \theta^1(t)\big)$ of the differential inclusion (4.28) satisfying the initial condition $\big(z^1(0), \theta^1(0)\big) = (z^1, \theta^1) \in Z' \times \mathbb{R}^j$ and a vector $(z^2, \theta^2) \in Z' \times \mathbb{R}^j$, there exists a solution $\big(z^2(t), \theta^2(t)\big)$ of (4.28) which satisfies the initial condition $\big(z^2(0), \theta^2(0)\big) = (z^2, \theta^2)$ such that estimates (4.13)–(4.14) will be valid.

It follows from (4.14) that

$$\mathrm{d_H}\big(\mathrm{cl}V_{\tilde{h}}(T, z^1), \mathrm{cl}V_{\tilde{h}}(T, z^2)\big) \leq b_2 T^{-1} \quad \forall z^i \in Z',\ i = 1, 2, \quad \forall T \geq 0. \tag{4.32}$$

Now applying results from [15] or [19, Proposition 3.2], one can establish the existence of a convex and compact set $V_{\tilde{h}}$ and a function $\mu_{\tilde{h}}(T) = O(T^{-1/2})$ which satisfy (4.29). This completes the proof of the theorem. □

REFERENCES

[1] O. Alvarez and M. Bardi, *Viscosity solutions methods for singular perturbations in deterministic and stochastic controls*, SIAM J. Control Optim., 40 (2001), pp. 1159–1188.

[2] Z. Artstein, *Invariant measures of differential inclusions*, J. Differential Equations, 152 (1999), pp. 289–307.

[3] Z. Artstein, *The chattering limit of singularly perturbed optimal control problems*, in Proceedings of CDC-2000, Control and Decision Conference, Sydney, 2000, pp. 564–569

[4] Z. Artstein and V. Gaitsgory, *Tracking fast trajectories along a slow dynamics: A singular perturbations approach*, SIAM J. Control Optim., 35 (1997), pp. 1487–1507.

[5] Z. Artstein and A. Vigodner, *Singularly perturbed ordinary differential equations with dynamical limits*, Proc. Roy. Soc. Edinburgh Sect. A, 126 (1996), pp. 541–569.

[6] J.-P. Aubin, *Viability Theory*, Birkhäuser, Boston, 1991.

[7] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser, Boston, 1990.

[8] F. Bagagiolo and M. Bardi, *Singular perturbation of a finite horizon problem with state-space constraints*, SIAM J. Control Optim., 36 (1998), pp. 2040–2060.

[9] A. Bensoussan, *Perturbations Methods in Optimal Control Problems*, Wiley, New York, 1984.

[10] D.P. Bertsekas and S.E. Shreve, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.

[11] T. Donchev and I. Slavov *Averaging method for one-sided Lipschitz differential inclusions with generalized solutions*, SIAM J. Control Optim., 37 (1999), pp. 1600–1613.

[12] A. Dontchev, *Time-scale decomposition of the reachable set of constrained linear systems*, Math. Control Signals Systems, 5 (1992), pp. 327–340.

[13] A. Dontchev, T. Donchev, and I. Slavov, *A Tichonov-type theorem for singularly perturbed differential inclusions*, Nonlinear Analysis, 26 (1996), pp. 1547–1554.

[14] O.P. Filatov and M.M. Hapaev, *Averaging of Systems of Differential Inclusions*, Moscow University Publishing House, Moscow, 1998 (in Russian).

[15] V. Gaitsgory, *Use of the averaging method in control problems*, Differential Equations, 22 (1986), pp. 1290–1299 (translated from Russian).

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

[17] V. Gaitsgory and A. Leizarowitz, *Limit occupational measures set for a control system and averaging of singularly perturbed control systems*, J. Math. Anal. Appl., 233 (1999), pp. 461–475.

[18] V. Gaitsgory and M.-T. Nguyen, *Averaging of three time scale singularly perturbed control systems*, Systems Control Lett., 42 (2001), pp. 395–403.

[19] G. Grammel, *Averaging of singularly perturbed systems*, Nonlinear Analysis, 28 (1997), pp. 1855–1865.

[20] G. Grammel, *Order Reduction for Nonlinear Systems by Re-iterated Averaging*, in Proceedings of the IEEE Singapore International Symposium on Control Theory and Applications, Singapore, 1997, pp. 1855–1865.

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

[22] P.V. Kokotović, H.K. Khalil, and J. O'Reilly, *Singular Perturbation Methods in Control: Analysis and Design*, Academic Press, London, 1986.

[23] H. Kushner, *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, Birkhäuser, Boston, 1990.
[24] R.E. O'Malley, Jr., *Introduction to Singular Perturbations*, Academic Press, New York, 1974.
[25] R.E. O'Malley, Jr., *Singular perturbations and optimal control*, in Mathematical Control Theory, Lecture Notes in Math. 680, W.A. Copel, ed., Springer-Verlag, Berlin, 1978, pp. 170–218.
[26] Z. Pan and T. Basar, *Multi-time scale zero-sum differential games with perfect state measurements*, Dynam. Control, 5 (1995), pp. 7–29.
[27] V.A. Plotnikov, *Differential Equations with Multivalued Right-Hand Sides: Asymptotic Methods*, AstroPrint, Odessa, 1999 (in Russian).
[28] M. Quincampoix and H. Zhang, *Singular perturbations in non-linear optimal control systems*, Differential Integral Equations, 8 (1995), pp. 931–944.
[29] M. Quincampoix and M. Veliov, *Open-loop viable control under uncertain initial state information*, Set-Valued Anal., 7 (1999), pp. 55–87.
[30] A.B. Vasil'eva and A.F. Butuzov, *Asymptotic Expansions of Solutions of Singularly Perturbed Equations*, Nauka, Moscow, 1973 (in Russian).
[31] V. Veliov, *A generalization of Tikhonov theorem for singularly perturbed differential inclusions*, J. Dynam. Control Systems, 3 (1997), pp. 291–319.
[32] A. Vigodner, *Limits of singularly perturbed control problems with statistical dynamics of fast motions*, SIAM J. Control Optim., 35 (1997), pp. 1–28.
[33] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
[34] G.G. Yin and Q. Zhang, *Continuous-Time Markov Chains and Applications. A Singular Perturbations Approach*, Springer-Verlag, New York, 1998.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 3, pp. 975–990

A STOCHASTIC DECENTRALIZED CONTROL PROBLEM WITH NOISY COMMUNICATION*

KAMBIZ SHOARINEJAD†, JASON L. SPEYER‡, AND IOANNIS KANELLAKOPOULOS§

Abstract. A simple decentralized stochastic control problem is considered where the nonclassical nature of the information pattern is induced by the uncertainty on the information transmission in the system. This is, in fact, a reformulation of the Witsenhausen counterexample, where the first station is allowed to send its information to the second station through a noisy channel. Nonconvexity of the problem in this new formulation has been established, and it is shown how this formulation relates to a classical problem and the Witsenhausen problem, respectively, when the transmission noise intensity goes to zero or infinity. Assuming small transmission noise intensity, we then use an asymptotic approach in order to find an approximated cost. A necessary condition for asymptotically optimal strategies has been obtained using a variational approach, and it is shown that the linear strategies, with slightly different coefficients than the noiseless transmission case, satisfy the necessary condition.

Key words. optimal stochastic control, decentralized systems, asymptotic analysis

AMS subject classifications. 93E20, 93A14

PII. S0363012901385629

1. Introduction.

Coordinating and controlling dynamic systems in spatial networks has always been a challenging problem for system designers. It is now attracting more attention as various new applications are emerging in a very wide range from autonomous vehicles in formation to flow and congestion control in computer networks. However, there are still some major difficulties in dealing with such systems. The main characteristics of any decentralized system is that the information is distributed among different stations and the performance of the system depends highly on the corresponding information pattern, i.e., *who knows what and when*. The stations may communicate with each other possibly by signaling through noisy channels. Even though there might be some physical constraints on the information structure of the system (e.g., locations of the sensors, the actuators, and the transmitters), in general, an optimal information pattern should be obtained. Then, based on the locally available information, a set of coordinated local strategies should be designed in order to achieve a common objective. In many cases, however, we will end up with nonconvex functional optimization problems, which are usually very difficult to solve.

One such class of problems is when a decentralized system has a *nonclassical* information pattern which is not partially nested. The information pattern is called nonclassical when the distributed stations do not have access to the same information and/or some stations do not have perfect recall (i.e., they lose information). Moreover,

*Received by the editors February 26, 2001; accepted for publication (in revised form) March 2, 2002; published electronically September 19, 2002. The results in this paper were partially presented at the 38th IEEE Conference on Decision and Control and the 1999 American Control Conference. This research was supported in part by the National Science Foundation under grant ECS-9502945, the Air Force Office of Scientific Research under grant F49620-97-1-0272, and the Office of Naval Research under award N00014-97-1-0939.

http://www.siam.org/journals/sicon/41-3/38562.html

†Innovics Wireless Inc., 11500 Olympic Boulevard, Suite 398, Los Angeles, CA 90064 (kambiz@innovics.wireless.com).

‡UCLA Mechanical and Aerospace Engineering, Box 951597, Los Angeles, CA 90095-1597 (speyer@seas.ucla.edu).

§Voyan Technology, 3255-7 Scott Boulevard, Santa Clara, CA 95054 (ioannis@voyan.com).

a nonclassical information pattern is not partially nested when some stations cannot reconstruct the previous actions of other stations which have affected their own local information. Unfortunately, this happens in many decentralized systems.

In 1968, Witsenhausen provided a simple example in [1] in which there are only two stations, the dynamics are linear, the underlying uncertainties are additive and Gaussian, and the cost is quadratic. The information pattern, however, is nonclassical. This example motivated much research on the links between decentralized stochastic control problems and team theory and the effects of different information patterns on decentralized systems. Although it is a very simple example, it demonstrates the main difficulties induced by nonclassical information patterns. In this example, one station acts first and affects the information available to the next station, while there is no way for the second station to determine the action of the first station. The existence of the optimal design was established in [1], where a nonlinear set of strategies was also proposed which showed that no affine strategy could be optimal.

This seemingly simple example, which is also called Witsenhausen's counterexample, turned out to be extremely hard. It is still outstanding after more than 30 years. It was later shown in [2] that when the uncertainty on the information available to the first station is small, linear strategies would still be optimal over a large class of nonlinear strategies. Intuitively, when the uncertainty on the information of the first station is small, the second station will also be able to *guess* what that information was. Therefore, since the problem is cooperative in the sense that the stations are aware of each others' strategies, the second station can almost reconstruct the action of the first station, and there is no need for any kind of signaling among the stations through the dynamics of the system. In Witsenhausen's problem, the nonclassical nature of the information pattern is a result of the fact that the information available to the first station is completely inaccessible for the second station. However, recent advances in computing and communication technologies make it possible for the stations in many decentralized systems to communicate different pieces of information. But communications can never be perfect, and there is always some uncertainty involved. Unfortunately, such uncertainty will again induce a nonclassical nature on the information pattern of the system.

In this paper, we reformulate Witsenhausen's problem by allowing the first station to communicate its information with the second station through a noisy channel. Then we show that as long as there is noise in the transmission, the main difficulties will persist. Specifically, the cost might still be nonconvex with respect to the strategies. We then consider the two limit cases where the transmission uncertainty becomes either very large or negligible. We show how this new formulation covers a wide range of problems, from the classical linear quadratic Gaussian (LQG) problem to the Witsenhausen counterexample.

When the transmission noise intensity is small, one would expect the optimal strategies to be very close to the corresponding strategies for the noiseless transmission case. Our next objective in this paper is to investigate this case through an asymptotic analysis.

In section 2, we present the problem formulation. In section 3, we obtain an alternative form for the performance index, which clearly shows the possible nonconvexity of the cost with respect to the strategies. In section 4, we consider the two limit cases, i.e., when the transmission noise intensity goes to zero or infinity. In section 5 we assume a small uncertainty on the transmission and approximate the cost by expanding it in terms of the small transmission noise intensity. In section 6, we use a variational approach in order to find a necessary condition for the strategies that

minimize the approximated cost. As we shall see, we will actually have a singular optimization problem. We then show that the asymptotically optimal strategies can still be linear with slightly different coefficients than the corresponding strategies for the noiseless transmission case. We provide concluding remarks in the final section.

2. Problem description. Consider a two-stage stochastic problem with the following state equations:

$$x_1 = x_0 + u_1, \tag{2.1}$$

$$x_2 = x_1 - u_2, \tag{2.2}$$

where x_0 is the initial state, which is assumed to be a zero mean Gaussian random variable with variance σ_0^2. The information pattern of the system is specified by the following output equations:

$$z_1 = x_0, \tag{2.3}$$

$$z_2 = \begin{bmatrix} x_0 + v_t \\ x_0 + u_1 + v_2 \end{bmatrix} := \begin{bmatrix} z_{21} \\ z_{22} \end{bmatrix}, \tag{2.4}$$

where v_2 is the measurement noise for the second station, which is also assumed to be a zero mean Gaussian random variable with unit variance. As we can see, the information available to the first station is being transmitted to the second station, and the communication uncertainty is modeled by an additive Gaussian noise $v_t \sim \mathcal{N}\left(0, \epsilon^2\right)$. Also, x_0, v_2, and v_t are all assumed to be independent of each other. It is clear that we have simply modeled the received information signal as the transmitted signal plus the Gaussian transmission noise. While this model can be quite realistic for analog communication systems, it may not be well justified when digital communication is used. In digital communication systems the signal is quantized, coded, and sent through the channel. Still, the channel noise may realistically be assumed to be additive and Gaussian, but sophisticated modulation and coding schemes make it difficult to assume a simple additive Gaussian uncertainty for the received *information signal.* However, if we try to incorporate the quantization effects along with the bit error probability distribution for some *good* coding and modulation schemes in order to model the communication uncertainties, we will end up with models which could still be approximated, to some degree, by simple additive Gaussian models. Moreover, since there are already major difficulties in dealing with decentralized nonclassical information patterns, using more complex models for communication uncertainties may not seem very reasonable at this point. Furthermore, we believe that the results obtained under such a simplifying assumption would still serve as a guideline for finding the true nature of the optimal decentralized strategies. The objective is now to design the control strategies γ_1 and γ_2,

$$u_1 = \gamma_1\left(z_1\right), \tag{2.5}$$

$$u_2 = \gamma_2\left(z_2\right), \tag{2.6}$$

in order to minimize the cost function

$$J = E\left[k^2 u_1^2 + x_2^2\right], \tag{2.7}$$

where $k^2 > 0$ is a given constant. Note that this is a sequential stochastic control problem in the sense that the second station acts after the first station. In other words,

the order in which the stations apply their control actions does not depend on the uncertainties in the system. We see that the first controller has perfect information but its action is costly. In contrast, the second controller has inexpensive control but noisy information. Since the second station does not know what the first station knew, due to the transmission noise, we do not have perfect recall, and hence we still have a nonclassical pattern. If there was no transmission noise, we would have a classical information pattern for which the unique optimal strategies are known to be linear in the information.

3. An alternative form for the performance index. In this section, we show how the performance index may be expressed in terms of the Fisher information matrix, which indicates that the cost may not be convex in the strategies.

For simplicity, and similarly to the Witsenhausen problem, we define

$$f(z_1) := z_1 + \gamma_1(z_1) = x_0 + u_1, \tag{3.1}$$

$$g(z_2) := \gamma_2(z_2) = u_2. \tag{3.2}$$

Then the cost can be expressed as

$$\begin{aligned} J &= E\left[k^2 u_1^2 + x_2^2\right] \\ &= E\left[k^2 (z_1 - f(z_1))^2 + (f(z_1) - g(z_2))^2\right] \\ &:= J(f, g). \end{aligned} \tag{3.3}$$

If we fix the function f, the optimal strategy g will clearly be obtained as the conditional expectation, i.e.,

$$g^*(z_2) = \arg\min_g J(f, g) = E[f(z_1) \,|\, z_2]. \tag{3.4}$$

Substituting the above equation back in the cost, we get

$$\begin{aligned} J^*(f) &:= J(f, g^*) \\ &= k^2 E\left[(z_1 - f(z_1))^2\right] + E\left[(f(z_1) - g^*(z_2))^2\right] \\ &= k^2 E\left[(z_1 - f(z_1))^2\right] + E\left[(f(z_1))^2\right] - E\left[(g^*(z_2))^2\right], \end{aligned} \tag{3.5}$$

where we have used the orthogonality property of the conditional expectation

$$E[(f(z_1) - g^*(z_2))\, g^*(z_2)] = 0. \tag{3.6}$$

It is important to note the minus sign in the third term in (3.5). As we shall see, this minus sign could indeed destroy the convexity of the cost with respect to the strategies.

The objective is now to express the cost $J^*(f)$ in terms of only one strategy f. In doing so, we use the following lemma, which shows how $g^*(z_2)$ may be expressed in terms of information z_2 and its probability density function.

LEMMA 3.1. *The optimal strategy $g^*(z_2)$ can be expressed as*

$$g^*(z_2) = z_{22} + \frac{\partial}{\partial z_{22}} \ln p(z_2), \tag{3.7}$$

where $p(z_2) = p(z_{21}, z_{22})$ is the probability density function for the information available to the second station.

Proof. We have

$$\begin{aligned} g^*(z_2) &= \int f(z_1)\, p(z_1 | z_2)\, dz_1 \\ &= \frac{\int f(z_1)\, p(z_1, z_2)\, dz_1}{\int p(z_1, z_2)\, dz_1}, \end{aligned} \tag{3.8}$$

where $p(z_1, z_2)$ is the joint probability density of z_1 and z_2. At the same time, one can write

$$f(z_1)\, p(z_1, z_2) = z_{22} p(z_1, z_2) + \frac{\partial}{\partial z_{22}} p(z_1, z_2). \tag{3.9}$$

This can be shown as

$$\begin{aligned} z_{22} p(z_1, z_2) + \frac{\partial}{\partial z_{22}} p(z_1, z_2) &= z_{22} p(z_1, z_2) + \frac{\partial}{\partial z_{22}} p(z_2 | z_1)\, p(z_1) \\ &= z_{22} p(z_1, z_2) + \frac{\partial}{\partial z_{22}} p_{(v_t, v_2)} \left(\begin{bmatrix} z_{21} \\ z_{22} \end{bmatrix} - \begin{bmatrix} z_1 \\ f(z_1) \end{bmatrix} \right) p(z_1) \\ &= z_{22} p(z_1, z_2) + \frac{\partial}{\partial z_{22}} \left(\frac{1}{2\pi\epsilon} \exp\left(-\frac{(z_{21} - z_1)^2}{2\epsilon^2} - \frac{(z_{22} - f(z_1))^2}{2} \right) \right) p(z_1) \\ &= f(z_1)\, p(z_1, z_2), \end{aligned} \tag{3.10}$$

where we have used the specific form of the information available to the second station and the fact that $v_t \sim \mathcal{N}(0, \epsilon^2)$ and $v_2 \sim \mathcal{N}(0, 1)$ are independent. By substituting for $f(z_1)\, p(z_1, z_2)$ from (3.9) back in (3.8) and integrating with respect to z_1, the expression in (3.7) is obtained. □

As we shall see, when we try to express the performance index in terms of only a single strategy f, a *Fisher information* term comes up in the cost. Fisher information is originally obtained in the Cramer–Rao bound, which is a measure for the minimum error in estimating a parameter based on the value of a random variable. However, by introducing a location parameter, an alternative form of the Fisher information may be defined for a random variable with a given distribution. This alternative form is, in fact, related to the entropy measure (see [3, p. 494]). We first present the definition for the Fisher information matrix.

DEFINITION 3.2. *The Fisher information matrix for a random vector Z is defined as*

$$I_f(Z) := E\left[\nabla_z^T \ln p(z) \cdot \nabla_z \ln p(z) \right], \tag{3.11}$$

where $p(z)$ is the probability density function for the random variable Z and ∇_z denotes the gradient vector with respect to z:

$$\nabla_z := \left[\frac{\partial}{\partial z_1} \cdots \frac{\partial}{\partial z_n} \right], \tag{3.12}$$

where z_i is the ith component in the random vector.

We are now ready to present the alternative expression for the performance index.

THEOREM 3.3. *The performance index* (3.5) *can be written as*

$$J^*(f) = k^2 E\left[(z_1 - f(z_1))^2 \right] + 1 - I_f(Z_2)_{22}, \tag{3.13}$$

where $I_f\,(Z_2)_{22}$ *is, in fact, the* $(2,2)$ *element of the* Fisher information matrix *for the random vector* Z_2*. The subscript* f *indicates the fact that it actually depends on the form of the strategy* f*, which is present in the definition of* z_2 *and would affect its probability density function.*

Proof. Using (3.7), we first obtain $E[(g^*(z_2))^2]$. We have

$$E\left[z_{22}^2\right] = E\left[\left(f\left(z_1\right)\right)^2\right] + 1, \tag{3.14}$$

and

$$E\left[z_{22}\frac{\partial}{\partial z_{22}}\ln p\left(z_2\right)\right] = \int\!\!\int_{-\infty}^{+\infty} z_{22}\frac{\partial}{\partial z_{22}}\ln\left(p\left(z_{21},z_{22}\right)\right)p\left(z_{21},z_{22}\right)dz_{21}dz_{22}. \tag{3.15}$$

If we integrate by parts with respect to z_{22}, we get

$$\begin{aligned}\int_{-\infty}^{+\infty} z_{22}\frac{\partial}{\partial z_{22}}\ln\left(p\left(z_{21},z_{22}\right)\right)p\left(z_{21},z_{22}\right)dz_{22} &= z_{22}p\left(z_{21},z_{22}\right)|_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} p\left(z_{21},z_{22}\right)dz_{22}\\ &= -p\left(z_{21}\right),\end{aligned} \tag{3.16}$$

where z_{22} is assumed to have a finite mean value, and therefore the first term becomes zero. Hence,

$$E\left[z_{22}\frac{\partial}{\partial z_{22}}\ln p\left(z_2\right)\right] = -1. \tag{3.17}$$

Therefore,

$$E\left[\left(g^*\left(z_2\right)\right)^2\right] = -1 + E\left[\left(f\left(z_1\right)\right)^2\right] + I_f\left(Z_2\right)_{22}, \tag{3.18}$$

where

$$I_f\left(Z_2\right)_{22} = E\left[\left(\frac{\partial}{\partial z_{22}}\ln p\left(z_2\right)\right)^2\right]. \tag{3.19}$$

Substituting (3.18) back in (3.5), we get (3.13) as an alternative form for representing the performance index. ∎

As we see, the cost is now expressed only in terms of one strategy f. Also, this somehow shows us that in order to minimize the cost, we need to get the lowest possible cost associated with the first station, while we transfer as much information as possible to the second station through the dynamics of the system. The possible nonconvexity of the cost with respect to f can also be seen from this alternative expression. It can be shown that the Fisher information term is a convex functional [4]. Therefore, $1 - I_f\left(Z_2\right)_{22}$ is concave and the sum of a convex and a concave functional may not be convex.

4. Limit cases. In this section we consider the two limit cases. First we consider the case where the transmission is noiseless, and then we investigate the case where the transmission noise intensity goes to infinity.

4.1. Noiseless transmission. Assume there is no uncertainty in transmitting information from the first to the second station, i.e., $\epsilon = 0$ and hence $z_{21} = z_1$. In this case, we have perfect recall and the information pattern is classical. We can write

$$\begin{aligned} p(z_2) &= p(z_{21}, z_{22}) = p(z_{22}|z_{21})\, p(z_{21}) \\ &= p(z_{22}|z_1)\, p(z_1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z_{22} - f(z_1))^2}{2}\right) p(z_1). \end{aligned} \tag{4.1}$$

Then, from (3.7), we have

$$g^*(z_2) = f(z_1) = f(z_{21}), \tag{4.2}$$

which could directly be obtained from the original definition for g^*, i.e.,

$$g^*(z_2) = E[f(z_1)|\, z_2] = f(z_1), \tag{4.3}$$

because z_1 is exactly known when z_2 is given. Substituting this back in (3.5) and minimizing with respect to the strategy f, we have

$$g^*(z_2) = f(z_1) = z_1, \tag{4.4}$$

and hence

$$\gamma_1(z_1) = 0, \tag{4.5}$$

$$\gamma_2(z_2) = z_1, \tag{4.6}$$

which is the unique linear set of optimal strategies. This indeed turns out to be a very simple example of the well-known classical LQG problem.

4.2. Infinite transmission noise intensity. Another limit case is when the transmission noise intensity increases to infinity. In this case, z_{21} and z_{22} become independent and we have

$$p(z_2) = p(z_{21}, z_{22}) = p(z_{21})\, p(z_{22}). \tag{4.7}$$

The Fisher information term can now be written as

$$\begin{aligned} I_f(Z_2)_{22} &= \int\int_{-\infty}^{+\infty} \left(\frac{\partial}{\partial z_{22}} \ln p(z_{21}, z_{22})\right)^2 p(z_{21}, z_{22})\, dz_{21} dz_{22} \\ &= \int_{-\infty}^{+\infty} \left(\frac{\partial}{\partial z_{22}} \ln p(z_{22})\right)^2 p(z_{22})\, dz_{22} \\ &= I_f(Z_{22}), \end{aligned} \tag{4.8}$$

which is indeed the Fisher information content of z_{22} only. Hence,

$$J^*(f) = k^2 E\left[(z_1 - f(z_1))^2\right] + 1 - I_f(Z_{22}). \tag{4.9}$$

This is the same result that was presented for the Witsenhausen counterexample in [1]. Intuitively, when we have infinite transmission noise intensity, we might as well deny the access to z_1 for the second station, and this is exactly the case in Witsenhausen's counterexample. The optimal strategies for this case are still unknown. Witsenhausen

showed that the optimal solution exists, even if x_0 has a general distribution with a finite second moment [1]. He then showed that if one of the strategies is restricted to being affine, the other optimal strategy would also be affine. But then he provided a set of nonlinear strategies that could achieve a lower cost for some values of k^2 and σ_0.

Different approaches have been taken in order to find the optimal strategies. As mentioned before, an asymptotic approach was used in [2] for the case where σ_0 is small. More recently, in [5], [6], [7] it was shown how a neural network, trained by stochastic approximation techniques, can be employed as a nonlinear function approximator in order to approximate $f(z_1)$. It was demonstrated that the optimal $f^*(z_1)$ may not be strictly piecewise, as was suggested by Witsenhausen, but slightly sloped. Some researchers have tried to attack the problem numerically and use some sample and search techniques to find the solution. A discretized version of the problem was formulated in [8], which was later shown in [9] to be NP-complete and computationally intractable. It is recently asserted in [10] and [11] that a global optimum would be achieved by searching directly in the strategy space using the generalized step functions to approximate $f(z_1)$.

So far we have shown, through a simple example, how any uncertainty in the transmission of information between the stations in a distributed system can make the optimal control design very complicated and even intractable. Then, by considering the two limit cases, we showed how our example covers a very wide range of scenarios. Namely, we saw that for the noiseless transmission case, the unique optimal strategies, which are linear in the information, are easily obtained, whereas for the infinite transmission noise intensity, the optimal strategies are still unknown. Now a very feasible case to investigate is when the uncertainty on the information transmission is small. In fact, when the transmission noise intensity ϵ is small, one would still expect behavior similar to the noiseless transmission case for the optimal strategies. In the following sections, we consider this case. Namely, we assume a small intensity for v_t. Under this assumption, we obtain the first few terms in the expansion of the performance index in terms of ϵ. We then use the Hamiltonian approach in order to find a necessary condition for the strategies that minimize the approximated cost.

We show that the linear strategies, with slightly different coefficients than the corresponding coefficients for the noiseless transmission case, do indeed satisfy the necessary condition. This asymptotic analysis not only gives us insight on how the optimal strategies change as the transmission uncertainty is introduced but also provides us with a better sense of the complexities in the design procedure.

5. An expansion for the cost. Assume that the first station communicates with the second station through a low noise channel. In other words, the transmission noise intensity ϵ is assumed to be small. In this section, we will find an expansion for the cost in terms of ϵ. For this purpose, we first find an expansion for the probability density function of the information available to the second station, i.e., $p(z_2)$. Then we use (3.7) in order to find the corresponding expansion for $g^*(z_2)$. By substituting back in (3.5), we will obtain the expanded cost only in terms of f.

The probability density function for z_2 can be written as

$$p_\epsilon(z_2) := p(z_2) = \int_{-\infty}^{+\infty} p(z_{22}, z_{21}, z_1)\, dz_1 \tag{5.1}$$

$$= \int_{-\infty}^{+\infty} p(z_{22}|z_{21}, z_1)\, p(z_{21}|z_1)\, p(z_1)\, dz_1 \tag{5.2}$$

$$= \int_{-\infty}^{+\infty} p\left(z_{22}|z_1\right) p\left(z_{21}|z_1\right) p\left(z_1\right) dz_1 \tag{5.3}$$

$$= \int_{-\infty}^{+\infty} p\left(z_{22}|z_1\right) p_{v_t}\left(z_{21}-z_1\right) p\left(z_1\right) dz_1 \tag{5.4}$$

$$= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(z_{22}-f\left(z_1\right)\right)^2}{2}\right) \frac{1}{\sqrt{2\pi}\epsilon} \exp\left(-\frac{\left(z_{21}-z_1\right)^2}{2\epsilon^2}\right) \times \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{z_1^2}{2\sigma_0^2}\right) dz_1, \tag{5.5}$$

where for (5.3) we have used the facts that the σ-fields generated by $\{z_{21}, z_1\}$ and $\{z_1, v_t\}$ are the same and z_1, v_t, and v_2 are mutually independent. At this point, one should note that even though the joint probability density function $p\left(z_{22}, z_{21}, z_1\right)$ can be explicitly expressed as in (5.5), introduction into the performance index shows that determination of $f\left(z_1\right)$ still requires averaging over all random variables. This is another way of looking at the effect of a nonclassical information pattern, which is not partially nested. We therefore decide to follow an asymptotic approach.

For small ϵ, we now approximate $\ln p_\epsilon\left(z_2\right)$ by considering only the first three terms of its expansion around $\epsilon = 0$. Namely,

$$\ln p_\epsilon\left(z_2\right) \simeq \ln p_0\left(z_2\right) + \frac{\partial}{\partial\epsilon} \ln p_\epsilon\left(z_2\right)\Big|_{\epsilon=0} \epsilon + \frac{\partial^2}{\partial\epsilon^2} \ln p_\epsilon\left(z_2\right)\Big|_{\epsilon=0} \epsilon^2. \tag{5.6}$$

By making the change of variables

$$\epsilon y := z_1 - z_{21} \Rightarrow \epsilon dy = dz_1, \tag{5.7}$$

we can write $p_\epsilon\left(z_2\right)$ in the following form:

$$p_\epsilon\left(z_2\right) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(z_{22}-\bar{f}_\epsilon(y)\right)^2}{2}\right) \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{\left(z_{21}+\epsilon y\right)^2}{2\sigma_0^2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy, \tag{5.8}$$

where

$$\bar{f}_\epsilon(y) := f\left(\epsilon y + z_{21}\right). \tag{5.9}$$

It is now clear that

$$p_0\left(z_2\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(z_{22}-f\left(z_{21}\right)\right)^2}{2}\right) \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{z_{21}^2}{2\sigma_0^2}\right), \tag{5.10}$$

and hence

$$\ln p_0\left(z_2\right) = -\frac{\left(z_{22}-f\left(z_{21}\right)\right)^2}{2} - \frac{z_{21}^2}{2\sigma_0^2} + \ln\left(\frac{1}{2\pi\sigma_0}\right). \tag{5.11}$$

For the first order term, we have

$$\frac{\partial}{\partial\epsilon} \ln p_\epsilon\left(z_2\right)\Big|_{\epsilon=0} = \frac{1}{p_0\left(z_2\right)} \frac{\partial}{\partial\epsilon} p_\epsilon\left(z_2\right)\Big|_{\epsilon=0}. \tag{5.12}$$

On the other hand,

$$\begin{aligned}\frac{\partial}{\partial\epsilon}p_\epsilon(z_2)\bigg|_{\epsilon=0} &= \int_{-\infty}^{+\infty}\frac{\partial}{\partial\epsilon}\left\{\frac{1}{\sqrt{2\pi}}e^{-\frac{(z_{22}-\bar{f}_\epsilon(y))^2}{2}}\frac{1}{\sqrt{2\pi}\sigma_0}e^{-\frac{(z_{21}+\epsilon y)^2}{2\sigma_0^2}}\right\}\bigg|_{\epsilon=0}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy\\ &= \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}(z_{22}-f(z_{21}))\,yf'(z_{21})\,e^{-\frac{(z_{22}-f(z_{21}))^2}{2}}\frac{1}{\sqrt{2\pi}\sigma_0}e^{-\frac{z_{21}^2}{2\sigma_0^2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy\\ &+\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{(z_{22}-f(z_{21}))^2}{2}}\frac{1}{\sqrt{2\pi}\sigma_0}\left(-\frac{z_{21}}{\sigma_0^2}\right)ye^{-\frac{z_{21}^2}{2\sigma_0^2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy=0.\end{aligned} \tag{5.13}$$

Therefore,

$$\frac{\partial}{\partial\epsilon}\ln p_\epsilon(z_2)\bigg|_{\epsilon=0}=0. \tag{5.14}$$

We could somehow expect this result. This is because we would expect the behavior of $p_\epsilon(z_2)$ to depend only on the variance of the Gaussian transmission noise, i.e., ϵ^2. Using (5.14), we can now obtain the second order term as

$$\frac{\partial^2}{\partial\epsilon^2}\ln p_\epsilon(z_2)\bigg|_{\epsilon=0}=\frac{1}{p_0(z_2)}\,\frac{\partial^2}{\partial\epsilon^2}p_\epsilon(z_2)\bigg|_{\epsilon=0}. \tag{5.15}$$

After some tedious but straightforward manipulations, we get

$$\begin{aligned}\frac{\partial^2}{\partial\epsilon^2}\ln p_\epsilon(z_2)\bigg|_{\epsilon=0} &= -f'^2(z_{21})+f''(z_{21})(z_{22}-f(z_{21}))+f'^2(z_{21})(z_{22}-f(z_{21}))^2\\ &\quad+2f'(z_{21})(z_{22}-f(z_{21}))\left(-\frac{z_{21}}{\sigma_0^2}\right)-\frac{1}{2\sigma_0^2}+\frac{z_{21}^2}{\sigma_0^4}.\end{aligned} \tag{5.16}$$

We can now obtain a second order approximation for $\ln p_\epsilon(z_2)$ by substituting the corresponding terms from (5.11), (5.14), and (5.16) back into the expansion (5.6). In the next step, we substitute the expansion for $\ln p_\epsilon(z_2)$ in (3.7) in order to find the corresponding expansion for $g^*(z_2)$. Remember that $g^*(z_2)$ is the optimal strategy for the second station, assuming that the first station has a fixed strategy $\gamma_1(z_1)=f(z_1)-z_1$. We have

$$\begin{aligned}g^*(z_2) &= z_{22}+\frac{\partial}{\partial z_{22}}\ln p(z_2)\\ &\simeq z_{22}+\frac{\partial}{\partial z_{22}}\ln p_0(z_2)+\epsilon^2\frac{\partial}{\partial z_{22}}\left(\frac{\partial^2}{\partial\epsilon^2}\ln p_\epsilon(z_2)\bigg|_{\epsilon=0}\right)\\ &= z_{22}-(z_{22}-f(z_{21}))\\ &\quad+\epsilon^2\left[f''(z_{21})+2f'^2(z_{21})(z_{22}-f(z_{21}))+2f'(z_{21})\left(-\frac{z_{21}}{\sigma_0^2}\right)\right].\end{aligned} \tag{5.17}$$

Our goal is to get an expansion for the cost, which as we know from (3.5) can be written as

$$J^*(f)=k^2E\left[(z_1-f(z_1))^2\right]+E\left[(f(z_1))^2\right]-E\left[(g^*(z_2))^2\right]. \tag{5.18}$$

Using the expansion for $g^*(z_2)$ from (5.17), we have

$$E\left[(g^*(z_2))^2\right] \simeq E\left[(f(z_{21}))^2\right]$$
$$+2\epsilon^2 E\left[f(z_{21})\left(f''(z_{21})+2f'^2(z_{21})(z_{22}-f(z_{21}))+2f'(z_{21})\left(-\frac{z_{21}}{\sigma_0^2}\right)\right)\right], \tag{5.19}$$

where we have neglected the fourth order term in ϵ. Substituting this expansion back in (5.18), we will obtain the following expansion for the cost:

$$J^*(f) = k^2 E\left[(z_1-f(z_1))^2\right] + E\left[(f(z_1))^2\right] - E\left[(f(z_{21}))^2\right]$$
$$-2\epsilon^2 E\left[f(z_{21})\left(f''(z_{21})+2f'^2(z_{21})(z_{22}-f(z_{21}))+2f'(z_{21})\left(-\frac{z_{21}}{\sigma_0^2}\right)\right)\right]. \tag{5.20}$$

Note that when the transmission is noiseless, i.e., $\epsilon = 0$ and therefore $z_{21} = z_1$, we have

$$J^*(f) = k^2 E\left[(z_1-f(z_1))^2\right], \tag{5.21}$$

and $f(z_1) = z_1$ is the obvious unique optimal solution. The above expansion, however, is not exactly in our desired form yet. This is because the third term on the right-hand side, which is an average over z_{21}, still depends on ϵ. We shall now rewrite the expansion in (5.20) by explicitly expressing the expectations based on the corresponding probability densities:

$$\begin{aligned} J^*(f) = &\int_{-\infty}^{+\infty}\left[k^2(t-f(t))^2+f^2(t)\right]\frac{1}{\sqrt{2\pi}\sigma_0}e^{-\frac{t^2}{2\sigma_0^2}}dt \\ &-\int_{-\infty}^{+\infty}\left[f^2(t)+2\epsilon^2\left(f(t)f''(t)-2f(t)f'(t)\frac{t}{\sigma_0^2}\right)\right]\frac{1}{\sqrt{2\pi(\sigma_0^2+\epsilon^2)}}e^{-\frac{t^2}{2\left(\sigma_0^2+\epsilon^2\right)}}dt \\ &-\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}4\epsilon^2 f(t)f'^2(t)(\tau-f(t))\frac{1}{\sqrt{2\pi}}e^{-\frac{(\tau-f(t))^2}{2}}\frac{1}{\sqrt{2\pi}\sigma_0}e^{-\frac{t^2}{2\sigma_0^2}}dtd\tau, \end{aligned} \tag{5.22}$$

where we have substituted $p(z_2) = p(z_{22}, z_{21}) \simeq p_0(z_2)$ in the third term, since the higher order terms would be multiplied by ϵ^2 and would then be neglected. Now the third term turns out to be zero, because

$$\begin{aligned} &\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}4\epsilon^2 f(t)f'^2(t)(\tau-f(t))\frac{1}{\sqrt{2\pi}}e^{-\frac{(\tau-f(t))^2}{2}}\frac{1}{\sqrt{2\pi}\sigma_0}e^{-\frac{t^2}{2\sigma_0^2}}dt \\ &=\int_{-\infty}^{+\infty}4\epsilon^2 f(t)f'^2(t)\frac{1}{\sqrt{2\pi}\sigma_0}e^{-\frac{t^2}{2\sigma_0^2}}\left(\int_{-\infty}^{+\infty}(\tau-f(t))\frac{1}{\sqrt{2\pi}}e^{-\frac{(\tau-f(t))^2}{2}}d\tau\right)dt = 0. \end{aligned} \tag{5.23}$$

At the same time, we can expand the probability density of z_{21} up to the second order in ϵ. It is actually straightforward to obtain

$$\frac{1}{\sqrt{2\pi(\sigma_0^2+\epsilon^2)}}e^{-\frac{t^2}{2\left(\sigma_0^2+\epsilon^2\right)}} \simeq \frac{1}{\sqrt{2\pi}\sigma_0}e^{-\frac{t^2}{2\sigma_0^2}} + \epsilon^2\frac{1}{\sqrt{2\pi}\sigma_0^5}\left(t^2-\sigma_0^2\right)e^{-\frac{t^2}{2\sigma_0^2}}. \tag{5.24}$$

Substituting (5.23) and the above expansion back in (5.22) and neglecting the higher order terms in ϵ, we can finally get the following expansion for the cost:

$$J^*(f) = \int_{-\infty}^{+\infty} \left[k^2 \left(t - f(t)\right)^2\right] \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{t^2}{2\sigma_0^2}} dt + \epsilon^2 \int_{-\infty}^{+\infty} \left[4f(t)f'(t)\frac{t}{\sigma_0^2} - 2f(t)f''(t) + f^2(t)\frac{\sigma_0^2 - t^2}{\sigma_0^4}\right] \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{t^2}{2\sigma_0^2}} dt$$
$$:= J_0^* + \epsilon^2 J_1^*. \tag{5.25}$$

The objective is now to obtain the function f, which minimizes the above approximated cost. In the next section, we use a variational approach in order to find a necessary condition for such a function and show how the linear strategies still satisfy this necessary condition.

6. Minimizing the approximated cost. So far, we have obtained an expansion for the cost assuming that the transmission noise intensity is small. We have, in fact, approximated the cost by including only up to the second order term in ϵ. We should now try to minimize this approximated cost in order to find the asymptotically optimal f^*. Obviously, this strategy would be optimal only for a small transmission noise intensity. However, it would still be very helpful for the analysis of the behavior of the optimal strategies when we deviate a little bit from the classical information pattern by introducing a small communication uncertainty.

We now use the Hamiltonian approach in order to find the necessary conditions for the function $f(t)$, which minimizes our approximated cost. For simplicity, denote

$$x_1(t) := f(t), \tag{6.1}$$
$$x_2(t) := \dot{x}_1(t) = f'(t), \tag{6.2}$$
$$u(t) := \dot{x}_2(t) = \ddot{x}_1(t) = f''(t), \tag{6.3}$$
$$p(t) := \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{t^2}{2\sigma_0^2}}. \tag{6.4}$$

The Hamiltonian is then defined as [12]

$$\mathcal{H} = k^2 \left(t - x_1(t)\right)^2 p(t) + \epsilon^2 \left(4x_1(t)x_2(t)\frac{t}{\sigma_0^2} - 2x_1(t)u(t) + x_1^2(t)\frac{\sigma_0^2 - t^2}{\sigma_0^4}\right) p(t) + \lambda_1(t)x_2(t) + \lambda_2(t)u(t), \tag{6.5}$$

where λ_1 and λ_2 are the Lagrange multipliers that should satisfy

$$\dot{\lambda}_1(t) = -\mathcal{H}_{x_1} = \left(2k^2\left(t - x_1(t)\right) - 4\epsilon^2 x_2(t)\frac{t}{\sigma_0^2} - 2\epsilon^2 x_1(t)\frac{\sigma_0^2 - t^2}{\sigma_0^4} + 2\epsilon^2 u(t)\right) p(t), \tag{6.6}$$

$$\dot{\lambda}_2(t) = -\mathcal{H}_{x_2} = -4\epsilon^2 x_1(t)\frac{t}{\sigma_0^2} p(t) - \lambda_1(t). \tag{6.7}$$

But as we can see, the Hamiltonian is linear in $u(t)$ and we actually have a *singular* optimization problem. The singular surface will be characterized by setting $\mathcal{H}_u$ and its derivatives with respect to t equal to zero, that is,

$$\mathcal{H}_u = -2\epsilon^2 x_1(t)p(t) + \lambda_2(t) = 0, \tag{6.8}$$

and

$$\frac{d}{dt}\mathcal{H}_u = -2\epsilon^2\dot{x}_1(t)p(t) - 2\epsilon^2 x_1(t)\dot{p}(t) + \dot{\lambda}_2(t) = 0. \tag{6.9}$$

Substituting $\dot{p}(t) = -\frac{t}{\sigma_0^2}p(t)$ and also $\dot{\lambda}_2$ from (6.7), we get

$$\frac{d}{dt}\mathcal{H}_u = -2\epsilon^2 x_2(t)p(t) - 2\epsilon^2 x_1(t)\frac{t}{\sigma_0^2}p(t) - \lambda_1(t) = 0. \tag{6.10}$$

Differentiating again and substituting $\dot{\lambda}_1$ from (6.6), we have

$$\frac{d^2}{dt^2}\mathcal{H}_u = -4\epsilon^2 u(t)p(t) + 4\epsilon^2\frac{t}{\sigma_0^2}x_2(t)p(t) - 2k^2\left(t - x_1(t)\right)p(t) = 0. \tag{6.11}$$

Therefore, the corresponding $u(t)$ on the singular surface is

$$u(t) = x_2(t)\frac{t}{\sigma_0^2} - \frac{k^2}{2\epsilon^2}\left(t - x_1(t)\right). \tag{6.12}$$

Note that the first order generalized Legendre–Clebsch condition, which is a necessary condition for $u(t)$ to be minimizing on the singular surface, is also satisfied, namely,

$$\frac{\partial}{\partial u}\left(\frac{d^2}{dt^2}\mathcal{H}_u\right) \leq 0. \tag{6.13}$$

Therefore, the corresponding $x_1(t)$ and $x_2(t)$, which minimize our approximated cost, should necessarily satisfy the following differential equations:

$$\dot{x}_1(t) = x_2(t), \tag{6.14}$$

$$\dot{x}_2(t) = x_2(t)\frac{t}{\sigma_0^2} - \frac{k^2}{2\epsilon^2}\left(t - x_1(t)\right). \tag{6.15}$$

Since ϵ is assumed to be small, we may assume the following form in order to obtain the solutions for the above differential equations:

$$x_1(t) = a_0(t) + \epsilon^2 a_2(t) + \epsilon^4 a_4(t) + \cdots, \tag{6.16}$$

$$x_2(t) = b_0(t) + \epsilon^2 b_2(t) + \epsilon^4 b_4(t) + \cdots. \tag{6.17}$$

Interestingly enough, by substituting the above x_1 and x_2 back into the differential equations and comparing the coefficients of the terms with the same order in ϵ, we get

$$x_1(t) = \left[1 - \frac{2\epsilon^2}{k^2\sigma_0^2} + \left(\frac{2\epsilon^2}{k^2\sigma_0^2}\right)^2 - \left(\frac{2\epsilon^2}{k^2\sigma_0^2}\right)^3 + \cdots\right]t = \frac{t}{\left(1 + \frac{2\epsilon^2}{k^2\sigma_0^2}\right)}. \tag{6.18}$$

Back to our original notation, we actually have

$$f(z_1) = \frac{z_1}{\left(1 + \frac{2\epsilon^2}{k^2\sigma_0^2}\right)}. \tag{6.19}$$

As we can see, the solution is still linear with a coefficient which is slightly different than the corresponding coefficient for the noiseless transmission case. Remember that

$f(z_1) = z_1$ is the optimal solution when there is no transmission noise, and note that for $\epsilon = 0$ in (6.19) we get exactly the same solution as expected. Given the above function $f(z_1)$, the corresponding $g^*(z_2)$ can easily be obtained using (3.4). Note that it will also be linear because of the Gaussian assumption for the underlying uncertainties.

We could somehow expect the optimal strategies to be linear from the beginning. As we mentioned in section 2, linear strategies were shown to be asymptotically optimal for the Witsenhausen example when the uncertainty on the information available to the first station is small [2]. In this paper, however, we have considered a reformulation of Witsenhausen's problem where the first station sends its information to the second station through a low noise channel. These two scenarios are somewhat similar. Namely, in both scenarios, the second station can determine the information available to the first station fairly accurately. Specifically, in the first scenario, the second station almost knows z_1 because of its small uncertainty, while in the second scenario it can determine z_1 from the information that is transmitted through a low noise channel.

We would also expect the optimal strategies to approach the corresponding strategies for the noiseless transmission case as the value of z_1 and, in some sense, the *signal-to-noise ratio* increases. This does not seem to happen in the solution (6.19). One may justify this by looking at the exponential function in the cost (5.25). This function drives the integrand of the cost to zero exponentially fast for large values of z_1. Therefore, the structure of the cost really does not force the optimal solution to approach $f(z_1) = z_1$ as z_1 increases.

We shall now obtain the corresponding value of the cost. Substituting $f(t)$ from (6.19) back into the cost (5.25), we get

$$\begin{aligned} J^*(f) &= \int_{-\infty}^{+\infty} \left[k^2 \left(t - \frac{t}{1 + \frac{2\epsilon^2}{k^2\sigma_0^2}} \right)^2 \right] \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{t^2}{2\sigma_0^2}} dt \\ &\quad + \epsilon^2 \int_{-\infty}^{+\infty} \left[4 \frac{t}{\left(1 + \frac{2\epsilon^2}{k^2\sigma_0^2}\right)^2} \frac{t}{\sigma_0^2} + \frac{t^2}{\left(1 + \frac{2\epsilon^2}{k^2\sigma_0^2}\right)^2} \frac{\sigma_0^2 - t^2}{\sigma_0^4} \right] \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{t^2}{2\sigma_0^2}} dt \\ &= \frac{1}{\left(1 + \frac{2\epsilon^2}{k^2\sigma_0^2}\right)^2} \left(2\epsilon^2 + \frac{4\epsilon^4}{k^2\sigma_0^2} \right) \\ &\simeq 2\epsilon^2 - \frac{4\epsilon^4}{k^2\sigma_0^2}\,, \end{aligned} \tag{6.20}$$

where we have used

$$\int_{-\infty}^{+\infty} t^2 \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{t^2}{2\sigma_0^2}} dt = \sigma_0^2, \tag{6.21}$$

$$\int_{-\infty}^{+\infty} t^4 \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{t^2}{2\sigma_0^2}} dt = 3\sigma_0^4. \tag{6.22}$$

The optimal cost for the noiseless transmission case is zero. But if we use $f(z_1) = z_1$ when the transmission is noisy, we get the following cost:

$$J^*(f) = 2\epsilon^2. \tag{6.23}$$

In other words, if we fix the strategies to be the optimal strategies for the noiseless transmission case while introducing a small transmission noise, the increase in the cost will be proportional to the transmission noise intensity. However, if we use (6.19), we can indeed improve the cost by the fourth order in ϵ.

One should note from (6.19) and (6.20) that as the value of $k^2\sigma_0^2$ increases, the asymptotically optimal solution approaches $f(z_1) = z_1$, and the change in the cost becomes smaller. In other words, increasing $k^2\sigma_0^2$ has an effect similar to decreasing the communication uncertainty. To explain this, we note from the performance index that increasing k^2 implies a more expensive control action for the first station, which, in turn, results in smaller u_1. This then implies that the information available to the second station is less affected by the action of the first station. At the same time, increasing σ_0^2 implies a higher level of uncertainty on x_0, which, incidentally, is the piece of information that is being transmitted between the stations.

This brings up an example of a very interesting fundamental issue: the notion of *information value* and how it could be different for control and communication purposes. In fact, we know from information theory that a higher level of uncertainty for a piece of information implies a higher level of *entropy* and therefore a more valuable piece of information for transmission. On the other hand, however, a more uncertain piece of information would probably be less valuable for control purposes and would have smaller effect on the control strategies. In other words, a control designer would probably be willing to spend less on installing transmitters on the stations for communicating more uncertain pieces of information. While defining a notion for the value of information for control purposes has been occasionally addressed in the literature for quite a long time, it still remains an open problem. This is mostly because of the fact that the value of information for control purposes would highly depend on how the cost is defined for the control design, and this could be quite different in various applications.

7. Concluding remarks. We analyzed an example of a decentralized stochastic system. This example was a reformulation of the Witsenhausen counterexample where the first station was allowed to send its information to the second station through a noisy channel. The dynamics were linear, all the underlying uncertainties were assumed to be Gaussian, and the cost was quadratic. It was shown that as soon as any uncertainty is introduced in the communication among the stations, the information pattern again becomes nonclassical, which is not partially nested. We then showed how the performance index can be alternatively expressed such that the possible nonconvexity of the cost, with respect to the control strategies, becomes more transparent. Therefore, in general, we will end up with a nonconvex functional optimization problem when we try to obtain the decentralized optimal control algorithms. We then considered two limit cases. Namely, the case where there is no communication uncertainty and the case in which the transmission noise intensity increases to infinity. The former case was shown to be a trivial example of a classical LQG problem, whereas the latter case corresponds to Witsenhausen's counterexample, the optimal solution of which is still unknown.

We then focused on the case where the communication uncertainty was small. We followed an asymptotic approach where we approximated the cost based on its expansion in terms of the small transmission noise intensity. We showed how minimizing the approximated cost can be seen as a singular optimization problem. We then used a variational approach in order to find the necessary conditions for the asymptotically optimal strategies and showed that some reasonable linear strategies

would actually satisfy those conditions. We also provided some intuitive explanations for the behavior of those linear strategies and obtained the corresponding cost.

Note that while we have focused on the reformulated Witsenhausen counterexample, our main result is quite general. In fact, we have shown through an example that communication uncertainties in decentralized systems generally result in nonclassical information patterns, which, in turn, can destroy the convexity of the associated functional optimization problems. Moreover, our approach is indeed a very general approach, which have been applied to various other problems before. More specifically, expanding a cost function in terms of some small parameters is a common practice in variational and perturbation-based approaches. Furthermore, using Hamiltonian approach in order to obtain the necessary conditions for the optimal strategies obviously is not specific to our reformulated Witsenhausen problem. However, finding the exact function (6.19), which is obtained in closed form, satisfies the necessary condition for optimality, and shows how the optimal strategies could change upon introduction of some communication uncertainty, could be very specific to our problem.

All the derivations and the results in this paper show some of the difficulties involved in dealing with decentralized systems as soon as we deviate a little bit from a classical, or at least a partially nested, information pattern. On the other hand, even though we have modeled the communication uncertainty in the simplest possible way, we have tried to emphasize the role of communication uncertainties in generating such information patterns that are very difficult to handle.

Finally, it should be mentioned that even though the optimization problem is generally difficult for this class of systems, in some applications one might be able to exploit the specific structure of the system in order to obtain some reasonably good *suboptimal* strategies, which could yield an acceptable performance.

REFERENCES

[1] H. S. Witsenhausen, *A counterexample in stochastic optimum control*, SIAM J. Control, 6 (1968), pp. 131–147.

[2] D. A. Castanon and N. R. Sandell, Jr., *Signaling and uncertainty: A case study*, in Proceedings of the IEEE Conference on Decision and Control, 1978, pp. 1140–1144.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[4] M. L. Cohen, *The Fisher information and convexity*, IEEE Trans. Inform. Theory, 14 (1968), pp. 591–592.

[5] M. Baglietto, T. Parisini, and R. Zoppoli, *Numerical solutions to the Witsenhausen counterexample by approximating networks*, IEEE Trans. Automat. Control, 46 (2001), pp. 1471–1477.

[6] M. Baglietto, T. Parisini, and R. Zoppoli, *Nonlinear approximations for the solution of team optimal control problems*, in Proceedings of the 36th IEEE Conference on Decision and Control, Vol. 5, 1997, pp. 4592–4594.

[7] M. Baglietto, T. Parisini, and R. Zoppoli, *Neural networks for the solution of information distributed optimal control problems*, in Proceedings of 6th European Symposium on Artificial Neural Networks, ESANN'98, 1998, pp. 79–84.

[8] Y. C. Ho and T. S. Chang, *Another look at the non-classical information structure problem*, IEEE Trans. Automat. Control, 25 (1980), pp. 537–540.

[9] C. H. Papadimitriou and J. Tsitsiklis, *Intractable problems in control theory*, SIAM J. Control Optim., 24 (1986), pp. 639–654.

[10] J. T. Lee, E. Lau, and Y. C. Ho, *The Witsenhausen counter-example: A hierarchical search approach for nonconvex optimization problems*, IEEE Trans. Automat. Control, 46 (2001), pp. 382–397.

[11] J. T. Lee, E. Lau, and Y. C. Ho, *On the Global Optimum of the Witsenhausen Counter-Example*, manuscript.

[12] A. E. Bryson, Jr., and Y. C. Ho, *Applied Optimal Control*, Taylor and Francis, Bristol, PA, 1975.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 4, pp. 991–1014

STRONG OPTIMALITY FOR A BANG-BANG TRAJECTORY*

ANDREI A. AGRACHEV†, GIANNA STEFANI‡, AND PIERLUIGI ZEZZA§

Abstract. In this paper we give sufficient conditions for a bang-bang regular extremal to be a strong local optimum for a control problem in the Mayer form; strong means that we consider the C^0 topology in the state space. The controls appear linearly and take values in a polyhedron, and the state space and the end point constraints are finite-dimensional smooth manifolds. In the case of bang-bang extremals, the kernel of the first variation of the problem is trivial, and hence the usual second variation, which is defined on the kernel of the first one, does not give any information. We consider the finite-dimensional subproblem generated by perturbing the switching times, and we prove that the sufficient second order optimality conditions for this finite-dimensional subproblem yield local strong optimality. We give an explicit algorithm to check the positivity of the second variation which is based on the properties of the Hamiltonian fields.

Key words. optimal control, bang-bang controls, sufficient optimality condition, strong local optima

AMS subject classifications. Primary, 49K15; Secondary, 49K30, 58E25

PII. S036301290138866X

1. Introduction.

This paper is part of a general research program whose aim is to further extend the use of Hamiltonian methods in the study of optimal control problems. We believe that these methods can play a relevant role in control theory because they allow a general approach to sufficient conditions for strong local optimality, as we wish to show here.

The Hamiltonian approach to strong optimality consists of constructing a field of state extremals covering a neighborhood of a given trajectory which has to be tested. This field of extremals is obtained by projecting on the state manifold M the flow $\mathcal{H}_t$ of the maximized Hamiltonian emanating from the Lagrangian submanifold of the initial transversality conditions. If this projection admits a Lipschitz continuous local inverse, then we can estimate the variation of the cost function at a neighboring trajectory by a function ψ which depends only on the final point, and it is hence independent of the control differential equation; in this way we reduce the problem to a finite-dimensional one. The existence of a Lipschitz continuous local inverse is guaranteed by the surjectivity of the projection on M of the tangent map to the flow $\mathcal{H}_t$. This construction corresponds to the classical one of a nonselfintersecting family of state extremals. This is enough to obtain optimality if the final point is fixed since the submanifold of the final end points reduces to a singleton; otherwise we need some further optimality condition on the function ψ.

We use the relations existing between a suitable second variation and the symplectic properties of the Hamiltonian flow to show that when this second variation is positive definite then the projection on the state manifold M of the tangent map to

*Received by the editors May 2, 2001; accepted for publication (in revised form) February 25, 2002; published electronically October 8, 2002.
http://www.siam.org/journals/sicon/41-4/38866.html

†Steklov Math. Inst., Gubkina ul.8, 117966 Moscow, Russia, and SISSA, Via Beirut 4, 34014 Trieste, Italy (agrachev@sissa.it).

‡Dipartimento di Matematica Applicata, G. Sansone, Via S. Marta 3, 50139 Firenze, Italy (stefani@dma.unifi.it).

§Dipartimento di Matematica per le Decisioni, Via C. Lombroso 6/17, 50134 Firenze, Italy (pzezza@unifi.it, http://www.dmd.unifi.it/zezza).

the flow $\mathcal{H}_t$ is surjective; moreover the positivity of this second variation leads also to the sufficient optimality conditions for the function ψ.

To make this general approach possible we need an intrinsic formulation of the second variation as an accessory linear-quadratic minimization problem on the tangent space; this will allow us to exploit one of the crucial ideas underlying the Hamiltonian approach: the tangent map of the flow of the maximized Hamiltonian is the linear Hamiltonian flow of an associated linear-quadratic problem, i.e., the flow of the Jacobi system.

Another important issue is that, when the initial point is not free, it is not possible to cover a neighborhood of the initial point by the projection of the Hamiltonian flow. In the calculus of variations this problem has been solved by perturbing the initial time, but this method does not always work in optimal control because the projection could be singular for a time interval of positive length; this is always the case for bang-bang controls if there is a constraint on the initial point. We propose a different approach: when the second variation is positive we add a penalty term, which allows us to reduce the original problem to another one without constraints on the initial point.

Some of these issues have already been addressed. In [ASZ98b] we stated sufficient conditions for strong local optimality for an optimal control problem in $\mathbb{R}^n$ with unbounded controls, while in [ASZ98a] we gave an intrinsic expression of the accessory problem and studied the relations between the Hamiltonian flow and the index of the second variation. The geometric properties of the field of extremals necessary for proving sufficient conditions for strong optimality were studied in [ASZ99].

In this paper we study a control problem in the Mayer form where the controls appear linearly and take values in a polyhedron, the state space and the end point constraints are finite-dimensional smooth submanifolds, and we give sufficient conditions for a bang-bang extremal to be a strong local minimizer.

In the bang-bang case we have to face some new problems. Since the maximized Hamiltonian is not smooth at the switching points we need to give conditions (see Assumptions 2.1, 2.2, 2.3) which assure us that its flow is defined and piecewise smooth around the reference adjoint covector. Moreover, in the case of bang-bang extremals, the kernel of the first variation of the problem is trivial, and hence the usual second variation, which is defined on the kernel of the first one, does not give any information. We solve this problem by considering the finite-dimensional subproblem generated by perturbing the switching times. The usual (finite-dimensional) second order optimality conditions for this problem give an appropriate second variation. Indeed we prove that the positivity of this second variation yields that the Hamiltonian flow has the properties we have described so that we can prove strong local optimality for the reference trajectory. The set of admissible variations on which we test the second variation can be very small, its dimension can be less than the state space dimension, and when it is zero, we directly have optimality.

By introducing an analogue of the strict Legendre condition, Assumption 2.3, we can eliminate the control from the extremality conditions for the second variation, and its extremals are then described by a discrete version of the Jacobi system, (2.9); the flow of this system describes the tangent subspaces to the flow of the maximized Hamiltonian at the points of nonsmoothness. Since the optimality can be lost only at these points, then the positivity of the second variation can be checked by an algorithm (see Lemma 2.8) which is based on the properties of the discrete flow of the bang-bang Jacobi system. For analogous conditions in the case of unbounded controls, see [ASZ98b].

The literature on second order sufficient conditions for the optimality of a bang-bang trajectory is scarce; we refer to [PS00] and the references therein for results based on the existence of a regular synthesis, and to [Sar92] and [Sar97], where the author studies local minima in the L^1 norm on the control in the time-optimal case. For a general description of the classical study of strong local optimality in the one-dimensional calculus of variations, see [GH96a, GH96b].

2. Statement of the results. Let $X_i,\ i = 1, 2, \dots, m$, be distinct C^∞ vector fields defined on the C^∞ finite-dimensional manifold M and let $\Delta = co\,\{e_1,\, e_2,\, \dots,\, e_m\}$ be the unitary simplex in $\mathbb{R}^m$.

We are interested in the optimal control problem

$$\text{Minimize} \quad J(\xi) := c_0(\xi(0)) + c_T(\xi(T))$$

subject to

$$\dot\xi(t) = \sum_{i=1}^{m} u_i(t)\, X_i(\xi(t)), \quad u(t) \in \Delta \tag{2.1}$$

$$\xi(0) \in N_0\ ,\ \xi(T) \in N_T, \tag{2.2}$$

where the time interval $[0, T]$ is fixed, N_0, N_T are given C^∞ submanifolds of M, and c_0, c_T are real-valued smooth functions. We will give sufficient conditions for a trajectory to be a strong local optimum, where strong means that we consider the C^0 topology in the state space.

As a candidate optimal solution we are given a *bang-bang Pontryagin extremal* $(\hat\xi, \hat u)$, that is, an absolutely continuous solution $\hat\xi : [0, T] \to M$ of system (2.1)–(2.2) with corresponding control $\hat u$ satisfying the Pontryagin maximum principle (PMP); moreover there is a partition of $[0, T]$

$$0 = t_0 < t_1 < t_2 < \cdots < t_r < t_{r+1} = T$$

such that

$$\hat u(t) = e_{j_i}, \quad t \in (t_{i-1}, t_i),$$

for some $j_i \in \{1, 2, \dots, m\}$. Therefore $\hat\xi$ is a solution of

$$\dot\xi(t) = X_{j_i}(\xi(t)), \quad t \in [t_{i-1}, t_i], \tag{2.3}$$

in each subinterval. The values t_i for $i = 1, 2, \dots, r$ will be called *switching times*, and we set

$$x_0 := \hat\xi(0), \quad x_T := \hat\xi(T)$$

to simplify notation. Corresponding to the reference extremal we define the time-dependent vector field

$$\hat h : [0, T] \times M \to TM \quad \text{as} \quad \hat h_{|(t_{i-1}, t_i)} := X_{j_i},$$

and we set $h_i := X_{j_i}$. Therefore the reference trajectory is a solution of the differential equation

$$\dot\xi(t) = \hat h_t(\xi(t)), \quad t \in [0, T]. \tag{2.4}$$

By lifting the vector field $\hat{h}_t$ to the cotangent bundle, we define the time-dependent Hamiltonian

$$\widehat{H} : [0,T] \times T^*M \to \mathbb{R}, \quad (t,\ell) \mapsto \langle \ell, \hat{h}_t(\pi\ell)\rangle,$$

where $\pi : T^*M \to M$ is the canonical projection; let us set $H_i := \widehat{H}_{|(t_{i-1},t_i)}$.

For our problem the maximized Hamiltonian

$$H : T^*M \to \mathbb{R}, \quad \ell \mapsto \max_{u\in\Delta} \left\langle \ell, \sum_{i=1}^{m} u_i\, X_i(\pi\ell) \right\rangle$$

is well defined and Lipschitz.

Recall that any piecewise smooth Hamiltonian $H_t : T^*M \to \mathbb{R}$ defines a Hamiltonian vector field $\vec{H}_t$ whose flow will be denoted by $\mathcal{H}_t$. Moreover for any time-independent vector field Y we denote its flow by $(t,x) \mapsto \exp t\, Y(x)$; see [Arn80].

We can express the PMP by saying that there exist $p_0 \in \{0,1\}$ and a lift $\hat{\lambda}$ of $\hat{\xi}$ to the cotangent bundle, which is a solution of

$$\dot{\lambda}(t) = \vec{\widehat{H}}_t(\lambda(t)),$$

$$\lambda(0) = p_0\, dc_0(x_0) \quad \text{on } T_{x_0}N_0, \tag{2.5}$$

$$\lambda(T) = -p_0\, dc_T(x_T) \quad \text{on } T_{x_T}N_T \tag{2.6}$$

such that

$$|p_0| + \|\hat{\lambda}\| \neq 0,$$
$$\widehat{H}_t(\hat{\lambda}(t)) = H(\hat{\lambda}(t)).$$

Let us now introduce our first assumption.

Assumption 2.1 (bang-bang regular extremal). *The maximum*

$$\max_{u\in\Delta} \left\langle \hat{\lambda}(t), \sum_{i=1}^{m} u_i\, X_i(\hat{\xi}(t)) \right\rangle$$

is attained at a vertex of Δ *for all* $t \in [0,T],\ t \neq t_1, t_2, \dots, t_r$.

Assumption 2.1 means that on each subinterval $(t_{i-1},\, t_i)$ there is a unique index j_i such that

$$H(\hat{\lambda}(t)) = \langle \hat{\lambda}(t),\, X_{j_i}(\hat{\xi}(t))\rangle.$$

The smooth functions $p_0\, c_0$ and $p_0\, c_T$ are defined on N_0 and N_T, respectively, but they can be extended to the whole manifold M in such a way that the transversality conditions (2.5) and (2.6) hold on the whole tangent space. We denote by $\alpha, \beta : M \to \mathbb{R}$ two functions such that

$$\alpha = p_0\, c_0 \quad \text{on } N_0, \quad \beta = p_0\, c_T \quad \text{on } N_T,$$

$$\hat{\lambda}(0) = d\,\alpha(x_0) \quad \text{on } T_{x_0}M, \quad \hat{\lambda}(T) = -d\,\beta(x_T) \quad \text{on } T_{x_T}M. \tag{2.7}$$

Consider the two Lagrangian submanifolds

$$\Lambda_0 := \left\{ d\,\alpha(x) + (T_xN_0)^\perp \ \middle|\ x \in N_0 \right\},$$
$$\Lambda_T := \left\{ -d\,\beta(x) + (T_xN_T)^\perp \ \middle|\ x \in N_T \right\};$$

the transversality conditions (2.5) and (2.6) of the PMP can be equivalently stated by saying that

$$\lambda(0) \in \Lambda_0, \quad \lambda(T) \in \Lambda_T.$$

In the normal case ($p_0 = 1$) α, β are cost functions equivalent to the original ones, while in the abnormal case ($p_0 = 0$) they are extensions of the zero function. When $p_0 = 0$ all the costs disappear, and indeed we will study a problem with a zero cost; therefore, proving that $\hat{\xi}$ is a *strict* strong minimizer will imply that it is isolated with respect to the C^0 topology among the admissible trajectories. In the case of sub-Riemannian metrics isolated trajectories are called rigid geodesics.

The sufficient conditions will be derived by studying the following optimal control problem, which is equivalent to the original one:

$$\text{(P)} \qquad \text{Minimize} \quad J(\xi) := \alpha(\xi(0)) + \beta(\xi(T))$$

subject to (2.1) and (2.2).

The points

$$\ell_i := \hat{\lambda}(t_i), \quad i = 0, 1, \dots, r+1,$$

will be called the *switching points* of the adjoint covector $\hat{\lambda}$. From the PMP we can deduce the following relations, which represent necessary optimality conditions:

$$H_i(\ell_i) = H_{i+1}(\ell_i), \quad i = 1, 2, \dots, r,$$
$$\langle d(H_{i+1} - H_i), \vec{H}_{i+1} \rangle(\ell_i) \geq 0, \quad i = 1, 2, \dots, r.$$

To state sufficient conditions for $\hat{\xi}$ to be a strong local minimizer we need to strengthen these two conditions, and hence we assume the following.

Assumption 2.2 (simple switching points). *The maximum*

$$\max_{u \in \Delta} \left\langle \hat{\lambda}(t), \sum_{i=1}^{m} u_i \, X_i(\hat{\xi}(t)) \right\rangle$$

is attained along a one-dimensional edge of Δ *for* $t = t_1, t_2, \dots, t_r$.

Assumption 2.3 (strict bang-bang Legendre condition).

$$\langle d(H_{i+1} - H_i), \vec{H}_{i+1} \rangle(\ell_i) > 0, \quad i = 1, 2, \dots, r.$$

Remark 2.4. The PMP implies that the switching point ℓ_i belongs to the level set $H_{i+1} - H_i = 0$ for $i = 1, 2, \dots, r$. The strict bang-bang Legendre condition yields that, near the point ℓ_i, this level set is a hypersurface which will be called the *switching surface*.

Our assumptions are strictly related to the properties of the flow of the maximized Hamiltonian H, and they guarantee that the Hamiltonian flow is piecewise smooth; see Corollary 4.2. In particular, Assumptions 2.1 yields that locally around the reference extremal we can switch from one vector field to another only on the switching surfaces, while Assumption 2.2 yields that, on the switching surfaces, we can choose only between two vector fields, and the last one, Assumption 2.3, yields that we are forced to switch. Let $\boldsymbol{s}$ be the Liouville one form in T^*M and denote by $\boldsymbol{\sigma} = d\boldsymbol{s}$ the

canonical symplectic two form on T^*M; see [Arn80] for the definitions. Taking into account the basic properties of $\boldsymbol{\sigma}$, Assumption 2.3 can be equivalently written as

$$\begin{aligned}\langle d(H_{i+1}-H_i), \vec{H}_{i+1}\rangle(\ell_i) &= \boldsymbol{\sigma}\left(\vec{H}_i, \vec{H}_{i+1}\right)(\ell_i)\\ &= \{H_i, H_{i+1}\}(\ell_i)\\ &= \langle \ell_i, [h_i, h_{i+1}](\hat{\xi}(t_i)\rangle \geq 0,\end{aligned}$$

where $\{\ ,\ \}$ and $[\ ,\ ]$ denote the Poisson and Lie brackets, respectively.

2.1. A finite-dimensional subproblem. We are going to choose an appropriate r-dimensional family of variations corresponding to bang-bang trajectories; they are generated by perturbing the switching times. The optimality with respect to these variations will be not only necessary but also sufficient (under the previously stated assumptions) to prove that the reference trajectory is a strong local minimizer.

For a given $a > 0$ such that $\min_{i=1,\dots,r+1}(t_i - t_{i-1}) > 2a$, let $\varepsilon \in B(0,a) \subset \mathbb{R}^r$, set $\varepsilon_0 = \varepsilon_{r+1} = 0$, and consider the time-dependent vector field

$$(\varepsilon, x) \mapsto h_t(\varepsilon, x) = h_i(x) \quad \text{if} \quad t \in (t_{i-1} + \varepsilon_{i-1}, t_i + \varepsilon_i).$$

This new vector field is obtained from the reference one by moving the switching time t_i by ε_i.

Remark 2.5. A small ε corresponds to a control variation which is small in the L^1 norm but not in the L^∞ norm.

Denote the flow of $\dot{\xi}(t) = h_t(\varepsilon, \xi(t))$ by

$$S_t : M \times B(0,a) \to M$$

and consider the following finite-dimensional subproblem of problem (P):

$$\text{(sub-P)} \qquad \text{Minimize} \quad \gamma(x,\varepsilon) := \alpha(x) + \beta(S_T(x,\varepsilon))$$

subject to

$$x \in N_0, \quad S_T(x,\varepsilon) \in N_T.$$

Note that $\widehat{S}_t := S_t(\cdot, 0)$ is the flow of $\hat{h}_t$, $S_t(x_0, 0) = \hat{\xi}(t)$ and $(x_0, 0)$ is the candidate optimal solution for the subproblem.

By using the relations (2.7) and the extremality properties of the reference trajectory, it is easy to prove that $(x_0, 0)$ is a critical point for γ, that is,

$$d\gamma(x_0, 0) = 0,$$

and hence

$$J'' := \frac{1}{2} D^2\gamma(x_0, 0)$$

is a well-defined quadratic form on $T_{x_0}M \times \mathbb{R}^r$, which gives the second order approximation of γ. The second variation of (sub-P) is the restriction of J'' to the linearization of the constraints; namely, if we set

$$\mathcal{N} = \left\{(\delta x, \varepsilon) \in T_{x_0}N_0 \times \mathbb{R}^r : S_{T*}(\delta x, \varepsilon) \in T_{x_T}N_T\right\}, \tag{2.8}$$

then the second variation of (sub-P) is $J''_{|\mathcal{N}}$, and it will be called the *second variation at the switching points*. Let us remark that $J''_{|\mathcal{N}} \geq 0$ is a necessary optimality condition when the subproblem is normal.

The main result of the paper states that under the regularity conditions on the maximized Hamiltonian previously stated, the positivity of the second variation at the switching points is sufficient to prove that the reference trajectory $(\hat{\xi}, \hat{u})$ is a strict local minimizer for the original problem in the C^0 topology on the state (strong minimizer).

From an intuitive point of view the idea underlying this result can be summarized by saying that the flow of the maximized Hamiltonian projects onto the trajectories of the finite-dimensional subproblem, which then generates a field of extremals that can be used to prove sufficiency.

THEOREM 2.6. *Assume that the given bang-bang Pontryagin extremal $\hat{\xi}$ is regular and has simple switching points and that the strict bang-bang Legendre condition is satisfied. If the second variation at the switching points is positive definite, then $\hat{\xi}$ is a strict strong local minimizer, i.e., a strict local minimizer in the C^0 topology. In the abnormal case $\hat{\xi}$ is an isolated admissible solution of the constrained control system.*

Remark 2.7. If $\mathcal{N}$ reduces to $\{0\}$, then the second variation at the switching points is positive definite, and hence we obtain a first order sufficient condition.

2.2. The bang-bang Jacobi system. We can check the positivity of the second variation at the switching points in a complete Hamiltonian form. Let

$$\Pi := T_{\pi\,\ell} M \hookrightarrow T_\ell T^* M$$

be the *vertical* subspace and define

$$L_0 := T_{\ell_0} \Lambda_0, \quad L_T := T_{\ell_T} \Lambda_T.$$

The regularity assumptions on the maximized Hamiltonian yield that $\mathcal{H}_t$ is smooth everywhere except at the switching times where it is left and right smooth; see Corollary 4.2. The positivity of the second variation at the switching points can be checked through the properties of the tangent subspaces to $\mathcal{H}_{t_k}(\Lambda_0)$ from the left and from the right and by their relative positions with respect to Π. Thanks to the strict bang-bang Legendre condition, these Lagrangian subspaces can be described through the flow of the following discrete version of the Jacobi system:

$$\text{(2.9)} \qquad \begin{cases} \delta\ell_k^- = \left(\exp(t_{k+1} - t_k)\, \vec{H}_{k+1}\right)_*\, \delta\ell_{k-1}^+, \\ \delta\ell_k^+ = \delta\ell_k^- + \dfrac{\boldsymbol{\sigma}(\delta\ell_k^-, (\vec{H}_k - \vec{H}_{k+1})(\ell_k))}{\boldsymbol{\sigma}(\vec{H}_k, \vec{H}_{k+1})(\ell_k)}\, (\vec{H}_k - \vec{H}_{k+1})(\ell_k). \end{cases}$$

Denote the flow of $\delta\ell_k^-$ and $\delta\ell_k^+$ by Δ_k^-, Δ_k^+ and set

$$L_k^- := \Delta_k^- L_0, \quad L_k^+ := \Delta_k^+ L_0.$$

In section 4 we prove that L_k^+, L_k^- are the left and right tangent subspaces to $\mathcal{H}_{t_k}(\Lambda_0)$; see Remark 4.6.

LEMMA 2.8. *The positivity of the second variation at the switching points can be checked through the following algorithm.*

STEP 1: Set $k = 1$.

STEP 2: If $k \leq r$, then go to STEP 3
else go to STEP 4.

STEP 3: If $(\vec{H}_k - \vec{H}_{k+1})(\ell_k) \notin L_k^- + \Pi$ or $(\vec{H}_k - \vec{H}_{k+1})(\ell_k) \in L_k^-$, then set $k = k+1$ and go to STEP 2.
else
if $L_k^+ \cap \Pi \subseteq L_k^- \cap \Pi$ and for every $\delta\ell^+ \in L_k^+$, $\delta\ell^- \in L_k^-$ such that $\pi_* \delta\ell^+ = \pi_* \delta\ell^-$ we have that $\boldsymbol{\sigma}(\delta\ell^-, \delta\ell^+) \geq 0$, then set $k = k+1$ and go to STEP 2.
else $J''_{|\mathcal{N}}$ is not positive definite, STOP.

STEP 4: If for every $\delta\ell \in L_{r+1}^-$, $\delta\ell_T \in L_T$ such that $\pi_* \delta\ell = \pi_* \delta\ell_T \neq 0$ we have that $\boldsymbol{\sigma}(\delta\ell, \delta\ell_T) > 0$, then $J''_{|\mathcal{N}}$ is positive definite, END.
else $J''_{|\mathcal{N}}$ is not positive definite, STOP.

Remark 2.9. Let us explain the meaning of each step of this algorithm.

1. The algorithm first checks the positivity of the second variation associated to the corresponding problem with fixed final point (STEP 3).
2. Each iteration of the algorithm is associated to a new variation obtained by perturbing the corresponding switching time; this procedure generates an increasing family of variations.
3. STEP 3 deserves some comment: if $(\vec{H}_k - \vec{H}_{k+1})(\ell_k) \notin L_k^- + \Pi$, then there is no new variation, and hence there is no condition to check; if $(\vec{H}_k - \vec{H}_{k+1})(\ell_k) \in L_k^-$, then the flow $\mathcal{H}_t$ is differentiable also at t_k and the properties of the second variation remain unchanged.
4. STEP 4 checks the positivity conditions related to the presence of a nontrivial final cost, and hence when the final point is fixed, STEP 4 is void and the algorithm becomes the following.

   STEP 1: $k = 1$.

   STEP 2: If $k \leq r$, then go to STEP 3
   else $J''_{|\mathcal{N}}$ is positive definite, END.

   STEP 3: If $(\vec{H}_k - \vec{H}_{k+1})(\ell_k) \notin L_k^- + \Pi$ or $(\vec{H}_k - \vec{H}_{k+1})(\ell_k) \in L_k^-$, then set $k = k+1$ and go to STEP 2.
   else
   if $L_k^+ \cap \Pi \subseteq L_k^- \cap \Pi$ and for every $\delta\ell^+ \in L_k^+$, $\delta\ell^- \in L_k^-$ such that $\pi_* \delta\ell^+ = \pi_* \delta\ell^-$ we have $\boldsymbol{\sigma}(\delta\ell^-, \delta\ell^+) \geq 0$, then set $k = k+1$ and go to STEP 2.
   else $J''_{|\mathcal{N}}$ is not positive definite, STOP.

Since the maximized Hamiltonian is a piecewise lift of a vector field on M, then the vertical directions remain vertical under the action of the flow, and at the switching points the dimension of the projection increases at most by one. To obtain a flow which projects locally onto M we will reduce the problem to an equivalent one with free initial point; for this reason we describe explicitly the algorithm in this special case.

COROLLARY 2.10. *If the initial point is free, we have that* $L_0 \cap \Pi = \{0\}$, *and hence the algorithm becomes the following.*

STEP 1: Set $k = 1$.

STEP 2: If $k \leq r$, then go to STEP 3
else go to STEP 4.

STEP 3: If $(\vec{H}_k - \vec{H}_{k+1})(\ell_k) \in L_k^-$, then set $k = k+1$ and go to STEP 2.
else
if $\pi_* L_k^+ = T_{\hat{\xi}(t_k)} M$ and for every $\delta\ell^+ \in L_k^+$, $\delta\ell^- \in L_k^-$ such that $\pi_*\delta\ell^+ = \pi_*\delta\ell^-$ we have that $\boldsymbol{\sigma}(\delta\ell^-, \delta\ell^+) \geq 0$, then set $k = k+1$ and go to STEP 2.
else $J''_{|\mathcal{N}}$ is not positive definite, STOP.

STEP 4: If for every $\delta\ell \in L_{r+1}^-$, $\delta\ell_T \in L_T$ such that $\pi_*\delta\ell = \pi_*\delta\ell_T \neq 0$ we have that $\boldsymbol{\sigma}(\delta\ell, \delta\ell_T) > 0$, then $J''_{|\mathcal{N}}$ is positive definite, END.
else $J''_{|\mathcal{N}}$ is not positive definite, STOP.

Remark 2.11. In STEP 3 we check the fixed final point problem, and the algorithm stops when we find a direction on which the quadratic form is negative or zero. For this reason we call the corresponding switching time t_k the *conjugate point*; a conjugate point can occur only at a switching time.

2.3. The Bolza problem. We deal with an optimal control problem in the Mayer form only for simplicity; all the results can be stated for a problem in the Bolza form when the cost function includes an integral term, that is,

$$\text{Minimize} \quad J(\xi) := c_0(\xi(0)) + c_T(\xi(T)) + \int_0^T \sum_{i=1}^m u_i(t)\, X_i^0(\xi(t))\, dt$$

subject to (2.1) and (2.2), where X_i^0, $i = 1, 2, \ldots, m$, are C^∞ functions defined on M.

The same proofs can be carried out using as reference and maximized Hamiltonian those defined as

$$\widehat{H} : \ell \mapsto \langle \ell, \hat{h}(\pi\ell) \rangle - p_0 \sum_{i=1}^m \hat{u}_i(t)\, X_i^0(\pi\ell),$$

$$H : \ell \mapsto \max_{u \in \Delta} \left(\left\langle \ell, \sum_{i=1}^m u_i\, X_i(\pi\ell) \right\rangle - p_0 \sum_{i=1}^m u_i\, X_i^0(\pi\ell) \right).$$

3. The second variation at the switching points. This section is necessarily technical, but it contains the main ideas and the technical lemmas needed to carry out this kind of approach.

To study the relations existing between the second variation at the switching points and the properties of the Hamiltonian flow, let us reduce (sub-P) to a single-input affine problem with piecewise constant control maps having the t_i's as switching times. This reduction can be achieved by the following time reparametrization:

$$\dot{\varphi}(\tau) = 1 + \nu(\tau), \quad \nu \in (-1, 1),$$
$$\varphi(0) = 0, \quad \varphi(T) = T,$$

where ν is piecewise constant, i.e., $\nu(\tau) \equiv \nu_i$, $\tau \in [t_{i-1}, t_i)$.

Any solution of this boundary value problem is an increasing isomorphism of the interval $[0, T]$ onto itself. If we set $\varepsilon_i := \varphi(t_i) - t_i$, $i = 1, 2, \ldots, r$, then we have that $S_{\varphi(\tau)}(x, \varepsilon)$ is the solution of the differential equation

$$\dot{\zeta}(\tau) = [1 + \nu(\tau)]\, \hat{h}_\tau(\zeta(\tau)), \quad \zeta(0) = x. \tag{3.1}$$

If we set $u_i := \nu_i\,(t_i - t_{i-1})$, then from our construction it follows that $\varepsilon_i = \sum_{j=1}^{i} u_j$ and $\sum_{j=1}^{r+1} u_j = 0$ as it follows from the boundary condition $\varphi(T) = T$. Therefore we can take as control space the r-dimensional vector space

$$U := \left\{ (u_1, u_2, \dots, u_{r+1}) \in \mathbb{R}^{r+1} \;\middle|\; \sum_{j=1}^{r+1} u_j = 0 \right\}.$$

For a given $u \in U$ we denote by ν_u the corresponding control map. Since there is a one-to-one correspondence between ε and u we still denote by $\mathcal{N}$ the subset of $T_{x_0}N_0 \times U$ corresponding to $\mathcal{N} \subseteq T_{x_0}N_0 \times \mathbb{R}^r$, which is defined in (2.8). We can now study the second variation of the problem of minimizing $\alpha(\zeta(0)) + \beta(\zeta(T))$ subject to (3.1) with the boundary conditions $\zeta(0) \in N_0$ and $\zeta(T) \in N_T$.

Following the same approach used in [ASZ98a] we can define the second variation as a linear quadratic problem on $T_{x_0}\,M$ by the using pull-back system defined through the time-dependent vector field

$$\hat{g}_t := \widehat{S}_{t*}^{-1}\,\hat{h}_t \circ \widehat{S}_t. \tag{3.2}$$

$\hat{g}_t$ is piecewise constant with the same switching times as $\hat{h}_t$, and we set $g_i := \hat{g}_{|(t_{i-1},t_i)}$. Consider the pull-back control system

$$\dot{\eta}(t) = \nu(t)\,\hat{g}_t(\eta(t)) \tag{3.3}$$

and the associated linearized equation at $\eta(t) \equiv x_0$,

$$\dot{\delta\eta}(t) = \nu(t)\,\hat{g}_t(x_0). \tag{3.4}$$

If we also pull back the costs by setting

$$\hat{\beta} = \beta \circ \widehat{S}_T, \quad \hat{\gamma} = \alpha + \hat{\beta},$$

then, reasoning as in [ASZ98a], the second variation at the switching points can be equivalently written as the restriction to $\mathcal{N}$ of the linear-quadratic form

$$J''[\delta e]^2 = \frac{1}{2} D^2\hat{\gamma}(x_0)[\delta x]^2 + \int_0^T \nu_u(s)\,\langle Q_s, \delta\eta_s(\delta e)\rangle\,ds,$$

where $\delta e := (\delta x, u) \in T_{x_0}M \times U$ and

$$\langle Q_t, \delta x\rangle = \langle D\langle D\,\hat{\beta}, \hat{g}_t\rangle(x_0),\, \delta x\rangle.$$

The Hamiltonian associated with this linear-quadratic problem is

$$(\omega, \delta x, u) \mapsto G_t''(\omega, \delta x)\nu_u(t),$$

where G_t'' is the following piecewise constant linear Hamiltonian:

$$G_t'' : T_{x_0}^*M \times T_{x_0}M \to \mathbb{R}, \quad (\omega, \delta x) \mapsto \langle \omega, \hat{g}_t(x_0)\rangle + \langle Q_t, \delta x\rangle.$$

With notation analogous to previous ones, we set $G_i'' := G''_{\,|(t_{i-1},t_i)}$ and define the Lagrangian subspace of the initial and final transversality conditions as

$$L_0'' := \left\{ \left(-D^2\hat{\gamma}(x_0)(\delta x, \cdot) + \omega,\, \delta x\right) \;|\; \delta x \in T_{x_0}N_0,\; \omega \in \left(T_{x_0}N_0\right)^{\perp} \right\},$$

$$L_T'' := \left(T_{x_0}\widehat{S}_T^{-1}(N_T)\right)^{\perp} \times T_{x_0}\widehat{S}_T^{-1}(N_T).$$

We want to express the value of the form J'' in Hamiltonian notation. Let $\delta e = (\delta x, u)$ and $\delta f = (\delta y, v)$ belong to $T_{x_0}M \times U$ and let $\delta\ell \in L_0''$ be such that

$$\pi_* \delta\ell = \delta x.$$

If we denote by $\mathcal{G}_t''(\delta\ell, u) := (\omega_t(\delta\ell, u), \delta\eta_t(\delta e))$ the solution of the Hamiltonian system

$$\dot\lambda(t) = \vec G_t''(\lambda(t))\, \nu_u(t), \quad \lambda(0) = \delta\ell, \tag{3.5}$$

then we obtain, as in the proof of Lemma 4 in [ASZ98a, p. 700],

$$\begin{aligned} J''(\delta e, \delta f) &= D^2\hat\gamma(x_0)(\delta x, \delta y) + \int_0^T \langle Q_s, \nu_u(s)\delta\eta_s(\delta f) + \nu_v(s)\delta\eta_s(\delta e)\rangle\, ds \\ &= D^2\hat\gamma(x_0)(\delta x, \delta y) - \langle \omega_T(\delta\ell, u), \delta\eta_T(\delta f)\rangle + \langle \omega_0(\delta\ell, u), \delta y\rangle \\ &\quad + \int_0^T G_t''\Big(\mathcal{G}_t''(\delta\ell, u)\Big)\, \nu_v(t)\, dt. \end{aligned} \tag{3.6}$$

The positivity of the second variation at the switching points will be checked in two steps. We first consider the problem with fixed final point and check the positivity of the corresponding second variation, that is, J'' restricted to

$$V := \{(\delta x, u) \in T_{x_0}N_0 \times U \mid \delta\eta(\delta x, \nu_u, T) = 0\} \subseteq \mathcal{N}.$$

Afterwards we check the positivity of J'' on $\mathcal{N} \cap V^{\perp_{J''}}$, where $\perp_{J''}$ means orthogonality with respect to J''.

To study the signature of the second variation on V we take an increasing sequence of subspaces $V_k \subset V$ obtained by considering as admissible controls those u for which ν_u is zero from t_{k+1} on; i.e., we will study the second variation on each

$$V_k := \{(\delta x, u) \in V \mid u_j = 0 \text{ for } j \geq k+2\}.$$

The extremals of J'' on V are essential in the study of its signature, and they are those δe belonging to $V \cap V^{\perp_{J''}}$. For this reason we characterize the J''-orthogonality in the following integral version of the Jacobi system.

LEMMA 3.1. *For $k \in \{1, 2, \ldots, r\}$, $\delta e = (\delta x, u) \in \mathcal{N} \cap V_k^{\perp_{J''}}$ if and only if there exists $\delta\ell \in L_0''$ such that*

$$\pi_* \delta\ell = \delta x, \quad \pi_*\, \mathcal{G}_T''(\delta\ell, u) \in \pi_*\, L_T'', \tag{3.7}$$

$$\int_0^T G_t''\Big(\mathcal{G}_t''(\delta\ell, u)\Big)\, \nu_v(t)\, dt = 0 \quad \forall v:\ v_j = 0,\ j \geq k+2. \tag{3.8}$$

Proof. $\delta e \in \mathcal{N} \cap V_k^{\perp_{J''}}$ if and only if there exist $\bar\omega_0 \in (T_{x_0}N_0)^\perp$ and $\bar\omega_T \in \Pi$ such that

$$J''(\delta e, \delta f) = \langle \bar\omega_0,\, \delta y\rangle + \langle \bar\omega_T,\, \delta\eta_T(\delta f)\rangle$$

for all $\delta f = (\delta y, v) \in T_{x_0}M \times U$ such that $v_j = 0,\ j \geq k+2$.

If we choose

$$\delta\ell = (-D^2\hat\gamma_0(\delta x, \cdot) + \bar\omega_0,\, \delta x),$$

then $\delta\ell \in L_0''$ and (3.7) is satisfied. For $v = 0$, from (3.6) we obtain

$$-\omega_T(\delta\ell, u) = \bar{\omega}_T$$

and hence (3.8).

To prove the converse, let us remark that (3.7) yields that $\delta e \in \mathcal{N}$; moreover, using (3.6) for $\delta f \in V_k$, from (3.7) and (3.8), it follows that $J''[\delta e, \delta f] = 0$. ∎

COROLLARY 3.2. *Let $\delta e = (\delta x, u) \in \mathcal{N} \cap V^{\perp_{J''}}$ and let $\delta\ell \in L_0''$ be the one given in Lemma* 3.1. *If $\delta\ell_1 \in L_T''$ is such that $\pi_*\delta\ell_1 = \pi_*\mathcal{G}_T''(\delta\ell, u)$, then*

$$J''\,[\delta e]^2 = \boldsymbol{\sigma}\Big(\delta\ell_1, \mathcal{G}_T''(\delta\ell, u)\Big). \tag{3.9}$$

Let $\delta e = (\delta x, u) \in V_k \cap V_{k-1}^{\perp_{J''}}$ and let $\delta\ell \in L_0''$ be the one given in Lemma 3.1. *Then*

$$J''[\delta e]^2 = \boldsymbol{\sigma}\Big(\mathcal{G}_{t_k}''(\delta\ell, u), u_{k+1}(\vec{G}_{k+1}'' - \vec{G}_k'')\Big). \tag{3.10}$$

Proof. Equality (3.9) is an easy consequence of Lemma 3.1 and (3.6). Integrating by parts and using the symplectic properties of the Hamiltonian flow, again from (3.6), it follows that

$$\begin{aligned} J''[\delta e]^2 &= \int_{t_{k-1}}^{t_k} G_k''\Big(\mathcal{G}_t''(\delta\ell, u)\Big)\,\frac{-u_{k+1}}{t_k - t_{k-1}}\,dt + \int_{t_k}^{t_{k+1}} G_{k+1}''\Big(\mathcal{G}_t''(\delta\ell, u)\Big)\,\nu_u(t)\,dt \\ &= -u_{k+1}G_k''\Big(\mathcal{G}_{t_k}''(\delta\ell, u)\Big) + u_{k+1}G_{k+1}''\Big(\mathcal{G}_{t_{k+1}}''(\delta\ell, u)\Big) \\ &= u_{k+1}\Big(G_{k+1}'' - G_k''\Big)\Big(\mathcal{G}_{t_k}''(\delta\ell, u)\Big). \end{aligned}$$

Equality (3.10) now follows thanks to the symplectic properties of the Hamiltonian flow. ∎

Let us remark that (3.7) characterizes those δe in $\mathcal{N}$, while (3.8) characterizes those in $V_k^{\perp_{J''}}$. In particular the extremals of the second variation are described by those $\delta\ell \in L_0''$, $u \in U$ such that $\mathcal{G}_T''(\delta\ell, u) \in L_T''$, and

$$\int_0^T G_t''\Big(\mathcal{G}_t''(\delta\ell, u)\Big)\,\nu_v(t)\,dt = 0 \quad \forall\ v \in U. \tag{3.11}$$

The relations between the second variation and the Hamiltonian of the original problem can be better understood by using the following map:

$$\boxed{\imath : T_{x_0}^*M \times T_{x_0}M \to T_{\ell_0}T^*M, \quad (\omega, \delta x) \mapsto -\omega + d(-\hat{\beta})_*\delta x}$$

It is easy to check that the map $\imath$ is an antisymplectic isomorphism

$$\boldsymbol{\sigma}\,(\imath\,\delta\ell_1, \imath\,\delta\ell_2\,) = -\boldsymbol{\sigma}\,(\delta\ell_1, \delta\ell_2\,) \tag{3.12}$$

and that it is an isomorphism between L_0'' and L_0 which acts as

$$\imath\,\delta\ell = d\alpha_*\pi_*\delta\ell.$$

The map $\imath$ connects the Hamiltonians associated with the second variation with the original ones through the following relation:

$$\imath\,\vec{G}_k'' = \widehat{\mathcal{H}}_{t_{k*}}^{-1}\vec{H}_k(\ell_k) = \widehat{\mathcal{H}}_{t_{k-1}*}^{-1}\vec{H}_k(\ell_{k-1}). \tag{3.13}$$

Equation (3.13) can be proved starting from the equality

$$\vec{G}''_k = \Big(-D\langle D\,\hat{\beta}, g_k\rangle(x_0),\, g_k(x_0)\Big)$$

and applying the map $\imath$ to obtain, in coordinates,

$$\imath\,\vec{G}''_k = \Big(d\hat{\beta}(x_0)\,Dg_k(x_0), g_k(x_0)\Big);$$

finally since $d\hat{\beta}(x_0) = -d\alpha(x_0)$, then (3.13) follows.

Thanks to the above properties of the map $\imath$ we can restate the *strict bang-bang Legendre condition* as

$$\boldsymbol{\sigma}\left(\vec{G}''_k,\, \vec{G}''_{k+1}\right) < 0, \quad k = 1, 2, \ldots, r. \tag{3.14}$$

The strict bang-bang Legendre condition allows us to solve recursively equation (3.11) with respect to the control, and hence we are able to define a discrete version of the Jacobi system by substituting this control back into (3.5). The resulting system is defined below, and its construction is described in the subsequent Lemma 3.4.

Definition 3.3. *Suppose that the strict bang-bang Legendre condition is satisfied and consider the discrete dynamical system on* $\mathbb{R} \times T^*(T_{x_0} M)$,

$$\begin{cases} \mathrm{w}_k = \dfrac{\boldsymbol{\sigma}\Big(\delta\ell_{k-1}, \vec{G}''_k - \vec{G}''_{k+1}\Big)}{\boldsymbol{\sigma}\Big(\vec{G}''_k, \vec{G}''_{k+1}\Big)}, \\ \\ \delta\ell_k = \delta\ell_{k-1} + (\vec{G}''_k - \vec{G}''_{k+1})\,\mathrm{w}_k. \end{cases}$$

For $k = 1, 2, \ldots, r$ *we define the flows of* w_k *and* $\delta\ell_k$ *as the linear functions*

$$\omega_k : L''_0 \to \mathbb{R}$$

and the symplectic isomorphisms

$$\mathcal{G}^k_T : L''_0 \to T^*_{x_0} M \times T_{x_0} M.$$

Lemma 3.4. *Suppose that the strict bang-bang Legendre condition is satisfied and let* $(\delta\ell, u) \in L''_0 \times U$*; then* (3.8) *holds if and only if*

$$u_i = \langle(\omega_i - \omega_{i-1}),\, \delta\ell\rangle, \quad i = 1, 2, \ldots, k,$$
$$\mathcal{G}''_{t_i}(\delta\ell, u) = \mathcal{G}^i_T(\delta\ell) + \langle\omega_i,\, \delta\ell\rangle\, \vec{G}''_{i+1}, \quad i = 1, 2, \ldots, k.$$

Proof. From the properties of the Hamiltonian flows, by integrating by parts equality (3.8), it follows that

$$\int_0^T G''_t\Big(\mathcal{G}''_t(\delta\ell, u)\Big)\,\nu_v(t)\,dt$$
$$= \sum_{i=1}^{k} v_i \left(G''_i\Big(\mathcal{G}''_{t_i}(\delta\ell, u)\Big) - G''_{k+1}\Big(\mathcal{G}''_{t_{k+1}}(\delta\ell, u)\Big)\right) = 0$$

for all $v \in U$ such that $v_j = 0,\ j \geq k+2$. Hence (3.8) is equivalent to

$$G''_1\Big(\mathcal{G}''_{t_1}(\delta\ell, u)\Big) = G''_2\Big(\mathcal{G}''_{t_2}(\delta\ell, u)\Big) = \cdots = G''_{k+1}\Big(\mathcal{G}''_{t_{k+1}}(\delta\ell, u)\Big).$$

If we compute explicitly

$$\Big(G''_i - G''_{i+1}\Big)(\mathcal{G}''_{t_{i-1}}(\delta\ell, u) + u_i\,\vec{G}''_i) = 0,$$

we obtain, for each $i = 1, 2, \dots, k$,

$$\boldsymbol{\sigma}\Big(\mathcal{G}''_{t_{i-1}}(\delta\ell, u) + u_i\,\vec{G}''_i, \vec{G}'''_i - \vec{G}'''_{i+1}\Big) = 0.$$

The equivalence now follows by finite induction and from Definition 3.3. □

Remark 3.5. Let us remark that being J''-orthogonal to V_k implies that the values of the control maps $u_1, u_2, \dots, u_k$ are uniquely determined by the value of $\delta\ell$. Moreover we obtain the flow of the Hamiltonian system of the second variation up to time t_k through the flow of the discrete bang-bang Jacobi system. More precisely, if we define the control $\mathrm{w}^k : L''_0 \to U$ as

$$\begin{cases} \mathrm{w}^k_i := (\omega_i - \omega_{i-1}), & i = 1, 2, \dots, k, \\ \mathrm{w}^k_{k+1} := -\omega_k, & \\ \mathrm{w}^k_i := 0, & i \geq k+2, \end{cases} \tag{3.15}$$

then from Lemma 3.4 it follows that this control is such that $(\delta\ell, \langle \mathrm{w}^k,\, \delta\ell\rangle)$ satisfies (3.8) and

$$\mathcal{G}^k_T(\delta\ell) = \mathcal{G}''_T(\delta\ell, \langle \mathrm{w}^k,\, \delta\ell\rangle).$$

A possible way to check the positivity of J'' on V_k is to study the behavior of J'' on $V_k \cap V_{k-1}^{\perp_{J''}}$. Thanks to the properties of the bang-bang Jacobi system the variations belonging to $V_k \cap V_{k-1}^{\perp_{J''}}$ and the values of J'' can be described through the following subspaces:

$$L''_k := \mathcal{G}^k_T\, L''_0.$$

The results are given in the two following lemmas. Let us notice that the first statement of the next lemma states that the extremals of J'' on V_k are the solutions of the Jacobi system that become vertical at step k and that the third statement characterizes the occurrence of a new variation.

Lemma 3.6. *Suppose that the strict bang-bang Legendre condition is satisfied. The following statements hold:*

1. *$\delta e = (\delta x, u) \in V_k \cap V_k^{\perp_{J''}}$ if and only if there exists $\delta\ell \in L''_0$ such that*

$$\pi_*\delta\ell = \delta x, \quad u = \langle \mathrm{w}^k,\, \delta\ell\rangle, \quad \mathcal{G}^k_T(\delta\ell) \in \Pi.$$

2. *$\delta e \in V_k \cap V_{k-1}^{\perp_{J''}}$ if and only if there exists $\delta\ell \in L''_0$ such that*

$$\pi_*\delta\ell = \delta x, \quad u_j = \langle \mathrm{w}^{k-1}_j,\, \delta\ell\rangle, \quad j = 1, 2, \dots, k-1,$$
$$\mathcal{G}^{k-1}_T(\delta\ell) - u_{k+1}\,(\vec{G}''_k - \vec{G}''_{k+1}) \in \Pi,$$

and in this case we have that

$$J''[\delta e]^2 = \boldsymbol{\sigma}\Big(\mathcal{G}^{k-1}_T(\delta\ell) - u_{k+1}\vec{G}''_k, -u_{k+1}(\vec{G}''_k - \vec{G}'''_{k+1})\Big). \tag{3.16}$$

3. *If* $J''_{|V_{k-1}} > 0$, *then* $V_k = V_{k-1}$ *if and only if* $\vec{G}''_k - \vec{G}''_{k+1} \notin L''_{k-1} + \Pi$.
4. *If* $\delta\ell_{k-1} \in L''_{k-1}$ *and* $\delta\ell_k \in L''_k \setminus L''_{k-1}$ *are such that*

$$\pi_* \delta\ell_{k-1} = \pi_* \delta\ell_k,$$

then there exists a nontrivial $\delta e \in V_k \cap V_{k-1}^{\perp_{J''}}$ *such that*

$$J[\delta e]^2 = \boldsymbol{\sigma}\left(\delta\ell_k,\, \delta\ell_{k-1}\right).$$

Proof. 1. From (3.7) and (3.8) and from the properties of the control (3.15) it follows that

$$u = \langle \mathrm{w}^k,\, \delta\ell \rangle \quad \text{and} \quad \mathcal{G}''_T(\delta\ell, u) = \mathcal{G}^k_T(\delta\ell) \in \Pi.$$

2. Letting $\delta e \in V_k \cap V_{k-1}^{\perp_{J''}}$ be the first part is an immediate consequence of Lemmas 3.1 and 3.4. Once again from the properties of the control (3.15) we have that

$$\begin{aligned}\mathcal{G}''_{t_{k+1}}(\delta\ell, u) &= \mathcal{G}^{k-1}_T(\delta\ell) + \langle \omega_{k-1},\, \delta\ell \rangle\, \vec{G}''_k + u_k\, \vec{G}''_k + u_{k+1}\, \vec{G}''_{k+1} \\ &= \mathcal{G}^{k-1}_T(\delta\ell) - u_{k+1}\,(\vec{G}''_k - \vec{G}''_{k+1}).\end{aligned}$$

From (3.10) it follows that

$$\begin{aligned}J''[\delta e]^2 &= \boldsymbol{\sigma}\Big(\mathcal{G}''_{t_k}(\delta\ell, u),\, -u_{k+1}(\vec{G}''_k - \vec{G}''_{k+1})\Big) \\ &= \boldsymbol{\sigma}\Big(\mathcal{G}^{k-1}_T(\delta\ell) - u_{k+1}\vec{G}''_k, -u_{k+1}(\vec{G}''_k - \vec{G}''_{k+1})\Big).\end{aligned}$$

3. $\vec{G}''_k - \vec{G}''_{k+1} \notin L''_{k-1} + \Pi$ if and only if

$$\mathcal{G}^{k-1}_T(\delta\ell) - u_{k+1}\,(\vec{G}''_k - \vec{G}''_{k+1}) \in \Pi \Rightarrow u_{k+1} = 0.$$

From statement 2 we have that $J[\delta e]^2 = 0$, and hence the statement follows.

4. By definition there are $\delta\ell_0,\, \delta\ell_1 \in L''_0$ such that

$$\delta\ell_{k-1} = \mathcal{G}^{k-1}_T(\delta\ell_0), \quad \delta\ell_k = \mathcal{G}^k_T(\delta\ell_1).$$

From the assumptions we have that

$$\mathcal{G}^k_T(\delta\ell_1) - \mathcal{G}^{k-1}_T(\delta\ell_0) = \mathcal{G}^{k-1}_T(\delta\ell_1) - \mathcal{G}^{k-1}_T(\delta\ell_0) + \langle \omega_k, \delta\ell_1 \rangle\, (\vec{G}''_k - \vec{G}''_{k+1}) \in \Pi.$$

If we define $\delta\ell := \delta\ell_1 - \delta\ell_0$ and $\delta e := (\pi_*\delta\ell, \langle \mathrm{w}^k,\, \delta\ell_1 \rangle - \langle \mathrm{w}^{k-1},\, \delta\ell_0 \rangle)$, then we have that $u_{k+1} = -\langle \omega_k, \delta\ell_1 \rangle$, and from statement 2 we have that $\delta e \in V_k \cap V_{k-1}^{\perp_{J''}}$, and it is nontrivial because if $u_{k+1} = 0$, then $\delta\ell_k \in L''_{k-1}$. Moreover we have that

$$\begin{aligned}J''[\delta e]^2 &= \boldsymbol{\sigma}\left(\mathcal{G}^{k-1}_T(\delta\ell_1 - \delta\ell_0) - u_{k+1}\,\vec{G}''_k,\, -u_{k+1}(\vec{G}''_k - \vec{G}''_{k+1})\right) \\ &= \boldsymbol{\sigma}\left(-\mathcal{G}^{k-1}_T(\delta\ell_0),\, \mathcal{G}^k_T(\delta\ell_1)\right) + \boldsymbol{\sigma}\left(\mathcal{G}^{k-1}_T(\delta\ell_0),\, \mathcal{G}^{k-1}_T(\delta\ell_1)\right) \\ &\quad - u_{k+1}\,\boldsymbol{\sigma}\left(\mathcal{G}^{k-1}_T(\delta\ell_1) + \langle \omega_k,\, \delta\ell_1 \rangle\, \vec{G}''_k,\, \vec{G}''_k - \vec{G}''_{k+1}\right).\end{aligned}$$

The final statement now follows because the second addend is zero since both the arguments belong to the same Lagrangian subspace L''_{k-1} and the third one is zero by the properties of ω_k. ∎

LEMMA 3.7. *$J''_{|V} > 0$ if and only if one of the following statements holds for each $k = 1, 2, \dots, r$:*

1. $\vec{G}''_k - \vec{G}''_{k+1} \notin L''_{k-1} + \Pi$.
2. $\vec{G}''_k - \vec{G}''_{k+1} \in L''_{k-1}$.
3. $L''_k \cap \Pi \subseteq L''_{k-1} \cap \Pi$ *and for all* $\delta\ell_k \in L''_k$ *and* $\delta\ell_{k-1} \in L''_{k-1}$ *such that* $\pi_* \delta\ell_k = \pi_* \delta\ell_{k-1}$, *we have that* $\boldsymbol{\sigma}(\delta\ell_k, \delta\ell_{k-1}) \geq 0$.

Proof. The idea of the proof is the following: we first show that these conditions together with $J''_{|V_{k-1}} > 0$ imply that

$$J''_{|V_k \cap V_{k-1}^{\perp_{J''}}} > 0$$

and hence $J''_{|V_k} > 0$; since $V_0 = \{0\}$, then the lemma will follow by finite induction on k. Let us show that the induction step is valid for each k if and only if one of the statements of the lemma holds.

Assume that $J''_{|V_{k-1}} > 0$.

- From statement 2 of Lemma 3.6 it follows that $\vec{G}''_k - \vec{G}''_{k+1} \notin L''_{k-1} + \Pi$ is equivalent to $V_k \cap V_{k-1}^{\perp_{J''}} = \{0\}$ and the induction step is trivial.
- If $\vec{G}''_k - \vec{G}''_{k+1} \in L''_{k-1}$, then $L''_k = L''_{k-1}$; moreover we can choose $u_{k+1} = 1$ in part 2 of Lemma 3.6 to show that $V_k \cap V_{k-1}^{\perp_{J''}} \neq \{0\}$ to obtain a nontrivial $\delta e \in V_k \cap V_{k-1}^{\perp_{J''}}$ such that

$$\begin{aligned} J''[\delta e]^2 &= \boldsymbol{\sigma}\Big(\mathcal{G}_T^{k-1}(\delta\ell) - \vec{G}''_k, -(\vec{G}''_k - \vec{G}''_{k+1})\Big) \\ &= \boldsymbol{\sigma}\Big(\vec{G}''_{k+1}, \vec{G}''_k\Big). \end{aligned}$$

  Equation (3.14) completes the proof.
- If $\vec{G}''_k - \vec{G}''_{k+1} \in \{L''_{k-1} + \Pi\} \setminus L''_{k-1}$, then $\dim V_k \cap V_{k-1}^{\perp_{J''}} = 1$. From the first statement of Lemma 3.6 it follows that the condition $L''_k \cap \Pi \subseteq L''_{k-1} \cap \Pi$ is equivalent to

$$V_k \cap V_k^{\perp_{J''}} = V_{k-1} \cap V_{k-1}^{\perp_{J''}} = \{0\},$$

  and hence it will be enough to prove that $J''_{|V_k \cap V_{k-1}^{\perp_{J''}}} \geq 0$.

Under our assumptions there is $\delta\ell \in L''_{k-1}$ such that $\langle \omega_k, \delta\ell \rangle = 1$. If we set

$$\delta\ell_k := \delta\ell + (\vec{G}''_k - \vec{G}''_{k+1}) \in L''_k,$$

then we can find $\delta\ell_{k-1} \in L''_{k-1}$ such that $\pi_* \, \delta\ell_{k-1} = \pi_* \, \delta\ell_k$. From the fourth statement of Lemma 3.6 it follows that

$$J[\delta e]^2 = \boldsymbol{\sigma}\left(\delta\ell_k, \delta\ell_{k-1}\right).$$

Since $\dim V_k \cap V_{k-1}^{\perp_{J''}} = 1$ then $J_{|V_k \cap V_{k-1}^{\perp_{J''}}} \geq 0$ if and only if $\boldsymbol{\sigma}\left(\delta\ell_k, \delta\ell_{k-1}\right) \geq 0$. ∎

LEMMA 3.8. *Assume that $J''_{|V} > 0$; then the quadratic form $J''_{|\mathcal{N} \cap V^{\perp_{J''}}}$ is positive definite if and only if for every $\delta\ell \in L''_0$ and $\delta\ell_T \in L''_T$ such that*

$$\pi_* \, \delta\ell_T = \pi_* \, \mathcal{G}_T^r(\delta\ell) \neq 0$$

we have

$$\boldsymbol{\sigma}(\delta\ell_T, \mathcal{G}^r_T(\delta\ell)) > 0.$$

Proof. From Lemma 3.1 and the properties of the control w^k (see (3.15)), we have that $\delta e = (\delta x, u) \in \mathcal{N} \cap V^{\perp_{J''}}$ if and only if there is $\delta\ell \in L_0''$ such that

$$\pi_*\delta\ell = \delta x, \quad u = \langle \mathrm{w}^r, \delta\ell\rangle, \quad \pi_*\mathcal{G}^r_T(\delta\ell) \in \pi_* L''_T.$$

If $\pi_*\mathcal{G}^r_T(\delta\ell) = 0$, then $\delta e \in V \cap V^{\perp_{J''}}$ and hence $\delta e = 0$; otherwise we can use equation (3.9). □

3.1. The algorithm. We have essentially already shown that the algorithm can be used to check the positivity of the second variation at the switching points. This can be easily seen since from (3.12), (3.13), it follows that

$$L_k^- = \widehat{\mathcal{H}}_{t_{k*}} \imath\, L''_{k-1} \quad \text{for } k = 1, 2, \dots, r+1,$$
$$L_k^+ = \widehat{\mathcal{H}}_{t_{k*}} \imath\, L''_k \quad \text{for } k = 1, 2, \dots, r.$$

Moreover STEP 3 follows from Lemma 3.7, while STEP 4 follows from Lemma 3.8, taking into account (3.12).

4. Proof of the theorem. In order to demonstrate the Hamiltonian method we now give the proof of our main result step by step following the approach described in the introduction.

All the proofs make strong use of the properties (see [Arn80]) of the Poincaré–Cartan form $\boldsymbol{\omega} = \boldsymbol{s} - H\,dt$ on $I \times T^*M$ associated to the Hamiltonian H. Namely,

- $\boldsymbol{\omega}$ evaluated along a lift of a solution of (2.1) is nonpositive and it is zero along $\hat{\lambda}$;
- $\boldsymbol{\omega}$ is exact on the Legendre submanifold generated by the flow of $\vec{H}$ emanating from a Lagrangian submanifold.

4.1. Flow properties. The first step shows that our assumptions guarantee that the flow of the maximized Hamiltonian is locally well defined and piecewise C^∞ and describes the structure of the switching surfaces.

LEMMA 4.1. *There exists a neighborhood $\mathcal{U}$ of ℓ_0 such that we can define recursively for $i = 1, \dots, r$ the C^∞-maps*

$$\tau_i : \mathcal{U} \to \mathbb{R} \quad \textit{and} \quad \phi_i : \mathcal{U} \to T^*M$$

in the following way: set

$$\tau_0 := 0, \quad \phi_0 := Id.$$

The τ_i's are implicitly defined by

$$\begin{cases} (H_i - H_{i+1})\left(\exp \tau_i(\ell)\vec{H}_i(\phi_{i-1}(\ell))\right) = 0, \\ \tau_i(\ell_i) = t_i, \end{cases}$$

while the ϕ_i's are defined as

$$\phi_i := \ell \mapsto \exp(-\tau_i(\ell)\vec{H}_{i+1}) \circ \exp \tau_i(\ell)\vec{H}_i\,(\phi_{i-1}(\ell)).$$

The neighborhood $\mathcal{U}$ can be chosen such that

$$\sup_{\ell\in\mathcal{U}} \tau_i(\ell) < \inf_{\ell\in\mathcal{U}} \tau_{i+1}(\ell), \tag{4.1}$$

and the ϕ_i's are C^∞ symplectic diffeomorphisms.

Proof. Thanks to the strict bang-bang Legendre condition we can apply the implicit function theorem to show that the τ_i's are well defined and C^∞. Therefore, by continuity, we can guarantee that (4.1) holds. Let us show by induction that ϕ_i's are symplectic diffeomorphisms. From the definition of the τ_i's we have that

$$\boldsymbol{\sigma}\left(\vec{H}_i, \vec{H}_{i+1}\right) d\,\tau_i(\ell) = \boldsymbol{\sigma}\left((\exp \tau_i(\ell)\vec{H}_i)_* \phi_{i-1\,*}, \vec{H}_i - \vec{H}_{i+1}\right), \tag{4.2}$$

and from the definition of the ϕ_i's we have that

$$\begin{aligned} \phi_{i*} &= \exp(-\tau_i(\ell)\vec{H}_{i+1})_* \exp(\tau_i(\ell)\vec{H}_i)_* \phi_{i-1\,*} \\ &+\Big\{\exp(-\tau_i(\ell)\vec{H}_{i+1})_* \vec{H}_i\Big(\exp(\tau_i(\ell)\vec{H}_i(\phi_{i-1}(\ell))\Big) - \vec{H}_{i+1}(\phi_i(\ell))\Big\} d\,\tau_i(\ell). \end{aligned} \tag{4.3}$$

The result now follows from (4.2) and from the general fact that $\exp(s\,\vec{G})$ is a symplectic diffeomorphism for any Hamiltonian vector field $\vec{G}$. ∎

Let $\mathcal{U}$ be the neighborhood of ℓ_0 given in Lemma 4.1. If we set

$$\mathcal{O}_i := \Big\{(t,\ell) \mid \ell\in\mathcal{U},\ \tau_{i-1}(\ell) \le t \le \tau_i(\ell)\Big\} \subseteq [0,T]\times T^*M,$$

then the $\mathcal{O}_i$'s are $2\,n+1$-dimensional C^∞ submanifolds with boundary $\partial\mathcal{O}_i = S_{i-1}\cup S_i$, where

$$S_i := \mathcal{O}_i \cap \mathcal{O}_{i+1} = \Big\{(\tau_i(\ell),\ell)\ ,\quad \ell\in\mathcal{U}\Big\}.$$

From Lemma 4.1 we can easily deduce the following.

Corollary 4.2. *Under Assumptions* 2.1, 2.2, *and* 2.3 *the Hamiltonian system*

$$\begin{aligned} \dot\lambda(t) &= \vec{H}(\lambda(t)), \\ \lambda(0) &= \ell \end{aligned}$$

has a unique solution, which can be represented on $[0,T]\times\mathcal{U}$ by the map $\mathcal{H} : (t,\ell)\mapsto (t,\mathcal{H}_t(\ell))$ given by

$$\mathcal{H}_t(\ell) = \exp t\vec{H}_{i+1}(\phi_i(\ell)), \quad t\in[\tau_i(\ell),\tau_{i+1}(\ell)], \tag{4.4}$$

where $\tau_{r+1}\equiv T$; moreover the flow $\mathcal{H}$ is C^∞ on each $\mathcal{O}_i$.

Let us remark that every solution of the Hamiltonian system (4.4) has the same number of switches as the reference trajectory $\hat\xi$; moreover, from the above equation (4.4), we can deduce that for $t\in[\tau_{i-1}(\ell),\tau_i(\ell)]$ we can write

$$\mathcal{H}_t(\ell) = \exp(t-\tau_{i-1}(\ell))\vec{H}_i \circ\cdots\circ \exp(\tau_2(\ell)-\tau_1(\ell))\vec{H}_2 \circ \exp\tau_1(\ell)\vec{H}_1\,(\ell)$$

and the ϕ_i's can be written as

$$\phi_i(\ell) = \exp(-\tau_i(\ell)\vec{H}_{i+1}) \circ \mathcal{H}_{\tau_i(\ell)}\,(\ell).$$

Remark 4.3. We can interpret Lemma 4.1 as saying that, thanks to the strict bang-bang Legendre condition, we can define, in a tube around the adjoint covector, a time-dependent maximized Hamiltonian as

$$(t,\ell) \mapsto H_i(t,\ell) \quad \text{if} \quad \mathcal{H}^{-1}(t,\ell) \in \mathcal{O}_i.$$

This Hamiltonian switches from one vector field to another when its flow crosses the switching surfaces and hence when changing the vector field results in an energy increase. Assumptions 2.1 and 2.2 ensure that with this choice we obtain the maximized Hamiltonian.

4.2. Hamiltonian methods. For a general introduction to the use of these methods and their application to optimal control, we refer to [AG90, AG97].

Without loss of generality we can assume that $\Lambda_0 \subseteq \mathcal{U}$ and that Λ_0 is a smooth simply connected Lagrangian submanifold; if necessary we take the restriction to a neighborhood of x_0. Define

$$\Omega_i := \Big\{(t,\ell) \in \mathcal{O}_i \mid \ell \in \Lambda_0\Big\}, \quad \Sigma_i := \Omega_i \cap \Omega_{i+1},$$

and

$$\Omega := \bigcup_{i=1}^{r+1} \Omega_i.$$

The Ω_i's are $n+1$-dimensional C^∞ submanifolds with boundary $\partial\Omega_i = \Sigma_{i-1} \cup \Sigma_i$.

From (4.4) it follows that $\mathcal{H}_t(\Lambda_0)$ is a Lagrangian submanifold, although it might be not C^1 at the switching surfaces. We now investigate the properties of the Cartan form $\boldsymbol{\omega}$ and of the map

$$\pi_t := \pi \circ \mathcal{H}_t : \Lambda_0 \mapsto M.$$

LEMMA 4.4. *The form $\mathcal{H}^*\boldsymbol{\omega}$ is closed on each Ω_i and hence exact on Ω so that it can be written as*

$$\mathcal{H}^*\boldsymbol{\omega} = d\,\vartheta,$$

where ϑ is a continuous function on $[0,T] \times \Lambda_0$, which is C^∞ on each Ω_i. Moreover ϑ can be chosen such that

$$\vartheta(0,\cdot) := \vartheta_0 = \alpha \circ \pi.$$

If π_t is Lipschitz invertible, then

$$d(\vartheta_t \circ \pi_t^{-1}) = \mathcal{H}_t \circ \pi_t^{-1}.$$

Proof. The proof of the first statement is a standard consequence of the properties of $\boldsymbol{\omega}$ (see [Arn80]).

Let $\gamma : [a,b] \to M$ be a Lipschitz curve; then from the first part of the lemma it follows that

$$\int_\gamma \mathcal{H}_t \circ \pi_t^{-1} = \int_{\mathcal{H}_t \circ \pi_t^{-1} \circ \gamma} \boldsymbol{s} = \int_{\pi_t^{-1} \circ \gamma} \mathcal{H}_t^*\,\boldsymbol{s} = \vartheta_t \circ \pi_t^{-1}\Big|_{\gamma(a)}^{\gamma(b)},$$

and the statement follows. ∎

If $\lambda : [0,T] \to T^*M$ is a Lipschitz lift of a solution ξ of equation (2.1) such that $(t,\lambda(t)) \in \Omega$ for $t \in [0,T]$, then

$$\int_\lambda \boldsymbol{\omega} \le 0 \quad \text{and} \quad \int_{\hat\lambda} \boldsymbol{\omega} = 0 \tag{4.5}$$

because $\boldsymbol{\omega}$ is defined by the maximized Hamiltonian H. From this property and by the previous lemma, we obtain that

$$J(\pi\lambda) - J(\hat\xi) \ge \vartheta_T\Big(\mathcal{H}_T^{-1}(\lambda(T))\Big) - \vartheta_T(\ell_0) + \beta(\xi(T)) - \beta(x_T). \tag{4.6}$$

Hence, as we mentioned in the introduction, the variation of the cost is estimated from below by the function $\vartheta_T + \beta \circ \pi$, which depends only on the final point.

Let us remark that if π_T is invertible, then the same estimate can be obtained by the function

$$\chi := \vartheta_T \circ \pi_T^{-1} + \beta.$$

For this function, Lemma 4.4 and the transversality conditions imply that

$$d\,\chi(x_T) = 0.$$

4.3. An equivalent free initial point problem. To be able to lift to Ω any trajectory in a neighborhood of the reference one, we need that π_t is locally onto for each t and in particular for $t = 0$. This last condition can be fulfilled by constructing an equivalent problem with a free initial point. Let Q be any nonnegative quadratic form on $T_{x_0}M$, whose nullity is $T_{x_0}N_0$. We extend it to $T_{x_0}M \times \mathbb{R}^r$ by setting $Q[\delta x, \epsilon]^2 = Q[\delta x]^2$. If the quadratic form J'' is positive on $\mathcal{N}$, then we can find $\rho > 0$ such that

$$J'' + \frac{1}{2}\,\rho\, Q > 0 \quad \text{on } \Big\{(\delta x, \varepsilon) \in T_{x_0}M \times \mathbb{R}^r : S_{T*}(\delta x, \varepsilon) \in T_{x_T}N_T\Big\},$$

as can be easily proved by elementary arguments of linear algebra. Let us choose a function α_ρ such that

$$\begin{aligned} \alpha_\rho &= \alpha \quad \text{on } N_0, \\ d\alpha_\rho &= d\alpha \quad \text{on } T_{x_0}N_0, \\ D^2\alpha_\rho(x_0) &= D^2\alpha(x_0) + \rho\, Q \end{aligned}$$

and consider the problem

$$\text{Minimize} \quad \alpha_\rho(\xi(0)) + \beta(\xi(T)) \tag{4.7}$$

subject to

$$\begin{gathered} \dot\xi(t) = \sum_{i=1}^m u_i(t)\, X_i(\xi(t)), \quad u \in \Delta, \\ \xi(T) \in N_T. \end{gathered}$$

Since the reference trajectory satisfies the initial boundary conditions, then proving that it is optimal for this new problem yields its optimality for the original one. Therefore without loss of generality we can assume that the original problem has already free

initial point, i.e., $N_0 \equiv M$ and $\alpha \equiv \alpha_\rho$; in this case the initial Lagrangian submanifold is horizontal and its projection covers a neighborhood of the initial point x_0.

Remark 4.5. This reduction is possible because the new cost on the initial point contains an exact penalty which can be constructed assuming that the second variation is positive definite.

Let us now see which properties of the symplectic map $\mathcal{H}_{t*} : T_{\ell_0}\Lambda_0 \to T_{\hat\lambda(t)}M$ lead to the optimality of $\hat\xi$. Let us remark that (4.4) yields, for $\delta\ell \in L_0$,

$$\mathcal{H}_{t*}(\delta\ell) = (\exp t\vec{H}_{i+1})_*\phi_{i*}\delta\ell \quad \text{for } t \in (t_i, t_{i+1}), \tag{4.8}$$

$$\mathcal{H}_{t_i\,*}(\delta\ell) = \begin{cases} (\exp t_i\vec{H}_i)_*\phi_{i-1\,*}\delta\ell & \text{for } \langle d\,\tau_i(\ell_0), \delta\ell\rangle \le 0, \\ (\exp t_i\vec{H}_{i+1})_*\phi_{i*}\delta\ell & \text{for } \langle d\,\tau_i(\ell_0), \delta\ell\rangle \ge 0. \end{cases} \tag{4.9}$$

Remark 4.6. By (4.2) and (4.3) one can easily see that $L_k^- = (\exp t_k\vec{H}_k)_*\phi_{k-1\,*}\,L_0$ and $L_k^+ = (\exp t_k\vec{H}_{k+1})_*\phi_{k*}$; therefore L_k^- and L_k^+ are tangent to $\mathcal{H}_{t_k}(L_0)$ from the left and from the right, respectively. Moreover if $d\,\tau_k(\ell_0)_{|L_0} = 0$, then the flow is differentiable at (t_k, ℓ_0).

LEMMA 4.7. *If the map $\pi_*\,\mathcal{H}_{t*} : L_0 \to T_{\hat\xi(t)}M$ is onto for $t \in [0,T]$, then there exists a neighborhood $\mathcal{V} \subseteq \Lambda_0$ of ℓ_0 such that $[0,T]\times\mathcal{V}$ is mapped by $\pi\,\mathcal{H}$ onto a neighborhood of $\hat\xi$ in $[0,T]\times M$ and $\pi\,\mathcal{H}$ has a piecewise C^∞ local inverse. Without loss of generality we set $\mathcal{V} = \Lambda_0$.*

Proof. Thanks to the invertibility assumption on $\pi_*\,\mathcal{H}_{t_i\,*}$ and by possibly taking a smaller neighborhood of ℓ_0, we can apply the inverse function theorem on each submanifold with boundary Ω_i to show that the image under $\pi\,\mathcal{H}$ of Ω is a neighborhood of $\hat\xi$ in $[0,T]\times M$. □

THEOREM 4.8. *The equality*

$$\pi_*\mathcal{H}_{t*}\,L_0 = T_{\hat\xi(t)}M$$

holds for $t \in [0,T]$ if and only if the following statements hold for $i = 1, 2, \dots, r$:

1. $\pi_*\phi_{i*}L_0 = T_{x_0}M$.
2. *If $\delta\ell_1,\ \delta\ell_2 \in L_0$ are such that*

$$\pi_*\left[(\exp t_i\vec{H}_i)_*\phi_{i-1\,*}\delta\ell_1\right] = \pi_*\left[(\exp t_i\vec{H}_{i+1})_*\phi_{i*}\delta\ell_2\right],$$

then

$$\boldsymbol{\sigma}\left((\exp t_i\vec{H}_i)_*\phi_{i-1\,*}\delta\ell_1, (\exp t_i\vec{H}_{i+1})_*\phi_{i*}\delta\ell_2\right) \ge 0.$$

Proof. Since $\exp t\vec{H}_i$ transforms horizontal submanifolds into horizontal submanifolds, then (4.8)–(4.9) imply that the map $\pi_*\,\mathcal{H}_{t*}$ is onto for $t \in [0,T]$ if and only if it is onto for $t = t_i,\ i = 1, 2, \dots, r$.

Let us now check that conditions 1 and 2 are equivalent to

$$\pi_*\mathcal{H}_{t_i*}\,L_0 = T_{\hat\xi(t_i)}M. \tag{4.10}$$

If $d\,\tau_i(\ell_0) = 0$ on L_0, then the maps $(\exp t_i\vec{H}_i)_*\phi_{i-1\,*}$ and $(\exp t_i\vec{H}_{i+1})_*\phi_{i*}$ coincide, and hence (4.10) is equivalent to condition 1, and moreover condition 2 holds with the equality sign.

Otherwise we have that (4.10) holds if and only if

$$\pi_* \phi_{i-1\,*} L_0 = \pi_* \phi_{i*} L_0 = T_{x_0} M$$

and the two half-spaces

$$\left\{ \pi_* (\exp t_i \vec{H}_i)_* \phi_{i-1\,*} \delta\ell,\ \langle d\,\tau_i(\ell_0), \delta\ell \rangle \le 0 \right\},$$
$$\left\{ \pi_* (\exp t_i \vec{H}_{i+1})_* \phi_{i*} \delta\ell,\ \langle d\,\tau_i(\ell_0), \delta\ell \rangle \ge 0 \right\}$$

do not coincide. To check this we can show that for every $\delta\ell_1,\ \delta\ell_2 \in L_0$ such that

$$\pi_* \left[(\exp t_i \vec{H}_i)_* \phi_{i-1\,*} \delta\ell_1 \right] = \pi_* \left[(\exp t_i \vec{H}_{i+1})_* \phi_{i*} \delta\ell_2 \right]$$

one has

$$\langle d\,\tau_i(\ell_0), \delta\ell_1 \rangle \, \langle d\,\tau_i(\ell_0), \delta\ell_2 \rangle \ge 0.$$

By (4.3) we obtain that

$$\boldsymbol{\sigma} \left((\exp t_i \vec{H}_i)_* \phi_{i-1\,*} \delta\ell_1, (\exp t_i \vec{H}_{i+1})_* \phi_{i*} \delta\ell_2 \right)$$
$$= \boldsymbol{\sigma} \left((\exp t_i \vec{H}_i)_* \phi_{i-1\,*} \delta\ell_1, (\vec{H}_i - \vec{H}_{i+1})(\ell_i) \right) \langle d\,\tau_i(\ell_0), \delta\ell_2 \rangle.$$

Finally by the strict bang-bang Legendre condition and by (4.2) we obtain that

$$\boldsymbol{\sigma} \left((\exp t_i \vec{H}_i)_* \phi_{i-1\,*} \delta\ell_1, (\vec{H}_i - \vec{H}_{i+1})(\ell_i) \right)$$

has the same sign as $\langle d\,\tau_i(\ell_0), \delta\ell_1 \rangle$, and the statement is proved. ∎

Remark 4.9. Theorem 4.8 states that in the bang-bang case, a *conjugate point* can occur only at a switching time; moreover condition 1 states that the projection has full dimension, while condition 2 says that there is not a fold.

THEOREM 4.10. *Let the map* $\pi_* \mathcal{H}_{t*} : L_0 \to T_{\hat{\xi}(t)} M$ *be onto for* $t \in [0, T]$. *If the form*

$$\delta x \mapsto \boldsymbol{\sigma} \Big(d(\vartheta_T \circ \pi_T^{-1})_* \delta x, d(-\beta)_* \delta x \Big) = \boldsymbol{\sigma} \left((\mathcal{H}_T \circ \pi_T^{-1})_* \delta x, d(-\beta)_* \delta x \right)$$

is positive definite on $T_{x_T} N_T$, *then* $\hat{\xi}$ *is a strict strong local minimizer for the problem* (P).

Proof. By (4.6) if we prove that x_T is a local minimizer on N_T of $\chi = \vartheta_T \circ \pi_T^{-1} + \beta$, then $\hat{\xi}$ is a strong local minimizer for the problem (P). As we pointed out before, $d\,\chi(x_T) = 0$; thus the second derivative of χ is well defined at x_T, and we have that

$$D^2\,\chi(x_T)[\delta x]^2 = \boldsymbol{\sigma} \left(d(\vartheta_T \circ \pi_T^{-1})_* \delta x, d(-\beta)_* \delta x \right),$$

which ends the first part of the proof.

Let us now prove that the minimum is locally uniquely attained. First of all let us notice that under our assumptions $\hat{\xi}(T)$ is a strict local minimum for the function χ. Assume now by contradiction that there exists another admissible trajectory ξ with the same cost and the same final point $\xi(T) = \hat{\xi}(T)$; denote by $\lambda := t \mapsto \pi_t^{-1}(\xi(t))$ its lift. By (4.6) and (4.5) we have that

$$\int_\lambda \boldsymbol{\omega} = 0;$$

thus, since H is the maximized Hamiltonian, we have that

$$\langle \mathcal{H}_t(\lambda(t)), \dot{\xi}(t)\rangle - H(\mathcal{H}_t(\lambda(t))) = 0 \quad \text{a.e. } t \in [0,T].$$

Thanks to Assumption 2.1, for all t such that $(t,\lambda(t)) \in \operatorname{int} \Omega_i$ we have that

$$\dot{\xi}(t) = h_i(\xi(t));$$

this equation yields that

$$\dot{\lambda}(t) = 0,$$

and hence, from inequality (4.1), it follows that $\lambda(t)$ is constant, say k_i, for $t \in [\tau_{i-1}(k_i), \tau_i(k_i)]$. When $\tau_i(k_i) \le t \le \tau_i(k_{i+1})$ we have that $(t,\lambda(t)) \in \Sigma_i$, and hence there is $\mu \in [0,1]$ such that

$$\dot{\xi}(t) = (1-\mu)h_{i+1}(\xi(t)) + \mu h_i(\xi(t)).$$

For $t \in [\tau_r(k_r), T]$ we have that $\lambda(t) \in \Omega_{r+1}$, and hence λ is as smooth as ξ; moreover on the last time interval $[t_r, T]$ we have that $\lambda(t) = \ell_0$. Let us now show that λ cannot remain on the switching surface for a time interval of positive measure or, equivalently, that $\tau_r(k_r) = t_r$. By contradiction, if $k_r \neq \ell_0$, then

$$(t,\xi(t)) \in \pi\mathcal{H}(\Sigma_r) \subseteq \pi\mathcal{H}(\Omega_{r+1}), \quad t \in [\tau_r(k_r), t_r].$$

If we differentiate the identities

$$t = \tau_r(\lambda(t)), \quad \pi\mathcal{H}_t(\lambda(t)) = \xi(t),$$

we obtain, a.e. $t \in [\tau_r(k_r), t_r]$,

$$\langle d\,\tau_r(\lambda(t)), \dot{\lambda}(t)\rangle = 1, \quad \dot{\xi}(t) = h_{r+1}(\xi(t)) + \pi_* \mathcal{H}_{t\,*}\dot{\lambda}(t). \tag{4.11}$$

From Assumption 2.2 it follows that for a.a. $t \in [\tau_r(k_r), t_r]$ there exists $\mu_t \in [0,1]$ such that

$$\dot{\xi}(t) = (1-\mu_t)\, h_{r+1}(\xi(t)) + \mu_t\, h_r(\xi(t)),$$

and hence

$$\pi_* \mathcal{H}_{t\,*}\dot{\lambda}(t) = \mu_t \Big(h_r(\xi(t)) - h_{r+1}(\xi(t)) \Big).$$

Since $\dot{\lambda}(t)$ and μ_t are bounded we can take a sequence $t_i \to t_r$ such that $\dot{\lambda}(t_i) \to \delta\ell \in L_0$ and $\mu_{t_i} \to \mu$. Taking into account (4.11) we can say that $\delta\ell \neq 0$, and since we have that $\pi_* \mathcal{H}_{t_r\,*}$ is injective, then from

$$\pi_* \mathcal{H}_{t_r\,*}\delta\ell = \mu \Big(h_r(\xi(t_r)) - h_{r+1}(\xi(t_r)) \Big)$$

it follows that also $\mu \neq 0$. On the other hand from (4.4) and (4.3) we obtain

$$\begin{aligned} \pi_* \left(\exp t_r \vec{H}_r \right)_* \phi_{r-1\,*}\delta\ell &= \pi_* \mathcal{H}_{t_r\,*}\delta\ell - \Big(h_r(\xi(t_r)) - h_{r+1}(\xi(t_r)) \Big) \\ &= (\mu - 1) \Big(h_r(\xi(t_r)) - h_{r+1}(\xi(t_r)) \Big). \end{aligned}$$

Once again the last term has to be nonzero. If we set $\delta\ell_1 := \frac{\mu}{\mu-1}\delta\ell$, then

$$\langle d\,\tau_r(\ell_0), \delta\ell\rangle \, \langle d\,\tau_r(\ell_0), \delta\ell_1\rangle = \frac{\mu}{\mu-1} < 0,$$

and from Theorem 4.8 we obtain a contradiction. Therefore $k_r = \ell_0$, and hence $\tau_r(k_r) = \tau_r(\ell_0) = t_r$. We can do the same proof on each interval proceeding backwards in time to prove that the trajectory is constant. ∎

4.4. The proof. As we pointed out in Remark 4.6 if we have a regular bang-bang extremal with simple switching times and the strict bang-bang Legendre condition is satisfied, then

$$L_k^- = (\exp t_k \vec{H}_k)_* \varphi_{k-1\,*}\, L_0 \quad \text{for } k = 1, 2, \dots, r+1,$$
$$L_k^+ = (\exp t_k \vec{H}_{k+1})_* \varphi_{k\,*}\, L_0 \quad \text{for } k = 1, 2, \dots, r.$$

On the other hand we are considering a free initial point problem; therefore if $J''_{|V} > 0$, then from STEP 3 of the algorithm described in Corollary 2.10 it follows that we can apply Theorem 4.8, and hence we have that $\pi_* \mathcal{H}_{t*}\, L_0 = T_{\hat{\xi}(t)} M$ for all $t \in [0, T]$ and

$$L_{r+1}^- = \left(\mathcal{H}_T \circ \pi_T^{-1}\right)_* T_{x_T} M.$$

Therefore STEP 4 of the algorithm described in Corollary 2.10 yields that we can apply Theorem 4.10.

In the abnormal case the cost is zero, and hence the existence of a *strict* strong local minimizer is equivalent to the fact that the reference trajectory is isolated among the admissible trajectories in the C^0 topology.

REFERENCES

[AG90] A. A. Agrachev and R. V. Gamkrelidze, *Symplectic geometry for optimal control*, in Nonlinear Controllability and Optimal Control, Monogr. Textbooks Pure Appl. Math. 133, H. J. Sussmann, ed., Marcel Dekker, New York, 1990, pp. 263–277.

[AG97] A. A. Agrachev and R. V. Gamkrelidze, *Symplectic methods for optimization and control*, in Geometry of Feedback and Optimal Control, Monogr. Textbooks Pure Appl. Math. 207, B. Jacubzick and W. Respondek, eds., Marcel Dekker, New York, 1998, pp. 19–77.

[Arn80] V. I. Arnold, *Mathematical Methods in Classical Mechanics*, Springer, New York, 1980.

[ASZ98a] A. A. Agrachev, G. Stefani, and P. Zezza, *An invariant second variation in optimal control*, Internat. J. Control, 71 (1998), pp. 689–715.

[ASZ98b] A. A. Agrachev, G. Stefani, and P. Zezza, *Strong minima in optimal control*, Proc. Steklov Inst. Math., 220 (1998), pp. 4–26.

[ASZ99] A. A. Agrachev, G. Stefani, and P. Zezza, *A Hamiltonian approach to strong minima in optimal control*, in Differential Geometry and Control (Boulder, CO, 1997), AMS, Providence, RI, 1999, pp. 11–22.

[GH96a] M. Giaquinta and S. Hildebrandt, *Calculus of Variations* I, Springer, Berlin, 1996.

[GH96b] M. Giaquinta and S. Hildebrandt, *Calculus of Variations* II, Springer, Berlin, 1996.

[PS00] B. Piccoli and H. J. Sussmann, *Regular synthesis and sufficient conditions for optimality*, SIAM J. Control Optim., 39 (2000), pp. 359–410.

[Sar92] A. V. Sarychev, *Sufficient optimality conditions for Pontryagin extremals*, Systems Control Lett., 19 (1992), pp. 451–460.

[Sar97] A. V. Sarychev, *First- and second-order sufficient optimality conditions for bang-bang controls*, SIAM J. Control Optim., 35 (1997), pp. 315–340.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 4, pp. 1015–1041

A LEADER-FOLLOWER STOCHASTIC LINEAR QUADRATIC DIFFERENTIAL GAME*

JIONGMIN YONG†

Abstract. A leader-follower stochastic differential game is considered with the state equation being a linear Itô-type stochastic differential equation and the cost functionals being quadratic. We allow that the coefficients of the system and those of the cost functionals are random, the controls enter the diffusion of the state equation, and the weight matrices for the controls in the cost functionals are not necessarily positive definite. The so-called open-loop strategies are considered only. Thus, the follower first solves a stochastic linear quadratic (LQ) optimal control problem with the aid of a stochastic Riccati equation. Then the leader turns to solve a stochastic LQ problem for a forward-backward stochastic differential equation. If such an LQ problem is solvable, one obtains an open-loop solution to the two-person leader-follower stochastic differential game. Moreover, it is shown that the open-loop solution admits a state feedback representation if a new stochastic Riccati equation is solvable.

Key words. leader-follower stochastic differential game, linear quadratic optimal control problem, forward-backward stochastic differential equation, stochastic Riccati equation

AMS subject classifications. 93E20, 49K45, 49N10, 60H10

PII. S0363012901391925

1. Introduction.

Let $(\Omega, \mathcal{F}, \mathbf{P}, \{\mathcal{F}_t\}_{t\geq 0})$ be a complete filtered probability space on which a one-dimensional standard Brownian motion $W(\cdot)$ is defined such that $\{\mathcal{F}_t\}_{t\geq 0}$ is the natural filtration generated by $W(\cdot)$, augmented by all the $\mathbf{P}$-null sets in $\mathcal{F}$. We consider the following controlled linear stochastic differential equation (SDE):

$$\begin{cases} dx(t) = [A(t)x(t) + B_1(t)u_1(t) + B_2(t)u_2(t)]dt \\ \qquad\qquad + [C(t)x(t) + D_1(t)u_1(t) + D_2(t)u_2(t)]dW(t), \qquad t \in [0,T], \\ x(0) = \xi, \end{cases} \tag{1.1}$$

where $A(\cdot)$, $B_1(\cdot)$, $B_2(\cdot)$, $C(\cdot)$, $D_1(\cdot)$, and $D_2(\cdot)$ are matrix-valued $\{\mathcal{F}_t\}_{t\geq 0}$-adapted processes of suitable dimensions, and $\xi \in \mathbb{R}^n$, the standard n-dimensional Euclidean space. In the above, $x(\cdot)$ is the *state process* with values in $\mathbb{R}^n$, and $u_1(\cdot)$ and $u_2(\cdot)$ are *control processes* taken by the two players in the game, labeled 1 and 2, with values in $\mathbb{R}^{m_1}$ and $\mathbb{R}^{m_2}$, respectively. It is seen that the controls enter the diffusion of the state equation. Let $\mathcal{U}_1[0,T] \triangleq L^2_{\mathcal{F}}(0,T;\mathbb{R}^{m_1})$, the set of all $\mathbb{R}^{m_1}$-valued $\{\mathcal{F}_t\}_{t\geq 0}$-adapted processes $u_1 : [0,T]\times\Omega \to \mathbb{R}^{m_1}$ such that $E\int_0^T |u_1(t)|^2 dt < \infty$. The set $\mathcal{U}_2[0,T] \triangleq L^2_{\mathcal{F}}(0,T;\mathbb{R}^{m_2})$ is defined similarly. The control processes $u_1(\cdot)$ and $u_2(\cdot)$ are taken from $\mathcal{U}_1[0,T]$ and $\mathcal{U}_2[0,T]$, respectively.

Under some mild conditions on the coefficients, for any $(\xi, u_1(\cdot), u_2(\cdot)) \in \mathbb{R}^n \times \mathcal{U}_1[0,T]\times\mathcal{U}_2[0,T]$, there exists a unique (strong) solution $x(\cdot) \equiv x(\cdot\,;\xi, u_1(\cdot), u_2(\cdot)) \in$

*Received by the editors July 5, 2001; accepted for publication (in revised form) March 18, 2002; published electronically October 8, 2002. This work was supported in part by the NSFC under grant 10131030, the Chinese Education Ministry Science Foundation under grant 2000024605, and the Cheung Kong Scholars Programme.

http://www.siam.org/journals/sicon/41-4/39192.html

†Laboratory of Mathematics for Nonlinear Sciences, Department of Mathematics, and Institute of Mathematical Finance, Fudan University, Shanghai 200433, China (jyong@fudan.edu.cn).

$L^2_{\mathcal{F}}(0,T;\mathbb{R}^n)$ to (1.1). Thus, we can define the *cost functionals* for the players as follows:

$$J_i(\xi;u_1(\cdot),u_2(\cdot)) = E\left\{\int_0^T \big[\langle Q_i x(t),x(t)\rangle + \langle R_i u_i(t),u_i(t)\rangle\big]dt + \langle G_i x(T),x(T)\rangle\right\}, \qquad i=1,2, \tag{1.2}$$

where, for $i=1,2$, Q_i and R_i are $(n\times n)$ and $(m_i\times m_i)$ symmetric matrix-valued $\{\mathcal{F}_t\}_{t\geq 0}$-adapted processes, respectively, and G_i is an $(n\times n)$ symmetric matrix-valued $\mathcal{F}_T$-measurable random variable. We emphasize here that no nonnegative conditions are assumed on any of Q_i, R_i, and G_i. In particular, R_1 and R_2 are not necessarily positive definite. It is possible to consider more general forms of cost functionals (for example, $J_1(\cdot)$ also explicitly contains $u_2(\cdot)$, etc.). We take the above form for simplicity of the presentation in this paper.

Roughly speaking, in the above, Player i wants to minimize his/her own cost functional $J_i(\xi;u_1(\cdot),u_2(\cdot))$ by choosing a suitable control $u_i(\cdot)\in\mathcal{U}_i[0,T]$. We refer to this problem as a stochastic linear quadratic (LQ) differential game.

Let us now explain the leader-follower feature of the game. In the game, Player 2 is the leader, and Player 1 is the follower. For any choice $u_2(\cdot)\in\mathcal{U}_2[0,T]$ of Player 2 and a fixed initial state $\xi\in\mathbb{R}^n$, Player 1 would like to choose a $\overline{u}_1(\cdot)\in\mathcal{U}_1[0,T]$ so that $J_1(\xi;\overline{u}_1(\cdot),u_2(\cdot))$ is the minimum of $J_1(\xi;u_1(\cdot),u_2(\cdot))$ over $u_1(\cdot)\in\mathcal{U}_1[0,T]$. Knowing the follower would take such an optimal control $\overline{u}_1(\cdot)$ (supposing it exists, which depends on the choice $u_2(\cdot)$ of the leader and the initial state ξ, in general), Player 2 (the leader) would like to choose some $\overline{u}_2(\cdot)\in\mathcal{U}_2[0,T]$ to minimize $J_2(\xi;\overline{u}_1(\cdot),u_2(\cdot))$ over $u_2(\cdot)\in\mathcal{U}_2[0,T]$. We refer to such a problem as a *leader-follower stochastic LQ differential game*.

In a little more rigorous way, Player 1 wants to find a map $\overline{\alpha}_1:\mathcal{U}_2[0,T]\times\mathbb{R}^n\to\mathcal{U}_1[0,T]$ and Player 2 wants to find a control $\overline{u}_2(\cdot)\in\mathcal{U}_2[0,T]$ such that

$$\begin{cases} J_1(\xi;\overline{\alpha}_1[u_2(\cdot),\xi](\cdot),u_2(\cdot)) = \min\limits_{u_1(\cdot)\in\mathcal{U}_1[0,T]} J_1(\xi;u_1(\cdot),u_2(\cdot)) \quad \forall u_2(\cdot)\in\mathcal{U}_2[0,T], \\ J_2(\xi;\overline{\alpha}_1[\overline{u}_2(\cdot),\xi](\cdot),\overline{u}_2(\cdot)) = \min\limits_{u_2(\cdot)\in\mathcal{U}_2[0,T]} J_2(\xi;\overline{\alpha}_1[u_2(\cdot),\xi](\cdot),u_2(\cdot)). \end{cases} \tag{1.3}$$

If the above pair $(\overline{\alpha}_1[\cdot],\overline{u}_2(\cdot))$ exists, we refer to it as an *open-loop solution* to the above leader-follower differential game. Note that the map $\overline{\alpha}_1$ could be "anticipating"; i.e., $\overline{\alpha}_1[u_2(\cdot),\xi](t)$ could depend on the future value $u_2(s)$ ($s\in[t,T]$) of $u_2(\cdot)$ (see below). Thus, to make the open-loop solution (if it exists) more useful, we will make a great effort to find a state feedback representation for the open-loop solution. This can be regarded as one of the main contributions in this paper. See [26] for a similar result of feedback representation for the open-loop solution in a deterministic case.

We recall that there are several frameworks of studying differential games as far as the strategies are concerned. For open-loop strategies (or pure strategies) in finite-dimensional deterministic differential games, see [1, 3, 9, 26, 28]; the infinite-dimensional counterpart can be found in [2, 17, 22]; some corresponding stochastic cases can be found in [14, 15, 16]. For "nonanticipating" and/or closed-loop strategies of deterministic or stochastic differential games, there is substantial literature; see [1, 2, 4, 10, 11, 12, 13, 16, 17, 21, 22, 29, 30] and the references cited therein. In the present paper, we consider only open-loop strategies. The open-loop solution

and its state feedback representation will be discussed. It is also possible to consider problems with closed-loop strategies together with some kind of equilibrium. However, at this moment, we do not have publishable results for this. We hope to address these problems in future work.

Now, let us make some simple comparisons of our formulation with others in the literature. First, all the formulations of stochastic differential games in the above-mentioned literature have the restriction that the controls do not enter into the diffusion. Also, in most cases (except [14, 15, 16]), all the coefficients are deterministic. Hence, this paper might be the first one for random coefficient stochastic differential games with the diffusion that contains controls. Second, recently the stochastic LQ control problems in which the weight matrix for the control in the cost functionals are not necessarily nonnegative have been studied extensively [6, 7, 8, 34]. This kind of problem has some interesting applications in mathematical finance [34]. This paper might also be the first time that a similar kind of problem is formulated for differential games.

To summarize the above, we see that the novelty of the formulation in this paper is the following: (i) The controls enter into the diffusion, and all coefficients are possibly random; (ii) the weight matrices of the controls in the cost functionals are not necessarily nonnegative; and (iii) the state feedback representation of an open-loop solution to such a differential game is obtained via a new stochastic Riccati equation.

Next, let us briefly look at the procedure of finding an open-loop solution to the above leader-follower differential game. First, we solve an LQ problem for Player 1. For given $u_2(\cdot) \in \mathcal{U}_2[0,T]$, the follower (Player 1) wants to solve the following stochastic LQ problem.

Problem $(\mathrm{LQ})_1$. For given $\xi \in \mathbb{R}^n$, find a $\overline{u}_1(\cdot) \in \mathcal{U}_1[0,T]$ such that

$$J_1(\xi; \overline{u}_1(\cdot), u_2(\cdot)) = \inf_{u_1(\cdot) \in \mathcal{U}_1[0,T]} J_1(\xi; u_1(\cdot), u_2(\cdot)). \tag{1.4}$$

Any $\overline{u}_1(\cdot) \in \mathcal{U}_1[0,T]$ satisfying (1.4) is called an *optimal control*, the corresponding state process $\overline{x}(\cdot)$ is called an *optimal state process*, and $(\overline{x}(\cdot), \overline{u}_1(\cdot))$ is called an *optimal pair* of Problem $(\mathrm{LQ})_1$, respectively. This problem will be studied in section 2. Here, we state some relevant results only. To this end, let us introduce the following SDE:

$$\begin{cases} dP = -\big\{PA + A^TP + C^TPC + \Lambda C + C^T\Lambda + Q_1 \\ \qquad\quad - \big(PB_1 + \Lambda D_1 + C^TPD_1\big)\big(R_1 + D_1^TPD_1\big)^{-1}\big(B_1^TP + D_1^T\Lambda + D_1^TPC\big)\big\}dt \\ \qquad\qquad\qquad + \Lambda dW(t), \qquad t \in [0,T], \\ P(T) = G_1, \\ R_1 + D_1^TP(t)D_1 > 0, \qquad t \in [0,T], \text{ a.s.} \end{cases} \tag{1.5}$$

This is a terminal value problem for an SDE, which is known by now as a *backward stochastic differential equation* (BSDE; see [24, 27, 34] for details). We call (1.5) the *stochastic Riccati equation* for Problem $(\mathrm{LQ})_1$ (see [5, 6, 7, 8, 31, 34]). The unknown for (1.5) is the pair of $(n \times n)$ symmetric matrix-valued $\{\mathcal{F}_t\}_{t\geq 0}$-adapted processes $(P(\cdot), \Lambda(\cdot))$. If such a pair exists, we call it an *adapted solution* of (1.5). Note that (1.5) is a nonlinear BSDE with some singularities in the unknown $(P(\cdot), \Lambda(\cdot))$. In [7], a general local solvability result, as well as a couple of results on the solvability to some special case of such BSDEs, was presented. We should point out that the general solvability of (1.5) remains widely open. Now, suppose (1.5) admits an *adapted*

solution $(P(\cdot),\Lambda(\cdot))$. With such a pair at hand, we introduce another BSDE:

$$
\begin{cases}
d\varphi = -\Big\{[A^T-(PB_1+\Lambda D_1+C^TPD_1)(R_1+D_1^TPD_1)^{-1}B_1^T]\varphi \\
\qquad\quad +[C^T-(PB_1+\Lambda D_1+C^TPD_1)(R_1+D_1^TPD_1)^{-1}D_1^T]\theta \\
\qquad\quad +\big[-(PB_1+\Lambda D_1+C^TPD_1)(R_1+D_1^TPD_1)^{-1}D_1^TPD_2 \\
\qquad\quad +PB_2+\Lambda D_2+C^TPD_2\big]u_2\Big\}dt+\theta dW(t), \qquad t\in[0,T], \\
\varphi(T)=0.
\end{cases} \tag{1.6}
$$

Again, the unknown here is a pair $(\varphi(\cdot),\theta(\cdot))$ of $\mathbb{R}^n$-valued $\{\mathcal{F}_t\}_{t\geq 0}$-adapted processes. Now, let us formally state the following result for Problem $(\mathrm{LQ})_1$ (omitting some technical assumptions). A precise statement will be given in the next section.

THEOREM 1.1. *Suppose* (1.5)–(1.6) *admit adapted solutions* $(P(\cdot),\Lambda(\cdot))$ *and* $(\varphi(\cdot),\theta(\cdot))$*, respectively. Then Problem* $(\mathrm{LQ})_1$ *admits an optimal control* $\overline{u}_1(\cdot)$ *of a state feedback form,*

$$
\overline{u}_1(t)=-\widehat{R}_1(t)^{-1}\widehat{S}_1(t)x(t)-\widehat{R}_1(t)^{-1}f(t), \qquad t\in[0,T], \tag{1.7}
$$

where

$$
\begin{cases}
\widehat{R}_1\triangleq R_1+D_1^TPD_1, \qquad \widehat{S}_1\triangleq B_1^TP+D_1^T\Lambda+D_1^TPC, \\
f\triangleq B_1^T\varphi+D_1^T\theta+D_1^TPD_2u_2.
\end{cases} \tag{1.8}
$$

If we define the right-hand side of (1.7) as $\overline{\alpha}_1[u_2(\cdot),\xi](t)$, then the first equality in (1.3) holds. Note that in the above, $(P(\cdot),\Lambda(\cdot))$ does not depend on $(u_1(\cdot),u_2(\cdot))$, whereas $(\varphi(\cdot),\theta(\cdot))$ does depend on $u_2(\cdot)$ although it still does not depend on $u_1(\cdot)$. Moreover, since (1.6) is a BSDE, the value $(\varphi(t),\theta(t))$ of $(\varphi(\cdot),\theta(\cdot))$ at time t depends on $\{u_2(s)\bigm| s\in[0,T]\}$. Thus, $f(t)$ and hence $\overline{u}_1(t)$ depend on $\{u_2(s)\bigm| s\in[0,T]\}$ as well (see (1.7)–(1.8)). This means that $\overline{\alpha}_1[u_2(\cdot),\xi]\equiv\overline{u}_1(\cdot)$ is anticipating, in general (unless $B_1^T\varphi+D_1^T\theta=0$), as we pointed out earlier.

Once Problem $(\mathrm{LQ})_1$ is solved, we return to the leader (Player 2). Note that when the follower takes his optimal control $\overline{u}_1(\cdot)$ given by (1.7), the leader ends up with the following state equation:

$$
\begin{cases}
dx(t)=\big\{\widehat{A}x(t)+\widehat{F}_1\varphi(t)+\widehat{B}_1\theta(t)+\widehat{B}_2u_2(t)\big\}dt, \\
\qquad\qquad +\big\{\widehat{C}x(t)+\widehat{B}_1^T\varphi(t)+\widehat{D}_1\theta(t)+\widehat{D}_2u_2(t)\big\}dW(t), \\
d\varphi(t)=-\big\{\widehat{A}^T\varphi(t)+\widehat{C}^T\theta(t)+\widehat{F}_2^Tu_2(t)\big\}dt+\theta(t)dW(t), \\
x(0)=\xi, \qquad \varphi(T)=0,
\end{cases} \tag{1.9}
$$

where

$$
\begin{cases}
\widehat{A}\triangleq A-B_1\widehat{R}_1^{-1}\widehat{S}_1, \qquad \widehat{F}_1\triangleq -B_1\widehat{R}_1^{-1}B_1^T, \\
\widehat{B}_1\triangleq -B_1\widehat{R}_1^{-1}D_1^T, \qquad \widehat{B}_2\triangleq B_2-B_1\widehat{R}_1^{-1}D_1^TPD_2, \\
\widehat{C}\triangleq C-D_1\widehat{R}_1^{-1}\widehat{S}_1, \qquad \widehat{F}_2\triangleq \widehat{S}_2-D_2^TPD_1\widehat{R}_1^{-1}\widehat{S}_1, \\
\widehat{D}_1\triangleq -D_1\widehat{R}_1^{-1}D_1^T, \qquad \widehat{D}_2\triangleq D_2-D_1\widehat{R}_1^{-1}D_1^TPD_2, \\
\widehat{S}_2\triangleq B_2^TP+D_2^T\Lambda+D_2^TPC.
\end{cases} \tag{1.10}
$$

The cost functional is $J_2(\xi;\overline{u}_1(\cdot),u_2(\cdot))$, with $\overline{u}_1(\cdot)$ being of form (1.7), which is still a quadratic form. Equation (1.9) is a two-point boundary value problem for SDEs, which is what we call a *forward-backward stochastic differential equation* (FBSDE; see [24, 32, 33]). This FBSDE is decoupled. One can first solve the backward equation for $(\varphi(\cdot),\theta(\cdot))$, then solve the forward equation for $x(\cdot)$. Hence, we end up with an LQ problem for an FBSDE. Let us keep in mind that the "state" for (1.9) is the triple $(x(\cdot),\varphi(\cdot),\theta(\cdot))$. We mention here that the LQ control problem for BSDEs was studied in [23].

Solving the LQ problem for an FBSDE (1.9) itself is interesting. In solving this problem, we obtain an optimal control $\overline{u}_2(\cdot)$ of a "state" feedback via a new stochastic Riccati equation in which the "state" is $(x(\cdot),\varphi(\cdot),\theta(\cdot))$. Then $(\overline{u}_1(\cdot),\overline{u}_2(\cdot))$ gives an open-loop solution to the leader-follower differential game. We have seen that $\overline{u}_1(\cdot)$ is anticipating in general. Since $(\varphi(\cdot),\theta(\cdot))$ could be anticipating, then so could $\overline{u}_2(\cdot)$. Thus, the open-loop solution $(\overline{u}_1(\cdot),\overline{u}_2(\cdot))$ of the game is anticipating in general. The next big job in this paper is to represent $\overline{u}_1(\cdot)$ and $\overline{u}_2(\cdot)$ in terms of the original state $x(\cdot)$. This makes the whole current paper nontrivial and meaningful.

Readers are referred to [19] for some classical deterministic LQ problems, to [31, 5] for classical stochastic LQ problems, to [20, 25] for stochastic control systems with control entering the diffusion, and to [6, 7, 8, 34] for stochastic LQ problems with control entering into the diffusion of the state equation without assuming the nonnegativity of control weight in the cost functional. Finally, some classical results for differential games can be found in [18].

The rest of the paper is organized as follows. Section 2 is devoted to a brief study of the LQ problem for the follower, together with some other preliminary results. In section 3, we discuss the LQ problem for the leader. A new kind of Riccati equation is derived. In section 4, we present two one-dimensional cases. In section 5, we investigate the deterministic coefficients case, for which we obtain an open-loop solution and its state feedback representation to the leader-follower differential game.

2. LQ problem for the follower. Let us first introduce some notation which will be used throughout the paper.

Let $\mathbb{R}^{n\times m}$ be the set of all $(n\times m)$ matrices, and let $\mathcal{S}^n$ be the set of all $(n\times n)$ symmetric matrices. For any Banach space H (for example, $H=\mathbb{R}^n,\mathbb{R}^{n\times m},\mathcal{S}^n$), let $L^\infty_{\mathcal{F}}(0,T;H)$ (resp., $C_{\mathcal{F}}([0,T];H)$) be the set of all H-valued $\{\mathcal{F}_t\}_{t\geq 0}$-adapted bounded (resp., bounded continuous) processes. For any $1\leq p<\infty$, let $L^p_{\mathcal{F}}(0,T;H)$ be the set of all H-valued $\{\mathcal{F}_t\}_{t\geq 0}$-adapted processes $\varphi:[0,T]\times\Omega\to H$ such that $E\int_0^T|\varphi(t)|^p_H dt<\infty$, and let $L^p_{\mathcal{F}_T}(\Omega;H)$ be the set of all $\mathcal{F}_T$-measurable random variable $\eta:\Omega\to H$ such that $E|\eta|^p_H<\infty$. The spaces of deterministic functions $L^p(0,T;H)$ and $C([0,T];H)$ are defined in the usual way.

Let us introduce the following assumptions, which will be used later.

(S) Let

$$
(2.1)\quad \begin{cases} A,C\in L^\infty_{\mathcal{F}}(0,T;\mathbb{R}^{n\times n}),\\ B_i\in L^\infty_{\mathcal{F}}(0,T;\mathbb{R}^{n\times m_i}),\quad D_i\in C_{\mathcal{F}}([0,T];\mathbb{R}^{n\times m_i}),\qquad i=1,2,\\ Q_i\in C_{\mathcal{F}}([0,T];\mathcal{S}^n),\quad R_i\in C_{\mathcal{F}}([0,T];\mathcal{S}^{m_i}),\quad G_i\in L^\infty_{\mathcal{F}_T}(\Omega;\mathcal{S}^n),\qquad i=1,2. \end{cases}
$$

Note that under (S), the coefficients of the control system and the cost functional are random. Now, let us discuss Problem $(\mathrm{LQ})_1$. Similarly to [7, 8], we introduce the following notion.

DEFINITION 2.1. *For given $u_2(\cdot)\in\mathcal{U}_2[0,T]$, Problem* $(\mathrm{LQ})_1$ *is said to be*

(i) *finite at* $\xi \in \mathbb{R}^n$ *if*

$$\inf_{u_1(\cdot)\in\mathcal{U}_1[0,T]} J_1(\xi; u_1(\cdot), u_2(\cdot)) > -\infty; \tag{2.2}$$

(ii) *(uniquely) solvable at* $\xi \in \mathbb{R}^n$ *if there exists a (unique)* $\overline{u}_1(\cdot) \in \mathcal{U}_1[0,T]$ *such that* (1.4) *holds.*

Using a similar idea found in [8] (see [34] also), we are able to prove the following result.

PROPOSITION 2.2. *Let* (S) *hold and* $u_2(\cdot) \in \mathcal{U}_2[0,T]$ *be given.*

(i) *Let Problem* $(\mathrm{LQ})_1$ *be finite at some* $\xi \in \mathbb{R}^n$. *Then for any* $u_1(\cdot) \in \mathcal{U}_1[0,T]$, *the unique adapted solution* $(x^0(\cdot), p^0(\cdot), q^0(\cdot))$ *of the FBSDE*

$$\begin{cases} dx^0(t) = [Ax^0(t) + B_1u_1(t)]dt + [Cx^0(t) + D_1u_1(t)]dW(t), \\ dp^0(t) = -[A^Tp^0(t) + C^Tq^0(t) + Q_1x^0(t)]dt + q^0(t)dW(t), \\ x^0(0) = 0, \qquad p^0(T) = G_1x^0(T) \end{cases} \tag{2.3}$$

satisfies

$$E\int_0^T \langle\, R_1u_1(t) + B_1^Tp^0(t) + D_1^Tq^0(t), u_1(t)\,\rangle\, dt \geq 0. \tag{2.4}$$

(ii) *Let the conclusion of* (i) *hold. Then Problem* $(\mathrm{LQ})_1$ *is (uniquely) solvable with* $(\overline{x}(\cdot), \overline{u}_1(\cdot))$ *being a (the only) optimal pair if and only if there exists a (unique)* 4*-tuple* $(\overline{x}(\cdot), \overline{u}_1(\cdot), \overline{p}(\cdot), \overline{q}(\cdot))$ *satisfying the FBSDE*

$$\begin{cases} d\overline{x}(t) = [A\overline{x}(t) + B_1\overline{u}_1(t) + B_2u_2(t)]dt + [C\overline{x}(t) + D_1\overline{u}_1(t) + D_2u_2(t)]dW(t), \\ d\overline{p}(t) = -[A^T\overline{p}(t) + C^T\overline{q}(t) + Q_1\overline{x}(t)]dt + \overline{q}(t)dW(t), \\ \overline{x}(0) = \xi, \qquad \overline{p}(T) = G_1\overline{x}(T) \end{cases} \tag{2.5}$$

such that

$$R_1\overline{u}_1(t) + B_1^T\overline{p}(t) + D_1^T\overline{q}(t) = 0, \qquad t \in [0,T], \text{ a.s.} \tag{2.6}$$

Next, by the idea of "four-step scheme" found in [24] (see also [33, 34]), we can heuristically derive the stochastic Riccati equation (1.5) and BSDE (1.6). With these, we have the following sufficient condition for the solvability of Problem $(\mathrm{LQ})_1$, which makes the statement of Theorem 1.1 precise.

THEOREM 2.3. *Let* (S) *hold. Suppose* (1.5)–(1.6) *admit adapted solutions* $(P(\cdot), \Lambda(\cdot)) \in C_{\mathcal{F}}([0,T];\mathcal{S}^n) \times L^2_{\mathcal{F}}(0,T;\mathcal{S}^n)$ *and* $(\varphi(\cdot), \theta(\cdot)) \in C_{\mathcal{F}}([0,T];\mathbb{R}^n) \times L^2_{\mathcal{F}}(0,T;\mathbb{R}^n)$, *respectively, such that* $\widehat{R}_1 \equiv R_1 + D_1^TPD_1 > 0$, *and*

$$\begin{cases} B_1\widehat{R}_1^{-1}\widehat{S}_1, \qquad D_1\widehat{R}_1^{-1}\widehat{S}_1 \in L^\infty_{\mathcal{F}}(0,T;\mathbb{R}^{n\times n}), \\ B_1\widehat{R}_1^{-1}f, \qquad D_1\widehat{R}_1^{-1}f \in L^\infty_{\mathcal{F}}(0,T;\mathbb{R}^n), \end{cases} \tag{2.7}$$

where $\widehat{S}_1$ *and* f *are given by* (1.8). *Then Problem* $(\mathrm{LQ})_1$ *is solvable with the optimal control* $\overline{u}_1(\cdot)$ *being of a state feedback form* (1.7), *and*

$$\begin{aligned} \inf_{u_1(\cdot)\in\mathcal{U}_1[0,T]} J_1(\xi; u_1(\cdot), u_2(\cdot)) &= \langle\, P(0)\xi, \xi\,\rangle + 2\,\langle\, \varphi(0), \xi\,\rangle \\ &+ E\int_0^T \big\{2\,\langle\, B_2^T\varphi(t) + D_2^T\theta(t), u_2(t)\,\rangle + \langle\, D_2^TP(t)D_2u_2(t), u_2(t)\,\rangle \\ &\qquad - |\widehat{R}_1(t)^{-\frac{1}{2}}f(t)|^2\big\}dt \qquad \forall \xi \in \mathbb{R}^n. \end{aligned} \tag{2.8}$$

Proof. By our assumption, the following SDE admits a unique strong solution $\overline{x}(\cdot)$:

$$\begin{cases} d\overline{x}(t) = \big\{\big[A - B_1\widehat{R}_1^{-1}\widehat{S}_1\big]\overline{x}(t) - B_1\widehat{R}_1^{-1}f + B_2u_2(t)\big\}dt \\ \qquad\qquad + \big\{\big[C - D_1\widehat{R}_1^{-1}\widehat{S}_1\big]\overline{x}(t) - D_1\widehat{R}_1^{-1}f + D_2u_2(t)\big\}dW(t), \\ \overline{x}(0) = \xi. \end{cases} \tag{2.9}$$

It is clear that the control $\overline{u}_1(\cdot)$ defined by (1.7) is in $\mathcal{U}_1[0,T]$. Now, for any $u_1(\cdot) \in \mathcal{U}_1[0,T]$, $u_2(\cdot) \in \mathcal{U}_2[0,T]$, let $x(\cdot)$ be the corresponding state process. Applying Itô's formula to $\langle\, P(\cdot)x(\cdot), x(\cdot)\,\rangle$ and $\langle\, \varphi(\cdot), x(\cdot)\,\rangle$, with some computation, we have

$$\begin{aligned} &J_1(\xi; u_1(\cdot), u_2(\cdot)) - \langle\, P(0)\xi, \xi\,\rangle - 2\langle\, \varphi(0), \xi\,\rangle \\ &= E\int_0^T \Big\{|\widehat{R}_1^{\frac{1}{2}}[u_1 + \widehat{R}_1^{-1}\widehat{S}_1x + \widehat{R}_1^{-1}f]|^2 - |\widehat{R}_1^{-\frac{1}{2}}f|^2 \\ &\qquad\qquad + 2\langle\, B_2^T\varphi + D_2^T\theta, u_2\,\rangle + \langle\, D_2^TPD_2u_2, u_2\,\rangle\Big\}\,dt. \end{aligned} \tag{2.10}$$

Note that $P(\cdot)$ does not depend on $(u_1(\cdot), u_2(\cdot))$ and $(\varphi(\cdot), \theta(\cdot))$ does not depend on $u_1(\cdot)$ (although it might depend on $u_2(\cdot)$). By (1.8), $f(\cdot)$ does not depend on $u_1(\cdot)$. Consequently, we obtain

$$\begin{aligned} &J_1(\xi; u_1(\cdot), u_2(\cdot)) - J_1(\xi; \overline{u}_1(\cdot), u_2(\cdot)) \\ &= E\int_0^T \big|\widehat{R}_1^{\frac{1}{2}}[u_1 + \widehat{R}_1^{-1}\widehat{S}_1x + \widehat{S}_1f]\big|^2dt \ge 0, \end{aligned} \tag{2.11}$$

which implies that $\overline{u}_1(\cdot)$ is an optimal control. It also leads to (2.8). ∎

3. LQ problem for the leader. Now, let Problem $(\mathrm{LQ})_1$ be uniquely solvable for any given $(\xi, u_2(\cdot)) \in \mathbb{R}^n \times \mathcal{U}_2[0,T]$. Then the follower takes his optimal control $\overline{u}_1(\cdot)$ of form (1.7). Consequently, the leader has the state equation (1.9) with the coefficients given by (1.10). Note that $\widehat{F}_1$ and $\widehat{D}_1$ are symmetric. As we mentioned before, (1.9) is an FBSDE whose adapted solution is a triple $(x(\cdot), \varphi(\cdot), \theta(\cdot))$ of $\{\mathcal{F}_t\}_{t\ge0}$-adapted processes. The leader would like to choose his control so that his cost functional (with $u_1(\cdot) = \overline{u}_1(\cdot)$ being of the form (1.7)) is minimized. To be more precise, we define

$$\begin{aligned} &\widehat{J}_2(\xi; u_2(\cdot)) \triangleq J_2(\xi; \overline{u}_1(\cdot), u_2(\cdot)) \\ &= E\left\{\int_0^T \big[\,\langle\, Q_2x(t), x(t)\,\rangle + \langle\, R_2u_2(t), u_2(t)\,\rangle\,\big]dt + \langle\, G_2x(T), x(T)\,\rangle\right\}, \end{aligned} \tag{3.1}$$

where $x(\cdot)$ is the first component of the adapted solution $(x(\cdot), \varphi(\cdot), \theta(\cdot))$ of (1.9). The LQ problem for the leader can be stated as follows.

Problem $(\mathrm{LQ})_2$. For given $\xi \in \mathbb{R}^n$, find a $\overline{u}_2(\cdot) \in \mathcal{U}_2[0,T]$ such that

$$\widehat{J}_2(\xi, \overline{u}_2(\cdot)) = \min_{u_2(\cdot)\in\mathcal{U}_2[0,T]} \widehat{J}_2(\xi; u_2(\cdot)). \tag{3.2}$$

The above Problem $(\mathrm{LQ})_2$ is an LQ problem for an FBSDE. Similarly to the previous section, we can introduce the following notion.

Definition 3.1. *Problem* $(\mathrm{LQ})_2$ *is said to be*

(i) *finite at* $\xi \in \mathbb{R}^n$ *if*

$$\inf_{u_2(\cdot)\in\mathcal{U}_2[0,T]} \widehat{J}_2(\xi; u_2(\cdot)) > -\infty; \tag{3.3}$$

(ii) *(uniquely) solvable at* $\xi \in \mathbb{R}^n$ *if there exists a (unique)* $\overline{u}_2(\cdot) \in \mathcal{U}_2[0,T]$ *such that* (3.2) *holds.*

Any $\overline{u}_2(\cdot) \in \mathcal{U}_2[0,T]$ *satisfying* (3.2) *is called an optimal control, and the corresponding state* $(\overline{x}(\cdot), \overline{\varphi}(\cdot), \overline{\theta}(\cdot))$ *is called an optimal state process, and* $(\overline{x}(\cdot), \overline{\varphi}(\cdot), \overline{\theta}(\cdot), \overline{u}_2(\cdot))$ *is called an optimal* 4*-tuple, for the Problem* $(\mathrm{LQ})_2$*, respectively.*

Similarly to Proposition 2.2, we have the following Pontryagin-type maximum principle for Problem $(\mathrm{LQ})_2$.

THEOREM 3.2. *Let* (S) *hold.*

(i) *Let Problem* $(\mathrm{LQ})_2$ *be finite at some* $\xi \in \mathbb{R}^n$*. Then for any* $u_2(\cdot) \in \mathcal{U}_2[0,T]$*, the unique adapted solution* $(x^0(\cdot), \varphi^0(\cdot), \theta^0(\cdot), y^0(\cdot), z^0(\cdot), \psi^0(\cdot))$ *of the FBSDE*

$$\begin{cases} dx^0(t) = \big[\widehat{A}x^0(t) + \widehat{F}_1\varphi^0(t) + \widehat{B}_1\theta^0(t) + \widehat{B}_2u_2(t)\big]dt, \\ \qquad\qquad + \big[\widehat{C}x^0(t) + \widehat{B}_1^T\varphi^0(t) + \widehat{D}_1\theta^0(t) + \widehat{D}_2u_2(t)\big]dW(t), \\ d\varphi^0(t) = -\big[\widehat{A}^T\varphi^0(t) + \widehat{C}^T\theta^0(t) + \widehat{F}_2^Tu_2(t)\big]dt + \theta^0(t)dW(t), \\ dy^0(t) = -\big[\widehat{A}^Ty^0(t) + \widehat{C}^Tz(t) + Q_2x^0(t)\big]dt + z^0(t)dW(t), \\ d\psi^0(t) = \big[\widehat{A}\psi^0(t) + \widehat{F}_1y^0(t) + \widehat{B}_1z^0(t)\big]dt \\ \qquad\qquad + \big[\widehat{C}\psi^0(t) + \widehat{B}_1^Ty^0(t) + \widehat{D}_1z^0(t)\big]dW(t), \\ x^0(0) = 0, \qquad \varphi(T) = 0, \qquad y^0(T) = G_2x^0(T), \qquad \psi^0(0) = 0 \end{cases} \tag{3.4}$$

satisfies

$$\widehat{J}_2(0; u_2(\cdot)) = E\int_0^T \langle\, R_2u_2(t) + \widehat{B}_2^Ty^0(t) + \widehat{D}_2^Tz^0(t) + \widehat{F}_2\psi^0(t), u_2(t)\,\rangle\, dt \geq 0. \tag{3.5}$$

(ii) *Let the conclusion of* (i) *hold. Then Problem* $(\mathrm{LQ})_2$ *is solvable at* $\xi \in \mathbb{R}^n$ *with* $(\overline{x}(\cdot), \overline{\varphi}(\cdot), \overline{\theta}(\cdot), \overline{u}_2(\cdot))$ *being an optimal* 4*-tuple if and only if* $(\overline{x}(\cdot), \overline{\varphi}(\cdot), \overline{\theta}(\cdot))$ *is the adapted solution of* (1.9) *corresponding to* $(\xi, \overline{u}_2(\cdot))$ *and the FBSDE*

$$\begin{cases} d\overline{y} = -(\widehat{A}^T\overline{y} + \widehat{C}^T\overline{z} + Q_2\overline{x})dt + \overline{z}dW(t), \\ d\overline{\psi} = (\widehat{A}\overline{\psi} + \widehat{F}_1\overline{y} + \widehat{B}_1\overline{z})dt + (\widehat{C}\overline{\psi} + \widehat{B}_1^T\overline{y} + \widehat{D}_1\overline{z})dW(t), \\ \overline{y}(T) = G_2\overline{x}(T), \qquad \overline{\psi}(0) = 0 \end{cases} \tag{3.6}$$

admits a unique adapted solution $(\overline{y}(\cdot), \overline{z}(\cdot), \overline{\psi}(\cdot))$ *such that*

$$R_2\overline{u}_2 + \widehat{B}_2^T\overline{y} + \widehat{D}_2^T\overline{z} + \widehat{F}_2\overline{\psi} = 0. \tag{3.7}$$

Proof. For any fixed $\xi \in \mathbb{R}^n$ and $\overline{u}_2(\cdot) \in \mathcal{U}_2[0,T]$, FBSDE (1.9) admits a unique adapted solution $(\overline{x}(\cdot), \overline{\varphi}(\cdot), \overline{\theta}(\cdot))$, and FBSDE (3.6) also admits a unique adapted solution $(\overline{y}(\cdot), \overline{z}(\cdot), \overline{\psi}(\cdot))$. (One can solve the BSDE for $(\overline{y}(\cdot), \overline{z}(\cdot))$ first, then solve the FSDE for $\overline{\psi}(\cdot)$.) Similarly, for any $u_2(\cdot) \in \mathcal{U}_2[0,T]$, (3.6) admits a unique adapted

solution $(x^0(\cdot),\varphi^0(\cdot),\theta^0(\cdot),y^0(\cdot),z^0(\cdot),\psi^0(\cdot))$. By Itô's formula, we obtain

$$\begin{aligned}
&E\langle G_2\overline{x}(T),x^0(T)\rangle\\
&=E\big[\langle \overline{y}(T),x^0(T)\rangle-\langle y(0),x^0(0)\rangle-\langle \overline{\psi}(T),\varphi^0(T)\rangle+\langle \overline{\psi}(0),\varphi^0(0)\rangle\big]\\
&=E\int_0^T\big[-\langle \widehat{A}^T\overline{y}+\widehat{C}^T\overline{z}+Q_2\overline{x},x^0\rangle+\langle \overline{y},\widehat{A}x^0+\widehat{F}_1\varphi^0+\widehat{B}_1\theta^0+\widehat{B}_2u_2\rangle\\
&\qquad+\langle \overline{z},\widehat{C}x^0+\widehat{B}_1^T\varphi^0+\widehat{D}_1\theta^0+\widehat{D}_2u_2\rangle-\langle \widehat{A\psi}+\widehat{F}_1\overline{y}+\widehat{B}_1\overline{z},\varphi^0\rangle\\
&\qquad+\langle \overline{\psi},\widehat{A}^T\varphi^0+\widehat{C}^T\theta^0+\widehat{F}_2^Tu_2\rangle-\langle \widehat{C\psi}+\widehat{B}_1^T\overline{y}+\widehat{D}_1\overline{z},\theta^0\rangle\big]dt\\
&=E\int_0^T\big[-\langle Q_2\overline{x},x^0\rangle+\langle u_2,\widehat{B}_2^T\overline{y}+\widehat{D}_2^T\overline{z}+\widehat{F}_2\overline{\psi}\rangle\big]dt.
\end{aligned}\tag{3.8}$$

Similarly, we have

$$E\langle G_2x^0(T),x^0(T)\rangle=E\int_0^T\big[-\langle Q_2x^0,x^0\rangle+\langle u_2,\widehat{B}_2^Ty^0+\widehat{D}_2^Tz^0+\widehat{F}_2\psi^0\rangle\big]dt,\tag{3.9}$$

which implies

$$\widehat{J}_2(0;u_2(\cdot))=E\int_0^T\langle R_2u_2(t)+\widehat{B}_2^Ty^0(t)+\widehat{D}_2^Tz^0(t)+\widehat{F}_2\psi^0(t),u_2(t)\rangle\,dt.\tag{3.10}$$

Then, for any $\lambda\in\mathbb{R}$,

$$\begin{aligned}
&\widehat{J}_2(\xi;\overline{u}_2(\cdot)+\lambda u_2(\cdot))-\widehat{J}_2(\xi;\overline{u}_2(\cdot))\\
&=2\lambda E\left\{\int_0^T\big[\langle Q_2\overline{x},x^0\rangle+\langle R_2\overline{u}_2,u_2\rangle\big]dt+\langle G_2\overline{x}(T),x^0(T)\rangle\right\}\\
&\qquad+\lambda^2E\left\{\int_0^T\big[\langle Q_2x^0,x^0\rangle+\langle R_2u_2,u_2\rangle\big]dt+\langle G_2x^0(T),x^0(T)\rangle\right\}\\
&=2\lambda E\int_0^T\langle u_2,R_2\overline{u}_2+\widehat{B}_2^T\overline{y}+\widehat{D}_2^T\overline{z}+\widehat{F}_2\overline{\psi}\rangle\,dt\\
&\qquad+\lambda^2E\int_0^T\langle u_2,R_2u_2+\widehat{B}_2^Ty^0+\widehat{D}_2^Tz^0+\widehat{F}_2\psi^0\rangle\,dt.
\end{aligned}\tag{3.11}$$

Hence, when Problem $(\mathrm{LQ})_2$ is finite at some $\xi\in\mathbb{R}^n$, (3.5) must hold. Further, if (3.5) holds, then $\overline{u}_2(\cdot)$ is optimal if and only if (3.7) holds. □

We note that when

$$Q_2(t)\geq 0,\quad R_2(t)\geq 0,\quad G_2\geq 0,\qquad t\in[0,T],\ \text{a.s.}\tag{3.12}$$

holds, (3.5) holds automatically.

The above result gives us an equivalence between the solvability of Problem $(\mathrm{LQ})_2$ and that of an FBSDE. As usual, we refer to (3.6) as the adjoint equation of (1.9) along the optimal 4-tuple. Condition (3.7) together with (3.5) can be regarded as the maximum condition in Pontryagin's maximum principle. In the current case, due to the linear quadratic nature of the problem, (3.5) implies that the functional $u_2(\cdot)\mapsto\widehat{J}_2(\xi;u_2(\cdot))$ is convex. Thus, (3.7) becomes a sufficient condition for the existence of an optimal control. We refer to (1.9), (3.6)–(3.7) as the *optimality system*

of Problem $(\mathrm{LQ})_2$. It is clear that (1.9) and (3.6) are two (decoupled) FBSDEs which are coupled through condition (3.7).

Similar to typical (stochastic) LQ problems, the representation of optimal control $\overline{u}_2(\cdot)$ through (3.7) is not satisfactory since in order to determine $(\overline{y}(t),\overline{z}(t))$ at time $t\in[0,T)$, one needs to solve the BSDE in (3.6), for which $\overline{x}(T)$ (a future information of the state) has to be known. This is not realistic. We expect to have some kind of state feedback representation for the optimal control via a certain Riccati equation. To make the problem clearer, let us put (1.9) and (3.6) together, dropping bars in $\overline{x}$, etc. (but keeping the bar in $\overline{u}_2$), for notational simplicity:

$$
\begin{cases}
dx(t)=\big[\widehat{A}x(t)+\widehat{F}_1\varphi(t)+\widehat{B}_1\theta(t)+\widehat{B}_2\overline{u}_2(t)\big]dt,\\
\qquad\qquad +\big[\widehat{C}x(t)+\widehat{B}_1^T\varphi(t)+\widehat{D}_1\theta(t)+\widehat{D}_2\overline{u}_2(t)\big]dW(t),\\
d\varphi(t)=-\big[\widehat{A}^T\varphi(t)+\widehat{C}^T\theta(t)+\widehat{F}_2^T\overline{u}_2(t)\big]dt+\theta(t)dW(t),\\
dy(t)=-\big[\widehat{A}^Ty(t)+\widehat{C}^Tz(t)+Q_2x(t)\big]dt+z(t)dW(t),\\
d\psi(t)=\big[\widehat{A}\psi(t)+\widehat{F}_1y(t)+\widehat{B}_1z(t)\big]dt+\big[\widehat{C}\psi(t)+\widehat{B}_1^Ty(t)+\widehat{D}_1z(t)\big]dW(t),\\
x(0)=\xi,\qquad \varphi(T)=0,\qquad y(T)=G_2x(T),\qquad \psi(0)=0,\\
R_2\overline{u}_2+\widehat{B}_2^Ty+\widehat{D}_2^Tz+\widehat{F}_2\psi=0.
\end{cases}
\tag{3.13}
$$

Note that the equations for $(x(\cdot),\varphi(\cdot))$ form a coupled FBSDE, and those for $(y(\cdot),\psi(\cdot))$ form another coupled FBSDE. These two FBSDEs are further coupled through the last relation. Hence, the above is a coupled system of FBSDEs. We may look at the above in a different way. To this end, let us set

$$
\begin{cases}
X=\begin{pmatrix}x\\ \psi\end{pmatrix},\quad Y=\begin{pmatrix}y\\ \varphi\end{pmatrix},\quad Z=\begin{pmatrix}z\\ \theta\end{pmatrix},\quad X_0\triangleq\begin{pmatrix}\xi\\ 0\end{pmatrix},\\
\widehat{\mathcal{A}}\triangleq\begin{pmatrix}\widehat{A}&0\\ 0&\widehat{A}\end{pmatrix},\quad \widehat{\mathcal{F}}_1=\begin{pmatrix}0&\widehat{F}_1\\ \widehat{F}_1&0\end{pmatrix},\quad \widehat{\mathcal{B}}_1=\begin{pmatrix}0&\widehat{B}_1\\ \widehat{B}_1&0\end{pmatrix},\quad \widehat{\mathcal{B}}_2=\begin{pmatrix}\widehat{B}_2\\ 0\end{pmatrix},\\
\widehat{\mathcal{C}}\triangleq\begin{pmatrix}\widehat{C}&0\\ 0&\widehat{C}\end{pmatrix},\quad \widehat{\mathcal{D}}_1=\begin{pmatrix}0&\widehat{D}_1\\ \widehat{D}_1&0\end{pmatrix},\quad \widehat{\mathcal{D}}_2=\begin{pmatrix}\widehat{D}_2\\ 0\end{pmatrix},\\
\widehat{\mathcal{Q}}_2\triangleq\begin{pmatrix}Q_2&0\\ 0&0\end{pmatrix},\quad \widehat{\mathcal{F}}_2=\begin{pmatrix}0&\widehat{F}_2\end{pmatrix},\quad \widehat{\mathcal{G}}_2=\begin{pmatrix}G_2&0\\ 0&0\end{pmatrix}.
\end{cases}
\tag{3.14}
$$

Then (3.13) is equivalent to the FBSDE

$$
\begin{cases}
dX=\big(\widehat{\mathcal{A}}X+\widehat{\mathcal{F}}_1Y+\widehat{\mathcal{B}}_1Z+\widehat{\mathcal{B}}_2\overline{u}_2\big)dt\\
\qquad\quad +\big(\widehat{\mathcal{C}}X+\widehat{\mathcal{B}}_1^TY+\widehat{\mathcal{D}}_1Z+\widehat{\mathcal{D}}_2\overline{u}_2\big)dW(t),\\
dY=-\big(\widehat{\mathcal{Q}}_2X+\widehat{\mathcal{A}}^TY+\widehat{\mathcal{C}}^TZ+\widehat{\mathcal{F}}_2^T\overline{u}_2\big)dt+ZdW(t),\\
X(0)=X_0,\qquad Y(T)=\widehat{\mathcal{G}}_2X(T),
\end{cases}
\tag{3.15}
$$

together with the following condition:

$$
R_2\overline{u}_2+\widehat{\mathcal{B}}_2^TY+\widehat{\mathcal{F}}_2X+\widehat{\mathcal{D}}_2^TZ=0. \tag{3.16}
$$

FBSDE (3.15) is of a standard form. We should point out that only $\widehat{\mathcal{F}}_1$, $\widehat{\mathcal{D}}_1$, $\widehat{\mathcal{Q}}_2$, and $\widehat{\mathcal{G}}_2$ are symmetric; $\widehat{\mathcal{A}}$, $\widehat{\mathcal{C}}$, and $\widehat{\mathcal{B}}_1$ are not necessarily symmetric and/or skew-symmetric (although they look like so). Thus, the seemingly special form (3.14) of the coefficients

does not give us any additional helps in solving the linear FBSDE (3.15). We now use the idea of the four-step scheme (see [24, 32, 33]) to study the solvability of the above FBSDE. Suppose we have the relation

$$Y(t) = \widehat{P}(t)X(t), \qquad t \in [0,T], \tag{3.17}$$

with $\widehat{P}(\cdot)$ being an $\mathcal{S}^{2n}$-valued process satisfying

$$\begin{cases} d\widehat{P}(t) = \widehat{\Gamma}(t)dt + \widehat{\Lambda}(t)dW(t), & t \in [0,T], \\ \widehat{P}(T) = \widehat{\mathcal{G}}_2 \end{cases} \tag{3.18}$$

for some undetermined $\mathcal{S}^{2n}$-valued processes $\widehat{\Gamma}(\cdot)$ and $\widehat{\Lambda}(\cdot)$. Applying Itô's formula to (3.17), we obtain (suppressing t below)

$$\begin{aligned} &\big(\widehat{\Gamma}X + \widehat{P}\widehat{\mathcal{A}}X + \widehat{P}\widehat{\mathcal{F}}_1\widehat{P}X + \widehat{P}\widehat{\mathcal{B}}_1 Z + \widehat{P}\widehat{\mathcal{B}}_2\overline{u}_2 \\ &\qquad + \widehat{\Lambda}\widehat{\mathcal{C}}X + \widehat{\Lambda}\widehat{\mathcal{B}}_1^T\widehat{P}X + \Lambda\widehat{\mathcal{D}}_1 Z + \widehat{\Lambda}\widehat{\mathcal{D}}_2\overline{u}_2\big)dt \\ &\qquad + \big(\widehat{\Lambda}X + \widehat{P}\widehat{\mathcal{C}}X + \widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}X + \widehat{P}\widehat{\mathcal{D}}_1 Z + \widehat{P}\widehat{\mathcal{D}}_2\overline{u}_2\big)dW(t) \\ &= d[\widehat{P}X] = dY \\ &= -\big(\widehat{\mathcal{Q}}_2 X + \widehat{\mathcal{A}}^T\widehat{P}X + \widehat{\mathcal{C}}^T Z + \widehat{\mathcal{F}}_2^T\overline{u}_2\big)dt + ZdW(t). \end{aligned} \tag{3.19}$$

Hence, comparing the diffusion terms, we have

$$\big(I - \widehat{P}\widehat{\mathcal{D}}_1\big)Z = \big(\widehat{\Lambda} + \widehat{P}\widehat{\mathcal{C}} + \widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}\big)X + \widehat{P}\widehat{\mathcal{D}}_2\overline{u}_2, \tag{3.20}$$

and by comparing the drift terms in (3.19), we have

$$\begin{aligned} 0 &= \widehat{\Gamma}X + \widehat{P}\widehat{\mathcal{A}}X + \widehat{P}\widehat{\mathcal{F}}_1\widehat{P}X + \widehat{P}\widehat{\mathcal{B}}_1 Z + \widehat{P}\widehat{\mathcal{B}}_2\overline{u}_2 \\ &\quad + \widehat{\Lambda}\widehat{\mathcal{C}}X + \widehat{\Lambda}\widehat{\mathcal{B}}_1^T\widehat{P}X + \widehat{\Lambda}\widehat{\mathcal{D}}_1 Z + \widehat{\Lambda}\widehat{\mathcal{D}}_2\overline{u}_2 \\ &\quad + \widehat{\mathcal{Q}}_2 X + \widehat{\mathcal{A}}^T\widehat{P}X + \widehat{\mathcal{C}}^T Z + \widehat{\mathcal{F}}_2^T\overline{u}_2 \\ &= (\widehat{\Gamma} + \widehat{P}\widehat{\mathcal{A}} + \widehat{\mathcal{A}}^T\widehat{P} + \widehat{P}\widehat{\mathcal{F}}_1\widehat{P} + \widehat{\mathcal{Q}}_2 + \widehat{\Lambda}\widehat{\mathcal{C}} + \widehat{\Lambda}\widehat{\mathcal{B}}_1^T\widehat{P})X \\ &\quad + (\widehat{P}\widehat{\mathcal{B}}_1 + \widehat{\Lambda}\widehat{\mathcal{D}}_1 + \widehat{\mathcal{C}}^T)Z + (\widehat{P}\widehat{\mathcal{B}}_2 + \widehat{\Lambda}\widehat{\mathcal{D}}_2 + \widehat{\mathcal{F}}_2^T)\overline{u}_2. \end{aligned} \tag{3.21}$$

We should keep in mind that (3.19) is equivalent to (3.20)–(3.21). Next, assuming the existence of $\big(I - \widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}$, one obtains (from (3.20))

$$Z = \big(I - \widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big[\big(\widehat{\Lambda} + \widehat{P}\widehat{\mathcal{C}} + \widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}\big)X + \widehat{P}\widehat{\mathcal{D}}_2\overline{u}_2\big]. \tag{3.22}$$

Hence, it follows from (3.16) that

$$\begin{aligned} 0 &= R_2\overline{u}_2 + \widehat{\mathcal{B}}_2^T Y + \widehat{\mathcal{F}}_2 X + \widehat{\mathcal{D}}_2^T Z \\ &= \big[R_2 + \widehat{\mathcal{D}}_2^T\big(I - \widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\widehat{P}\widehat{\mathcal{D}}_2\big]\overline{u}_2 \\ &\quad + \big[\widehat{\mathcal{D}}_2^T\big(I - \widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big(\widehat{\Lambda} + \widehat{P}\widehat{\mathcal{C}} + \widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}\big) + \widehat{\mathcal{B}}_2^T\widehat{P} + \widehat{\mathcal{F}}_2\big]X. \end{aligned} \tag{3.23}$$

We point out that

$$\big(I - \widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\widehat{P} = \widehat{P}\big(I - \widehat{\mathcal{D}}_1\widehat{P}\big)^{-1} \tag{3.24}$$

is symmetric (which can be proved by multiplying both sides by $(I-\widehat{P}\widehat{\mathcal{D}}_1)$ from left and by $(I-\widehat{\mathcal{D}}_1\widehat{P})$ from right). Let us now assume that

$$\widehat{R}_2 \triangleq R_2+\widehat{\mathcal{D}}_2^T\big(I-\widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\widehat{P}\widehat{\mathcal{D}}_2 \tag{3.25}$$

is invertible (which is an $\mathcal{S}^{m_2}$-valued process). Consequently, by (3.23),

$$\overline{u}_2=-\widehat{R}_2^{-1}\big[\widehat{\mathcal{D}}_2^T\big(I-\widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big(\widehat{\Lambda}+\widehat{P}\widehat{\mathcal{C}}+\widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}\big)+\widehat{\mathcal{B}}_2^T\widehat{P}+\widehat{\mathcal{F}}_2\big]X. \tag{3.26}$$

Plugging the above into (3.22), one has

$$\begin{aligned} Z=&\big(I-\widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big\{\widehat{\Lambda}+\widehat{P}\widehat{\mathcal{C}}+\widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}\\ &-\widehat{P}\widehat{\mathcal{D}}_2\widehat{R}_2^{-1}\big[\widehat{\mathcal{D}}_2^T\big(I-\widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big(\widehat{\Lambda}+\widehat{P}\widehat{\mathcal{C}}+\widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}\big)+\widehat{\mathcal{B}}_2^T\widehat{P}+\widehat{\mathcal{F}}_2\big]\big\}X. \end{aligned} \tag{3.27}$$

Substituting (3.26)–(3.27) into (3.21), we end up with

$$\begin{aligned} 0=&(\widehat{\Gamma}+\widehat{P}\widehat{\mathcal{A}}+\widehat{\mathcal{A}}^T\widehat{P}+\widehat{P}\widehat{\mathcal{F}}_1\widehat{P}+\widehat{\mathcal{Q}}_2+\widehat{\Lambda}\widehat{\mathcal{C}}+\widehat{\Lambda}\widehat{\mathcal{B}}_1^T\widehat{P})X\\ &+(\widehat{P}\widehat{\mathcal{B}}_1+\widehat{\Lambda}\widehat{\mathcal{D}}_1+\widehat{\mathcal{C}}^T)\big(I-\widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big\{\widehat{\Lambda}+\widehat{P}\widehat{\mathcal{C}}+\widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}\\ &-\widehat{P}\widehat{\mathcal{D}}_2\widehat{R}_2^{-1}\big[\widehat{\mathcal{D}}_2^T\big(I-\widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big(\widehat{\Lambda}+\widehat{P}\widehat{\mathcal{C}}+\widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}\big)+\widehat{\mathcal{B}}_2^T\widehat{P}+\widehat{\mathcal{F}}_2\big]\big\}X\\ &-(\widehat{P}\widehat{\mathcal{B}}_2+\widehat{\Lambda}\widehat{\mathcal{D}}_2+\widehat{\mathcal{F}}_2^T)\widehat{R}_2^{-1}\big\{\widehat{\mathcal{D}}_2^T\big[I-\widehat{P}\widehat{\mathcal{D}}_1\big]^{-1}\big[\widehat{\Lambda}+\widehat{P}\widehat{\mathcal{C}}+\widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}\big]\\ &\qquad+\widehat{\mathcal{B}}_2^T\widehat{P}+\widehat{\mathcal{F}}_2\big\}X. \end{aligned} \tag{3.28}$$

Thus, process $\widehat{\Gamma}$ should be chosen as follows:

$$\begin{aligned} \widehat{\Gamma}=&-(\widehat{P}\widehat{\mathcal{A}}+\widehat{\mathcal{A}}^T\widehat{P}+\widehat{P}\widehat{\mathcal{F}}_1\widehat{P}+\widehat{\mathcal{Q}}_2+\widehat{\Lambda}\widehat{\mathcal{C}}+\widehat{\Lambda}\widehat{\mathcal{B}}_1^T\widehat{P})\\ &-(\widehat{P}\widehat{\mathcal{B}}_1+\widehat{\Lambda}\widehat{\mathcal{D}}_1+\widehat{\mathcal{C}}^T)\big(I-\widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big\{\widehat{\Lambda}+\widehat{P}\widehat{\mathcal{C}}+\widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}\\ &\qquad-\widehat{P}\widehat{\mathcal{D}}_2\widehat{R}_2^{-1}\big[\widehat{\mathcal{D}}_2^T\big(I-\widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big(\widehat{\Lambda}+\widehat{P}\widehat{\mathcal{C}}+\widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}\big)+\widehat{\mathcal{B}}_2^T\widehat{P}+\widehat{\mathcal{F}}_2\big]\big\}\\ &+(\widehat{P}\widehat{\mathcal{B}}_2+\widehat{\Lambda}\widehat{\mathcal{D}}_2+\widehat{\mathcal{F}}_2^T)\widehat{R}_2^{-1}[\widehat{\mathcal{D}}_2^T\big(I-\widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big(\widehat{\Lambda}+\widehat{P}\widehat{\mathcal{C}}+\widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}\big)+\widehat{\mathcal{B}}_2^T\widehat{P}+\widehat{\mathcal{F}}_2]. \end{aligned} \tag{3.29}$$

In other words, if $\widehat{\Gamma}(\cdot)$ is defined by (3.29), then (3.28) holds.

Next, by substituting (3.17), (3.26), and (3.27) into the equation for $X(\cdot)$ in (3.15), we obtain

$$\begin{cases} dX=\overline{\mathcal{A}}Xdt+\overline{\mathcal{C}}XdW(t),\\ X(0)=X_0, \end{cases} \tag{3.30}$$

where

$$\begin{cases} \overline{\mathcal{A}}=\widehat{\mathcal{A}}+\widehat{\mathcal{F}}_1\widehat{P}+\widehat{\mathcal{B}}_1\big(I-\widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big\{\widehat{\Lambda}+\widehat{P}\widehat{\mathcal{C}}+\widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}\\ \qquad-\widehat{P}\widehat{\mathcal{D}}_2\widehat{R}_2^{-1}\big[\widehat{\mathcal{D}}_2^T\big(I-\widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big(\widehat{\Lambda}+\widehat{P}\widehat{\mathcal{C}}+\widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}\big)+\widehat{\mathcal{B}}_2^T\widehat{P}+\widehat{\mathcal{F}}_2\big]\big\}\\ \qquad-\widehat{\mathcal{B}}_2\widehat{R}_2^{-1}\big[\widehat{\mathcal{D}}_2^T\big(I-\widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big(\widehat{\Lambda}+\widehat{P}\widehat{\mathcal{C}}+\widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}\big)+\widehat{\mathcal{B}}_2^T\widehat{P}+\widehat{\mathcal{F}}_2\big],\\ \overline{\mathcal{C}}=\widehat{\mathcal{C}}+\widehat{\mathcal{B}}_1^T\widehat{P}+\widehat{\mathcal{D}}_1\big(I-\widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big\{\widehat{\Lambda}+\widehat{P}\widehat{\mathcal{C}}+\widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}\\ \qquad-\widehat{P}\widehat{\mathcal{D}}_2\widehat{R}_2^{-1}\big[\widehat{\mathcal{D}}_2^T\big(I-\widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big(\widehat{\Lambda}+\widehat{P}\widehat{\mathcal{C}}+\widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}\big)+\widehat{\mathcal{B}}_2^T\widehat{P}+\widehat{\mathcal{F}}_2\big]\big\}\\ \qquad-\widehat{\mathcal{D}}_2\widehat{R}_2^{-1}\big[\widehat{\mathcal{D}}_2^T\big(I-\widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big(\widehat{\Lambda}+\widehat{P}\widehat{\mathcal{C}}+\widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}\big)+\widehat{\mathcal{B}}_2^T\widehat{P}+\widehat{\mathcal{F}}_2\big]. \end{cases} \tag{3.31}$$

Now, if $X(\cdot)$ is the unique strong solution of (3.30) and if $Y(\cdot)$, $Z(\cdot)$, and $u_2(\cdot)$ are defined by (3.16), (3.25), and (3.26), respectively (assuming that $(\widehat{P}(\cdot), \widehat{\Lambda}(\cdot))$ is an adapted solution of (3.18) and that $(I - \widehat{P}\widehat{\mathcal{D}}_1)^{-1}$ and $\widehat{R}_2^{-1}$ exist), then (3.15)–(3.16) hold.

We now make some rearrangement for the right-hand side of (3.29). First of all,

$$
\begin{aligned}
&(\widehat{P}\widehat{\mathcal{B}}_1 + \widehat{\Lambda}\widehat{\mathcal{D}}_1 + \widehat{\mathcal{C}}^T)\big(I - \widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big(\widehat{\Lambda} + \widehat{P}\widehat{\mathcal{C}} + \widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}\big)\\
&= (\widehat{P}\widehat{\mathcal{B}}_1 + \widehat{\mathcal{C}}^T)(I - \widehat{P}\widehat{\mathcal{D}}_1)^{-1}\widehat{P}(\widehat{\mathcal{B}}_1^T\widehat{P} + \widehat{\mathcal{C}}) + \widehat{\Lambda}\widehat{\mathcal{D}}_1(I - \widehat{P}\widehat{\mathcal{D}}_1)^{-1}\widehat{\Lambda}\\
&\qquad + (\widehat{P}\widehat{\mathcal{B}}_1 + \widehat{\mathcal{C}}^T)(I - \widehat{P}\widehat{\mathcal{D}}_1)^{-1}\widehat{\Lambda} + \widehat{\Lambda}\widehat{\mathcal{D}}_1(I - \widehat{P}\widehat{\mathcal{D}}_1)^{-1}\widehat{P}(\widehat{\mathcal{B}}_1^T\widehat{P} + \widehat{\mathcal{C}})\\
&= (\widehat{P}\widehat{\mathcal{B}}_1 + \widehat{\mathcal{C}}^T)(I - \widehat{P}\widehat{\mathcal{D}}_1)^{-1}\widehat{P}(\widehat{\mathcal{B}}_1^T\widehat{P} + \widehat{\mathcal{C}}) + \widehat{\Lambda}\widehat{\mathcal{D}}_1(I - \widehat{P}\widehat{\mathcal{D}}_1)^{-1}\widehat{\Lambda}\\
&\qquad + (\widehat{P}\widehat{\mathcal{B}}_1 + \widehat{\mathcal{C}}^T)(I - \widehat{P}\widehat{\mathcal{D}}_1)^{-1}\widehat{\Lambda} + \widehat{\Lambda}[(I - \widehat{P}\widehat{\mathcal{D}}_1)^{-1} - I](\widehat{\mathcal{B}}_1^T\widehat{P} + \widehat{\mathcal{C}}).
\end{aligned}
\tag{3.32}
$$

Next, one has

$$
\begin{aligned}
&(\widehat{P}\widehat{\mathcal{B}}_1 + \widehat{\Lambda}\widehat{\mathcal{D}}_1 + \widehat{\mathcal{C}}^T)\big(I - \widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\widehat{P}\widehat{\mathcal{D}}_2\widehat{R}_2^{-1}\big[\widehat{\mathcal{D}}_2^T\big(I - \widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big(\widehat{\Lambda} + \widehat{P}\widehat{\mathcal{C}} + \widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}\big)\\
&\qquad + \widehat{\mathcal{B}}_2^T\widehat{P} + \widehat{\mathcal{F}}_2\big]\\
&+ (\widehat{P}\widehat{\mathcal{B}}_2 + \widehat{\Lambda}\widehat{\mathcal{D}}_2 + \widehat{\mathcal{F}}_2^T)\widehat{R}_2^{-1}\big[\widehat{\mathcal{D}}_2^T\big(I - \widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big(\widehat{\Lambda} + \widehat{P}\widehat{\mathcal{C}} + \widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}\big) + \widehat{\mathcal{B}}_2^T\widehat{P} + \widehat{\mathcal{F}}_2\big]\\
&= \big[(\widehat{P}\widehat{\mathcal{B}}_1 + \widehat{\Lambda}\widehat{\mathcal{D}}_1 + \widehat{\mathcal{C}}^T)\big(I - \widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\widehat{P}\widehat{\mathcal{D}}_2 + (\widehat{P}\widehat{\mathcal{B}}_2 + \widehat{\Lambda}\widehat{\mathcal{D}}_2 + \widehat{\mathcal{F}}_2^T)\big]\widehat{R}_2^{-1}\\
&\quad\cdot\big[\widehat{\mathcal{D}}_2^T\big(I - \widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big(\widehat{\Lambda} + \widehat{P}\widehat{\mathcal{C}} + \widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}\big) + \widehat{\mathcal{B}}_2^T\widehat{P} + \widehat{\mathcal{F}}_2\big].
\end{aligned}
\tag{3.33}
$$

Note that

$$
\begin{aligned}
&(\widehat{P}\widehat{\mathcal{B}}_1 + \widehat{\Lambda}\widehat{\mathcal{D}}_1 + \widehat{\mathcal{C}}^T)\big(I - \widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\widehat{P}\widehat{\mathcal{D}}_2 + \widehat{P}\widehat{\mathcal{B}}_2 + \widehat{\Lambda}\widehat{\mathcal{D}}_2 + \widehat{\mathcal{F}}_2^T\\
&= (\widehat{P}\widehat{\mathcal{B}}_1 + \widehat{\Lambda}\widehat{\mathcal{D}}_1 + \widehat{\mathcal{C}}^T)\widehat{P}\big(I - \widehat{\mathcal{D}}_1\widehat{P}\big)^{-1}\widehat{\mathcal{D}}_2 + \widehat{\Lambda}\widehat{\mathcal{D}}_2 + \widehat{P}\widehat{\mathcal{B}}_2 + \widehat{\mathcal{F}}_2^T\\
&= (\widehat{P}\widehat{\mathcal{B}}_1\widehat{P} + \widehat{\mathcal{C}}^T\widehat{P})(I - \widehat{\mathcal{D}}_1\widehat{P})^{-1}\widehat{\mathcal{D}}_2 + \widehat{\Lambda}\big[\big(I - \widehat{\mathcal{D}}_1\widehat{P}\big)^{-1} - I\big]\widehat{\mathcal{D}}_2\\
&\qquad + \widehat{\Lambda}\widehat{\mathcal{D}}_2 + \widehat{P}\widehat{\mathcal{B}}_2 + \widehat{\mathcal{F}}_2^T\\
&= (\widehat{P}\widehat{\mathcal{B}}_1\widehat{P} + \widehat{\mathcal{C}}^T\widehat{P} + \widehat{\Lambda})\big(I - \widehat{\mathcal{D}}_1\widehat{P}\big)^{-1}\widehat{\mathcal{D}}_2 + \widehat{P}\widehat{\mathcal{B}}_2 + \widehat{\mathcal{F}}_2^T.
\end{aligned}
\tag{3.34}
$$

Hence, the right-hand side of (3.34) reads

$$
\begin{aligned}
&\big[(\widehat{P}\widehat{\mathcal{B}}_1\widehat{P} + \widehat{\mathcal{C}}^T\widehat{P} + \widehat{\Lambda})\big(I - \widehat{\mathcal{D}}_1\widehat{P}\big)^{-1}\widehat{\mathcal{D}}_2 + \widehat{P}\widehat{\mathcal{B}}_2 + \widehat{\mathcal{F}}_2^T\big]\widehat{R}_2^{-1}\\
&\quad\cdot\big[\widehat{\mathcal{D}}_2^T\big(I - \widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big(\widehat{P}\widehat{\mathcal{C}} + \widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P} + \widehat{\Lambda}\big) + \widehat{\mathcal{B}}_2^T\widehat{P} + \widehat{\mathcal{F}}_2\big].
\end{aligned}
\tag{3.35}
$$

Combining (3.31)–(3.35), we obtain

$$
\begin{aligned}
-\widehat{\Gamma} \equiv\ & -\widehat{\Gamma}(t, \widehat{P}, \widehat{\Lambda}) \triangleq \widehat{P}\widehat{\mathcal{A}} + \widehat{\mathcal{A}}^T\widehat{P} + \widehat{P}\widehat{\mathcal{F}}_1\widehat{P} + \widehat{\mathcal{Q}}_2\\
&+ (\widehat{P}\widehat{\mathcal{B}}_1 + \widehat{\mathcal{C}}^T)\big(I - \widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\widehat{P}\big(\widehat{\mathcal{C}} + \widehat{\mathcal{B}}_1^T\widehat{P}\big) + \widehat{\Lambda}\widehat{\mathcal{D}}_1\big(I - \widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\widehat{\Lambda}\\
&+ \big(\widehat{P}\widehat{\mathcal{B}}_1 + \widehat{\mathcal{C}}^T\big)\big(I - \widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\widehat{\Lambda} + \widehat{\Lambda}\big(I - \widehat{\mathcal{D}}_1\widehat{P}\big)^{-1}\big(\widehat{\mathcal{B}}_1^T\widehat{P} + \widehat{\mathcal{C}}\big)\\
&- \big[\big(\widehat{P}\widehat{\mathcal{B}}_1\widehat{P} + \widehat{\mathcal{C}}^T\widehat{P} + \widehat{\Lambda}\big)\big(I - \widehat{\mathcal{D}}_1\widehat{P}\big)^{-1}\widehat{\mathcal{D}}_2 + \widehat{P}\widehat{\mathcal{B}}_2 + \widehat{\mathcal{F}}_2^T\big]\widehat{R}_2^{-1}\\
&\quad\cdot\big[\widehat{\mathcal{D}}_2^T\big(I - \widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big(\widehat{P}\widehat{\mathcal{C}} + \widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P} + \widehat{\Lambda}\big) + \widehat{\mathcal{B}}_2^T\widehat{P} + \widehat{\mathcal{F}}_2\big]\\
=\ & \widehat{P}\widehat{\mathcal{A}} + \widehat{\mathcal{A}}^T\widehat{P} + \widehat{P}\widehat{\mathcal{F}}_1\widehat{P} + \widehat{\mathcal{Q}}_2\\
&+ \begin{pmatrix} \widehat{P}\widehat{\mathcal{B}}_1 + \widehat{\mathcal{C}}^T & \widehat{\Lambda} \end{pmatrix}\begin{pmatrix} (I - \widehat{P}\widehat{\mathcal{D}}_1)^{-1}\widehat{P} & (I - \widehat{P}\widehat{\mathcal{D}}_1)^{-1}\\ (I - \widehat{\mathcal{D}}_1\widehat{P})^{-1} & \widehat{\mathcal{D}}_1(I - \widehat{P}\widehat{\mathcal{D}}_1)^{-1}\end{pmatrix}\begin{pmatrix} \widehat{\mathcal{B}}_1^T\widehat{P} + \widehat{\mathcal{C}}\\ \widehat{\Lambda}\end{pmatrix}\\
&- \big[\big(\widehat{P}\widehat{\mathcal{B}}_1\widehat{P} + \widehat{\mathcal{C}}^T\widehat{P} + \widehat{\Lambda}\big)\big(I - \widehat{\mathcal{D}}_1\widehat{P}\big)^{-1}\widehat{\mathcal{D}}_2 + \widehat{P}\widehat{\mathcal{B}}_2 + \widehat{\mathcal{F}}_2^T\big]\widehat{R}_2^{-1}\\
&\quad\cdot\big[\widehat{\mathcal{D}}_2^T\big(I - \widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big(\widehat{P}\widehat{\mathcal{C}} + \widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P} + \widehat{\Lambda}\big) + \widehat{\mathcal{B}}_2^T\widehat{P} + \widehat{\mathcal{F}}_2\big].
\end{aligned}
\tag{3.36}
$$

Once we write the map $\widehat{\Gamma}$ in the above form, we have that for any $\widehat{P}, \widehat{\Lambda} \in \mathcal{S}^{2n}$, provided $(I - \widehat{P}\widehat{\mathcal{D}}_1)^{-1}$ and $\widehat{R}_2^{-1}$ exist, $\widehat{\Gamma}(t, \widehat{P}, \widehat{\Lambda}) \in \mathcal{S}^{2n}$, i.e., for any symmetric $\widehat{P}$ and $\widehat{\Lambda}$, image $\Gamma(t, \widehat{P}, \widehat{\Lambda})$ of $(\widehat{P}, \widehat{\Lambda})$ under the map $\widehat{\Gamma}$ is also symmetric. Here, we should note that similarly to (3.24),

$$\widehat{\mathcal{D}}_1\big(I - \widehat{P}\widehat{\mathcal{D}}_1\big)^{-1} = \big(I - \widehat{\mathcal{D}}_1\widehat{P}\big)^{-1}\widehat{\mathcal{D}}_1 \tag{3.37}$$

is symmetric and $\mathcal{S}^{2m_1}$-valued. Now, the Riccati equation for $(\widehat{P}(\cdot), \widehat{\Lambda}(\cdot))$ is given by

$$\begin{cases} d\widehat{P}(t) = \widehat{\Gamma}(t, \widehat{P}(t), \widehat{\Lambda}(t))dt + \widehat{\Lambda}(t)dW(t), \qquad t \in [0, T], \\ \widehat{P}(T) = \widehat{\mathcal{Q}}_2, \\ \det[I - \widehat{P}(t)\widehat{\mathcal{D}}_1(t)] \neq 0, \\ \det[R_2(t) + \widehat{\mathcal{D}}_2(t)^T\big(I - \widehat{P}(t)\widehat{\mathcal{D}}_1(t)\big)^{-1}\widehat{P}(t)\widehat{\mathcal{D}}_2(t)] \neq 0, \end{cases} \tag{3.38}$$

with $\widehat{\Gamma}(t, \widehat{P}, \widehat{\Lambda})$ given by (3.36). We summarize the above as follows.

THEOREM 3.3. *Let* (S) *hold. Let Riccati equation* (3.38) *admit an adapted solution* $(\widehat{P}(\cdot), \widehat{\Lambda}(\cdot))$. *Let* $X(\cdot)$ *be the solution of* (3.30). *Define* $(Y(\cdot), Z(\cdot), \overline{u}_2(\cdot))$ *by* (3.17), (3.26), *and* (3.27). *Then* (3.15)–(3.16) *holds. Moreover, for such a* $\overline{u}_2(\cdot)$, *the following holds:*

$$\widehat{J}_2(\xi; \overline{u}_2(\cdot)) = \langle\, \widehat{P}_2(0)\xi, \xi \,\rangle, \tag{3.39}$$

where $\widehat{P}(\cdot) = \begin{pmatrix} \widehat{P}_2(\cdot) & \widehat{P}_3(\cdot) \\ \widehat{P}_3(\cdot)^T & \widehat{P}_4(\cdot) \end{pmatrix}$. *When* (i) *of Theorem* 3.2 *holds, in particular, if* (3.12) *holds, the state feedback control* $\overline{u}_2(\cdot)$ *defined by* (3.26) *is an optimal control of Problem* $(\mathrm{LQ})_2$.

Proof. We need only prove (3.39). Similarly to (3.8)–(3.10), we are able to show that

$$\begin{aligned} \widehat{J}_2(\xi; u_2(\cdot)) &= \langle\, Y(0), X(0) \,\rangle + E\int_0^T \langle\, u_2, R_2u_2 + \widehat{\mathcal{B}}_2^TY + \widehat{\mathcal{D}}_2^TZ + \widehat{\mathcal{F}}_2X \,\rangle\, dt \\ &= \langle\, \widehat{P}_2(0)\xi, \xi \,\rangle. \end{aligned}$$

The rest is clear. □

Note that optimal control $\overline{u}_2(\cdot)$ has a "state" feedback representation (3.26) with the "state" $X(\cdot) \equiv \begin{pmatrix} x(\cdot) \\ \psi(\cdot) \end{pmatrix}$ being the solution of (3.30), or $X(\cdot)$ is the solution of the system

$$\begin{cases} dX = \big[\widetilde{\mathcal{A}}X + \widetilde{\mathcal{B}}_2\overline{u}_2\big]dt + \big[\widetilde{\mathcal{C}}X + \widetilde{\mathcal{D}}_2\overline{u}_2\big]dW(t), \\ X(0) = X_0 \end{cases} \tag{3.40}$$

corresponding to the feedback control $\overline{u}_2(\cdot)$ given by (3.26), where

$$\begin{cases} \widetilde{\mathcal{A}} = \widehat{\mathcal{A}} + \widehat{\mathcal{F}}_1\widehat{P} + \widehat{\mathcal{B}}_1(I - \widehat{P}\widehat{\mathcal{D}}_1)^{-1}(\widehat{\mathcal{A}} + \widehat{P}\widehat{\mathcal{C}} + \widehat{P}\widehat{\mathcal{B}}_1\widehat{P}), \\ \widetilde{C} = \widehat{\mathcal{C}} + \widehat{\mathcal{B}}_1^T\widehat{P} + \widehat{\mathcal{D}}_1(I - \widehat{P}\widehat{\mathcal{D}}_1)^{-1}(\widehat{\mathcal{A}} + \widehat{P}\widehat{\mathcal{C}} + \widehat{P}\widehat{\mathcal{B}}_1\widehat{P}), \\ \widetilde{\mathcal{B}}_2 = \widehat{\mathcal{B}}_2 + \widehat{\mathcal{B}}_1(I - \widehat{P}\widehat{\mathcal{D}}_1)^{-1}\widehat{P}\widehat{\mathcal{D}}_2, \quad \widetilde{\mathcal{D}}_2 = \widehat{\mathcal{D}}_2 + \widehat{\mathcal{D}}_1(I - \widehat{P}\widehat{\mathcal{D}}_1)^{-1}\widehat{P}\widehat{\mathcal{D}}_2. \end{cases} \tag{3.41}$$

The point here is that $\overline{u}_2(\cdot)$ given by (3.26) is nonanticipating. Likewise, for the follower, the optimal control $\overline{u}_1(\cdot)$ can also be represented in a nonanticipating way.

In fact, by (1.7)–(1.8), we have

$$
\begin{aligned}
\overline{u}_1 &= -\widehat{R}_1^{-1}\big\{\widehat{S}_1 x - \widehat{R}_1^{-1}\widehat{\mathcal{B}}_1^T Y + \widehat{\mathcal{D}}_1^T Z + D_1^T P D_2 u_2\big\} \\
&= -\widehat{R}_1^{-1}\Big\{ \begin{pmatrix} \widehat{S}_1 & 0 \end{pmatrix} - \widehat{R}_1^{-1}\widehat{\mathcal{B}}_1^T\widehat{P} + \widehat{\mathcal{D}}_1^T\big(I - \widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big\{\widehat{\Lambda} + \widehat{P}\widehat{\mathcal{C}} + \widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P} \\
&\quad - \widehat{P}\widehat{\mathcal{D}}_2\widehat{R}_2^{-1}\big[\widehat{\mathcal{D}}_2^T\big(I - \widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big(\widehat{\Lambda} + \widehat{P}\widehat{\mathcal{C}} + \widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}\big) + \widehat{\mathcal{B}}_2^T\widehat{P} + \widehat{\mathcal{F}}_2\big]\big\} \\
&\quad - D_1^T P D_2\widehat{R}_2^{-1}\big[\widehat{\mathcal{D}}_2^T\big(I - \widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big(\widehat{\Lambda} + \widehat{P}\widehat{\mathcal{C}} + \widehat{P}\widehat{\mathcal{B}}_1^T\widehat{P}\big) + \widehat{\mathcal{B}}_2^T\widehat{P} + \widehat{\mathcal{F}}_2\big]\Big\}X.
\end{aligned} \tag{3.42}
$$

Hence, under assumptions in Theorems 2.3 and 3.2, our differential game admits an open-loop solution $(\overline{u}_1(\cdot), \overline{u}_2(\cdot))$, and they admit a state feedback representation (3.26) and (3.42).

Remark 3.4. From (3.36), one can see that the equations for each block $\widehat{P}_2(\cdot)$, $\widehat{P}_3(\cdot)$, and $\widehat{P}_4(\cdot)$ of $\widehat{P}(\cdot)$ in (3.38) are heavily coupled (even for very simple cases; this is mainly due to the appearance of the controls in the diffusion and the randomness of the coefficients; see below). In particular, one might not be able to solve the equation for $\widehat{P}_2(\cdot)$ first, and then the remaining equations for $\widehat{P}_3(\cdot)$ and $\widehat{P}_4(\cdot)$. Although from (3.39), it seems that only $\widehat{P}_2(\cdot)$ plays a particular role, blocks $\widehat{P}_3(\cdot)$, $\widehat{P}_4(\cdot)$, and $\widehat{\Lambda}(\cdot)$ actually play equally important roles. We note that in (2.8), only $P(0)$ and $\varphi(0)$ appear. However, $\Lambda(\cdot)$ and $\theta(\cdot)$ play crucial roles as well. This is a similar situation. We expect that there might be some interesting relations among blocks $\widehat{P}_i(\cdot)$ $(i = 2, 3, 4)$, but we are currently not able to find them. We hope to say something about this in future papers.

We now look at a special case in which the follower does not appear. In this case the problem is reduced to a typical LQ problem. Let us still regard it as if the follower does appear but does not affect the game at all, i.e., we assume that

$$
B_1 = D_1 = 0, \quad Q_1 = 0, \quad G_1 = 0, \quad R_1 = I. \tag{3.43}
$$

In the above case, Riccati equation (1.5) admits a unique solution $(P(\cdot), \Lambda(\cdot)) = (0, 0)$. Consequently, by (1.10), one has

$$
\begin{cases}
\widehat{A} = A, \quad \widehat{F}_1 = 0, \quad \widehat{B}_1 = 0, \quad \widehat{B}_2 = B_2, \\
\widehat{C} = C, \quad \widehat{F}_2 = 0, \quad \widehat{D}_1 = 0, \quad \widehat{D}_2 = D_2.
\end{cases} \tag{3.44}
$$

The above further implies that (see (3.14))

$$
\begin{cases}
\widehat{\mathcal{A}} \triangleq \begin{pmatrix} A & 0 \\ 0 & A \end{pmatrix}, \quad \widehat{\mathcal{F}}_1 = 0, \quad \widehat{\mathcal{B}}_1 = 0, \quad \widehat{\mathcal{B}}_2 = \begin{pmatrix} B_2 \\ 0 \end{pmatrix}, \\
\widehat{\mathcal{C}} \triangleq \begin{pmatrix} C & 0 \\ 0 & C \end{pmatrix}, \quad \widehat{\mathcal{D}}_1 = 0, \quad \widehat{\mathcal{D}}_2 = \begin{pmatrix} D_2 \\ 0 \end{pmatrix}, \\
\widehat{\mathcal{Q}}_2 \triangleq \begin{pmatrix} Q_2 & 0 \\ 0 & 0 \end{pmatrix}, \quad \widehat{\mathcal{F}}_2 = 0, \quad \widehat{\mathcal{G}}_2 = \begin{pmatrix} G_2 & 0 \\ 0 & 0 \end{pmatrix}.
\end{cases} \tag{3.45}
$$

Then (3.36) becomes

$$
\begin{aligned}
-\widehat{\Gamma} \equiv -\widehat{\Gamma}(t, \widehat{P}, \widehat{\Lambda}) &\triangleq \widehat{P}\widehat{\mathcal{A}} + \widehat{\mathcal{A}}^T\widehat{P} + \widehat{\mathcal{Q}}_2 + \widehat{\mathcal{C}}^T\widehat{P}\widehat{\mathcal{C}} + \widehat{\mathcal{C}}^T\widehat{\Lambda} + \widehat{\Lambda}\widehat{\mathcal{C}} \\
&\quad - \big[\big(\widehat{\mathcal{C}}^T\widehat{P} + \widehat{\Lambda}\big)\widehat{\mathcal{D}}_2 + \widehat{P}\widehat{\mathcal{B}}_2\big]\widehat{R}_2^{-1}\big[\widehat{\mathcal{D}}_2^T\big(\widehat{P}\widehat{\mathcal{C}} + \widehat{\Lambda}\big) + \widehat{\mathcal{B}}_2^T\widehat{P}\big].
\end{aligned} \tag{3.46}
$$

If $(\widehat{P}(\cdot),\widehat{\Lambda}(\cdot))$ is an adapted solution of (3.38), and if we let

$$\widehat{P}(\cdot)=\begin{pmatrix}\widehat{P}_2(\cdot) & \widehat{P}_3(\cdot)\\ \widehat{P}_3(\cdot)^T & \widehat{P}_4(\cdot)\end{pmatrix},\quad \widehat{\Lambda}(\cdot)=\begin{pmatrix}\widehat{\Lambda}_2(\cdot) & \widehat{\Lambda}_3(\cdot)\\ \widehat{\Lambda}_3(\cdot)^T & \widehat{\Lambda}_4(\cdot)\end{pmatrix}, \tag{3.47}$$

then $(\widehat{P}_2(\cdot),\widehat{\Lambda}_2(\cdot))$ is an adapted solution of

$$\begin{cases} d\widehat{P}_2=-\big\{\widehat{P}_2A+A^T\widehat{P}_2+Q_2+C^T\widehat{P}_2C+C^T\widehat{\Lambda}_2+\widehat{\Lambda}_2C\\ \quad -[(C^T\widehat{P}_2+\widehat{\Lambda}_2)D_2+\widehat{P}_2B_2](R_2+D_2^T\widehat{P}_2D_2)^{-1}[D_2^T(\widehat{P}_2C+\widehat{\Lambda}_2)+B_2^T\widehat{P}]\big\}dt\\ \qquad\quad +\widehat{\Lambda}_2dW(t),\\ \widehat{P}_2(T)=Q_2. \end{cases} \tag{3.48}$$

This is the Riccati equation for the LQ problem (found in [6, 7, 8]) with the state equation (1.1) in which $B_1=D_1=0$ and with the cost functional (1.2) (with $i=2$). Thus, we might regard our differential game problem as a generalization of LQ problems.

To conclude this section, we would like to point out that the general solvability of Riccati equation (3.38) is very difficult. It remains a very challenging open question. We are not able to attack the general case in this paper. Instead, we would like to look at some special and interesting cases in section 5.

4. Two one-dimensional cases. In this section, we present two one-dimensional cases with some very special coefficients, in which we want to address two points: (i) Due to the randomness of the coefficients and the appearance of the controls in the diffusion, even for a very simple problem, the explicit solution might not be expected; and (ii) due to the appearance of the controls in the diffusion, the Isaacs-type condition does not necessarily hold, which results in that the "equilibrium" does not necessarily exist.

We now first assume the following:

$$\begin{cases} n=m_1=m_2=1,\quad A=C=B_1=B_2=Q_1=Q_2=0,\\ D_1=D_2=R_1=R_2=1,\quad G_1=G_2=\eta\triangleq e^{\frac{T}{2}+W(T)}-1, \end{cases} \tag{4.1}$$

where $W(T)$ is the value, at $t=T$, of the Brownian motion $W(\cdot)$ appearing in the state equation. We choose the above special form of η merely for the later computation to be easier. Other choices are also possible. The point here is that η is not deterministic. When (4.1) holds, we can rewrite the state equation (1.1) as follows:

$$\begin{cases} dx(t)=[u_1(t)+u_2(t)]dW(t), & t\in[0,T],\\ x(0)=\xi, \end{cases} \tag{4.2}$$

and the cost functionals can be written as follows:

$$J_i(\xi;u_1(\cdot),u_2(\cdot))=E\left\{\int_0^T|u_i(t)|^2dt+\eta x(T)^2\right\},\qquad i=1,2. \tag{4.3}$$

We see that the problem looks to be extremely simple. In this case, Riccati equation (1.5) reads

$$\begin{cases} dP=\dfrac{\Lambda^2}{P+1}dt+\Lambda dW(t), & t\in[0,T],\\ P(T)=\eta,\\ P(t)+1>0, & t\in[0,T],\ \text{a.s.} \end{cases} \tag{4.4}$$

One can check that the adapted solution of (4.4) is given by

$$\begin{cases} P(t) = e^{\frac{t}{2}+W(t)} - 1, \\ \Lambda(t) = P(t) + 1 \equiv e^{\frac{t}{2}+W(t)}. \end{cases} \tag{4.5}$$

Consequently, BSDE (1.6) takes the following form:

$$\begin{cases} d\varphi = (\theta - u_2)dt + \theta dW(t), \qquad t \in [0,T], \\ \varphi(T) = 0. \end{cases} \tag{4.5}$$

The optimal control $\overline{u}_1(\cdot)$ is given by

$$\overline{u}_1(t) = -x(t) - \frac{\theta(t) + P(t)u_2(t)}{P(t)+1}, \qquad t \in [0,T], \tag{4.6}$$

and the optimal cost is given by (note $P(0) = 0$)

$$\begin{aligned} &\inf_{u_1(\cdot)\in\mathcal{U}_1[0,T]} J_1(\xi, u_1(\cdot), u_2(\cdot)) = J_1(\xi, \overline{u}_1(\cdot), u_2(\cdot)) \\ &= 2\varphi(0)\xi + E\int_0^T \left\{ 2\theta(t)u_2(t) + P(t)u_2(t)^2 - \frac{|\theta(t) + P(t)u_2(t)|^2}{P(t)+1} \right\} dt \\ &= 2\varphi(0)\xi + E\int_0^T \frac{P(t)u_2(t)^2 + 2\theta(t)u_2(t) - \theta(t)^2}{P(t)+1} dt. \end{aligned} \tag{4.7}$$

We note that since $\theta(\cdot)$ (determined by (4.5)) is anticipating, so is $\overline{u}_1(\cdot)$ defined by (4.6). Next, with

$$\widehat{R}_1 \triangleq P + 1, \quad \widehat{S}_1 = \widehat{S}_2 \triangleq \Lambda \equiv P + 1, \tag{4.8}$$

we have

$$\begin{cases} \widehat{A} = \widehat{F}_1 = \widehat{B}_1 = \widehat{B}_2 = 0, \quad \widehat{C} \triangleq -\widehat{R}_1^{-1}\widehat{S}_1 = -1, \\ \widehat{D}_1 \triangleq -\widehat{R}_1^{-1} = -\dfrac{1}{P+1}, \quad \widehat{D}_2 \triangleq 1 - \dfrac{P}{P+1} = \dfrac{1}{P+1}, \\ \widehat{F}_2 \triangleq \widehat{S}_2 - P\widehat{R}_1^{-1}\widehat{S}_1 = 1. \end{cases} \tag{4.9}$$

Then

$$\begin{cases} \widehat{\mathcal{A}} = \widehat{\mathcal{F}}_1 = \widehat{\mathcal{B}}_1 = \widehat{\mathcal{Q}}_2 = 0, \quad \widehat{\mathcal{B}}_2 = 0, \\ \widehat{\mathcal{C}} = -I, \quad \widehat{\mathcal{D}}_1 = -\dfrac{1}{P+1}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \widehat{\mathcal{D}}_2 = \dfrac{1}{P+1}\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \\ \widehat{\mathcal{F}}_2 = \begin{pmatrix} 0 & 1 \end{pmatrix}, \quad \widehat{\mathcal{G}}_2 = \eta\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}. \end{cases} \tag{4.10}$$

Note that

$$\begin{aligned} (I - \widehat{P}\widehat{\mathcal{D}}_1)^{-1} &= \left(I + \frac{1}{P+1}\widehat{P}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right)^{-1} \\ &= (P+1)\begin{pmatrix} P+1+\widehat{P}_3 & \widehat{P}_2 \\ \widehat{P}_4 & P+1+\widehat{P}_3 \end{pmatrix}^{-1} \\ &= \frac{P+1}{(P+1+\widehat{P}_3)^2 - \widehat{P}_2\widehat{P}_4}\begin{pmatrix} P+1+\widehat{P}_3 & -\widehat{P}_2 \\ -\widehat{P}_4 & P+1+\widehat{P}_3 \end{pmatrix}. \end{aligned} \tag{4.11}$$

Thus,

$$\begin{aligned}
&(I-\widehat{P}\widehat{\mathcal{D}}_1)^{-1}\widehat{P}\\
&\quad=\frac{P+1}{(P+1+\widehat{P}_3)^2-\widehat{P}_2\widehat{P}_4}\begin{pmatrix}P+1+\widehat{P}_3 & -\widehat{P}_2\\ -\widehat{P}_4 & P+1+\widehat{P}_3\end{pmatrix}\begin{pmatrix}\widehat{P}_2 & \widehat{P}_3\\ \widehat{P}_3 & \widehat{P}_4\end{pmatrix}\\
&\quad=\frac{P+1}{(P+1+\widehat{P}_3)^2-\widehat{P}_2\widehat{P}_4}\begin{pmatrix}(P+1)\widehat{P}_2 & (P+1+\widehat{P}_3)\widehat{P}_3-\widehat{P}_2\widehat{P}_4\\ (P+1+\widehat{P}_3)\widehat{P}_3-\widehat{P}_2\widehat{P}_4 & (P+1)\widehat{P}_4\end{pmatrix}
\end{aligned}\tag{4.12}$$

and

$$\widehat{\mathcal{D}}_1(I-\widehat{P}\widehat{\mathcal{D}}_1)^{-1}=\frac{-1}{(P+1+\widehat{P}_3)^2-\widehat{P}_2\widehat{P}_4}\begin{pmatrix}-\widehat{P}_2 & P+1+\widehat{P}_3\\ P+1+\widehat{P}_3 & -\widehat{P}_4\end{pmatrix}.\tag{4.13}$$

Then it follows that

$$\begin{aligned}
\widehat{R}_2&\triangleq R_2+\widehat{\mathcal{D}}_2^T\big(I-\widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\widehat{P}\widehat{\mathcal{D}}_2\\
&=1+\frac{\widehat{P}_2}{(P+1+\widehat{P}_3)^2-\widehat{P}_2\widehat{P}_4}=\frac{(P+1+\widehat{P}_3)^2-\widehat{P}_2\widehat{P}_4+\widehat{P}_2}{(P+1+\widehat{P}_3)^2-\widehat{P}_2\widehat{P}_4}.
\end{aligned}\tag{4.14}$$

Hence, we can compute

$$\begin{aligned}
-\widehat{\Gamma}&\equiv-\widehat{\Gamma}(t,\widehat{P},\widehat{\Lambda})=\big(I-\widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\widehat{P}+\widehat{\Lambda}\widehat{\mathcal{D}}_1\big(I-\widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\widehat{\Lambda}\\
&\quad+\big(I-\widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\widehat{\Lambda}+\widehat{\Lambda}\big(I-\widehat{\mathcal{D}}_1\widehat{P}\big)^{-1}\\
&\quad-\big[\big(\widehat{P}+\widehat{\Lambda}\big)\big(I-\widehat{\mathcal{D}}_1\widehat{P}\big)^{-1}\widehat{\mathcal{D}}_2+\widehat{\mathcal{F}}_2^T\big]\widehat{R}_2^{-1}\big[\widehat{\mathcal{D}}_2^T\big(I-\widehat{P}\widehat{\mathcal{D}}_1\big)^{-1}\big(\widehat{P}+\widehat{\Lambda}\big)+\widehat{\mathcal{F}}_2\big]\\
&=\frac{P+1}{(P+1+\widehat{P}_3)^2-\widehat{P}_2\widehat{P}_4}\begin{pmatrix}(P+1)\widehat{P}_2 & (P+1+\widehat{P}_3)\widehat{P}_3-\widehat{P}_2\widehat{P}_4\\ (P+1+\widehat{P}_3)\widehat{P}_3-\widehat{P}_2\widehat{P}_4 & (P+1)\widehat{P}_4\end{pmatrix}\\
&\quad-\frac{1}{(P+1+\widehat{P}_3)^2-\widehat{P}_2\widehat{P}_4}\widehat{\Lambda}\begin{pmatrix}-\widehat{P}_2 & P+1+\widehat{P}_3\\ P+1+\widehat{P}_3 & -\widehat{P}_4\end{pmatrix}\widehat{\Lambda}\\
&\quad-\frac{P+1}{(P+1+\widehat{P}_3)^2-\widehat{P}_2\widehat{P}_4}\begin{pmatrix}P+1+\widehat{P}_3 & -\widehat{P}_2\\ -\widehat{P}_4 & P+1+\widehat{P}_3\end{pmatrix}\widehat{\Lambda}\\
&\quad-\frac{P+1}{(P+1+\widehat{P}_3)^2-\widehat{P}_2\widehat{P}_4}\widehat{\Lambda}\begin{pmatrix}P+1+\widehat{P}_3 & -\widehat{P}_4\\ -\widehat{P}_2 & P+1+\widehat{P}_3\end{pmatrix}\\
&\quad-\frac{(P+1+\widehat{P}_3)^2-\widehat{P}_2\widehat{P}_4}{(P+1+\widehat{P}_3)^2-\widehat{P}_2\widehat{P}_4+\widehat{P}_2}\\
&\quad\cdot\left\{\frac{1}{(P+1+\widehat{P}_3)^2-\widehat{P}_2\widehat{P}_4}(\widehat{P}+\widehat{\Lambda})\begin{pmatrix}P+1+\widehat{P}_3 & -\widehat{P}_4\\ -\widehat{P}_2 & P+1+\widehat{P}_3\end{pmatrix}\begin{pmatrix}1\\0\end{pmatrix}+\begin{pmatrix}0\\1\end{pmatrix}\right\}\\
&\quad\cdot\left\{\frac{1}{(P+1+\widehat{P}_3)^2-\widehat{P}_2\widehat{P}_4}\begin{pmatrix}1 & 0\end{pmatrix}\begin{pmatrix}P+1+\widehat{P}_3 & -\widehat{P}_2\\ -\widehat{P}_4 & P+1+\widehat{P}_3\end{pmatrix}(\widehat{P}+\widehat{\Lambda})+\begin{pmatrix}0 & 1\end{pmatrix}\right\}\\
&=\frac{P+1}{(P+1+\widehat{P}_3)^2-\widehat{P}_2\widehat{P}_4}\begin{pmatrix}(P+1)\widehat{P}_2 & (P+1+\widehat{P}_3)\widehat{P}_3-\widehat{P}_2\widehat{P}_4\\ (P+1+\widehat{P}_3)\widehat{P}_3-\widehat{P}_2\widehat{P}_4 & (P+1)\widehat{P}_4\end{pmatrix}\\
&\quad-\frac{1}{(P+1+\widehat{P}_3)^2-\widehat{P}_2\widehat{P}_4}\widehat{\Lambda}\begin{pmatrix}-\widehat{P}_2 & P+1+\widehat{P}_3\\ P+1+\widehat{P}_3 & -\widehat{P}_4\end{pmatrix}\widehat{\Lambda}
\end{aligned}$$

$$
\begin{aligned}
&- \frac{P+1}{(P+1+\widehat{P}_3)^2-\widehat{P}_2\widehat{P}_4}\begin{pmatrix} P+1+\widehat{P}_3 & -\widehat{P}_2 \\ -\widehat{P}_4 & P+1+\widehat{P}_3 \end{pmatrix}\widehat{\Lambda} \\
&- \frac{P+1}{(P+1+\widehat{P}_3)^2-\widehat{P}_2\widehat{P}_4}\widehat{\Lambda}\begin{pmatrix} P+1+\widehat{P}_3 & -\widehat{P}_4 \\ -\widehat{P}_2 & P+1+\widehat{P}_3 \end{pmatrix} \\
&- \frac{(P+1+\widehat{P}_3)^2-\widehat{P}_2\widehat{P}_4}{(P+1+\widehat{P}_3)^2-\widehat{P}_2\widehat{P}_4+\widehat{P}_2} \\
&\cdot \left\{ \frac{1}{[(P+1+\widehat{P}_3)^2-\widehat{P}_2\widehat{P}_4]^2}(\widehat{P}+\widehat{\Lambda})\begin{pmatrix} (P+1+\widehat{P}_3)^2 & -\widehat{P}_2(P+1+\widehat{P}_3) \\ -\widehat{P}_2(P+1+\widehat{P}_3) & \widehat{P}_2^2 \end{pmatrix}(\widehat{P}+\widehat{\Lambda}) \right. \\
&+ \frac{1}{(P+1+\widehat{P}_3)^2-\widehat{P}_2\widehat{P}_4}(\widehat{P}+\widehat{\Lambda})\begin{pmatrix} 0 & P+1+\widehat{P}_3 \\ 0 & -\widehat{P}_2 \end{pmatrix} \\
&\left. + \frac{1}{(P+1+\widehat{P}_3)^2-\widehat{P}_2\widehat{P}_4}\begin{pmatrix} 0 & 0 \\ P+1+\widehat{P}_3 & -\widehat{P}_2 \end{pmatrix}(\widehat{P}+\widehat{\Lambda}) + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right\}.
\end{aligned}
$$

From the above, we can easily see that in the Riccati equation

$$
\begin{cases}
d\widehat{P}(t) = \widehat{\Gamma}(t,\widehat{P}(t),\widehat{\Lambda}(t))dt + \widehat{\Lambda}(t)dW(t), & t\in[0,T], \\
\widehat{P}(T) = 0, & \\
P(t)+1+\widehat{P}_3(t)-\widehat{P}_2(t)\widehat{P}_4(t)\neq 0, & t\in[0,T], \\
[P(t)+1+\widehat{P}_3(t)]^2-\widehat{P}_2(t)\widehat{P}_4(t)+\widehat{P}_2(t)\neq 0, & t\in[0,T],
\end{cases} \tag{4.15}
$$

the components $\widehat{P}_2(\cdot)$, $\widehat{P}_3(\cdot)$, and $\widehat{P}_4(\cdot)$ are heavily coupled. Thus, an explicit solution of (4.16) is not expected. The reason for the complexity is the appearance of the controls in the diffusion and the randomness of the coefficients.

Let us now look at another special case:

$$
\begin{cases}
n = m_1 = m_2 = 1, \quad A = C = B_1 = B_2 = Q_1 = Q_2 = R_1 = R_2 = 0, \\
D_1 = D_2 = G_1 = -G_2 = 1.
\end{cases} \tag{4.16}
$$

Then the state equation is the same as (4.2) and the cost functionals are

$$
J(\xi;u_1(\cdot),u_2(\cdot)) \triangleq J_1(\xi;u_1(\cdot),u_2(\cdot)) = E\Big\{|x(T)|^2\Big\} = -J_2(\xi;u_1(\cdot),u_2(\cdot)). \tag{4.17}
$$

Thus, our leader-follower differential game becomes a zero-sum differential game with the state equation (4.2) and the cost functional $J(\xi;u_1(\cdot),u_2(\cdot))$, in which Player 1 is the minimizer and Player 2 is the maximizer. If we denote the upper and lower values (in the sense of Elliot–Kalton [10]) by $V^+(t,x)$ and $V^-(t,x)$, respectively, then a formal application of Bellman's dynamic programming principle leads to the following Isaacs equations for these two functions:

$$
\begin{cases}
V_t^+(t,x) + H^+(V_{xx}^+(t,x)) = 0, \\
V^+(T,x) = x^2,
\end{cases} \tag{4.18}
$$

$$
\begin{cases}
V_t^-(t,x) + H^-(V_{xx}^-(t,x)) = 0, \\
V^-(T,x) = x^2,
\end{cases} \tag{4.19}
$$

where

$$\begin{cases} H^+(K) \triangleq \inf\limits_{u_1\in\mathbb{R}^{m_1}} \sup\limits_{u_2\in\mathbb{R}^{m_2}} \left\{\dfrac{1}{2}(u_1+u_2)^2K\right\} = \begin{cases} 0, & K\le 0,\\ +\infty, & K>0,\end{cases}\\ H^-(K) \triangleq \sup\limits_{u_2\in\mathbb{R}^{m_2}} \inf\limits_{u_1\in\mathbb{R}^{m_1}} \left\{\dfrac{1}{2}(u_1+u_2)^2K\right\} = \begin{cases} 0, & K\ge 0,\\ -\infty, & K<0.\end{cases}\end{cases} \tag{4.20}$$

Clearly, the usual Isaacs condition does not hold. We can check that the solutions of (4.19) and (4.20) are given by the following:

$$V^+(t,x) = \begin{cases} +\infty, & (t,x)\in[0,T)\times\mathbb{R},\\ x^2, & t=T,\ x\in\mathbb{R},\end{cases} \tag{4.21}$$

and

$$V^-(t,x) = x^2, \qquad (t,x)\in[0,T]\times\mathbb{R}. \tag{4.22}$$

They are not equal. Thus, the value (in the sense of Elliot–Kalton) does not exist. On the other hand, we may look at the problem using open-loop strategies. Since

$$\begin{cases} J_1(\xi;u_1(\cdot),u_2(\cdot)) = \xi^2 + E\displaystyle\int_0^T |u_1(t)+u_2(t)|^2dt,\\ J_2(\xi;u_1(\cdot),u_2(\cdot)) = -\xi^2 - E\displaystyle\int_0^T |u_1(t)+u_2(t)|^2dt,\end{cases} \tag{4.23}$$

we see easily that

$$\begin{cases} \inf\limits_{u_1(\cdot)\in\mathcal{U}_1[0,T]} J_1(\xi;u_1(\cdot),u_2(\cdot)) = \xi^2 \quad \forall u_2(\cdot)\in\mathcal{U}_2[0,T],\\ \inf\limits_{u_2(\cdot)\in\mathcal{U}_2[0,T]} J_2(\xi;u_1(\cdot),u_2(\cdot)) = -\infty \quad \forall u_1(\cdot)\in\mathcal{U}_1[0,T].\end{cases} \tag{4.24}$$

This coincides with the above conclusion.

Remark 4.1. The above shows that if the controls enter into the diffusion, one might not expect to have the existence of the value for a zero-sum differential game (unless some other compatibility conditions hold between the state equation and the cost functional; we hope to address this in future work). The main reason is that the Isaacs-type condition does not necessarily hold in the current case. At this moment, we might realize why almost all formulations of stochastic differential games in the literature avoided the controls entering the diffusion. Due to this, we study only the leader-follower case in the present paper. This can be regarded as a first step to approaching the stochastic differential games with controls in the diffusion.

5. Deterministic coefficient cases. In this section, we concentrate on the case in which all the coefficients are deterministic. For simplicity, we will consider only the constant coefficient case. To be more precise, we introduce the following assumption.

(DI) Let

$$A,C\in\mathbb{R}^{n\times n}, \quad B_i,D_i\in\mathbb{R}^{n\times m_i}, \quad Q_i,G_i\in\mathcal{S}^n, \quad R_i\in\mathcal{S}^{m_i}, \qquad i=1,2. \tag{5.1}$$

According to [6, 7], under (DI), the Riccati equation for Problem $(\mathrm{LQ})_1$ takes the following form:

$$\begin{cases} \dot{P} + PA + A^T P + C^T PC + Q_1 \\ \quad - (PB_1 + C^T PD_1)(R_1 + D_1^T PD_1)^{-1}(B_1^T P + D_1^T PC) = 0, \qquad t \in [0,T], \\ P(T) = G_1, \\ R_1 + D_1^T P(t) D_1 > 0 \qquad \text{a.e. } t \in [0,T]. \end{cases} \tag{5.2}$$

In fact, if $P(\cdot)$ is a solution of (5.2), then $(P(\cdot), 0)$ is an adapted solution of (1.5), with which (1.6) becomes

$$\begin{cases} d\varphi = -\Big\{ [A^T - (PB_1 + C^T PD_1)(R_1 + D_1^T PD_1)^{-1} B_1^T]\varphi \\ \qquad\qquad + [C^T - (PB_1 + C^T PD_1)(R_1 + D_1^T PD_1)^{-1} D_1^T]\theta \\ \qquad\qquad + \big[- (PB_1 + C^T PD_1)(R_1 + D_1^T PD_1)^{-1} D_1^T PD_2 \\ \qquad\qquad + PB_2 + C^T PD_2 \big] u_2 \Big\} dt + \theta dW(t), \qquad t \in [0,T], \\ \varphi(T) = 0. \end{cases} \tag{5.3}$$

This is still a BSDE, since $u_2(\cdot) \in \mathcal{U}_2[0,T]$ is random in general. For the current case, we still have Proposition 2.2 and Theorem 2.3. Note that these results are incomplete in the sense that we still do not know when Problem $(\mathrm{LQ})_1$ is solvable, unless we know when either the FBSDE (2.5) (together with (2.6)) or the Riccati equation (5.2) is solvable. Our first goal of this section is to present some sufficient conditions for the solvability of (2.5) and/or (5.2). We now introduce the following further assumption.

(H1) Matrix R_1 has an inverse. Moreover,

$$B_1 R_1^{-1} D_1^T = 0, \quad C = 0, \tag{5.4}$$

and

$$R_1 + D_1^T G_1 D_1 > 0. \tag{5.5}$$

We emphasize here that R_1 is not necessarily positive semidefinite. Also, B_1 and D_1 are not necessarily zero. Here is a simple example:

$$B_1 = D_1 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad R_1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{5.6}$$

One can easily check that (H1) holds for B_1, D_1 and R_1 given in (5.6). The following lemma tells us the role that assumption (H1) is playing.

LEMMA 5.1. *Let R_1^{-1} exist and let the first relation in* (5.4) *hold. Then for any $P \in \mathbb{R}^{n\times n}$ such that $(R_1 + D_1^T PD_1)^{-1}$ exists, the following holds:*

$$B_1(R_1 + D_1^T PD_1)^{-1} = B_1 R_1^{-1}. \tag{5.7}$$

Proof. We observe the following:

$$\begin{aligned} &B_1(R_1 + D_1^T PD_1)^{-1} - B_1 R_1^{-1} = B_1\big[(R_1 + D_1^T PD_1)^{-1} - R_1^{-1}\big] \\ &= B_1 R_1^{-1}[R_1 - R_1 - D_1^T PD_1](R_1 + D_1^T PD_1)^{-1} \\ &= -B_1 R_1^{-1} D_1^T PD_1 (R_1 + D_1^T PD_1)^{-1} = 0, \end{aligned} \tag{5.8}$$

proving (5.7). □

Under assumption (H1) (noting Lemma 5.1), Riccati equation (5.2) takes the following form:

$$\begin{cases} \dot{P} + PA + A^T P + Q_1 - PB_1 R_1^{-1} B_1^T P = 0, & t \in [0,T], \\ P(T) = G_1, \\ R_1 + D_1^T P(t) D_1 > 0 \qquad \text{a.e. } t \in [0,T], \end{cases} \tag{5.9}$$

and BSDE (5.3) becomes

$$\begin{cases} d\varphi = -\Big\{[A^T - PB_1 R_1^{-1} B_1^T]\varphi - PB_1 R_1^{-1} D_1^T \theta \\ \qquad\qquad - [PB_1 R_1^{-1} D_1^T P D_2 - PB_2]u_2\Big\}dt + \theta dW(t), \qquad t \in [0,T], \\ \varphi(T) = 0. \end{cases} \tag{5.10}$$

On the other hand, the FBSDE (2.5) becomes (the bars are dropped)

$$\begin{cases} dx(t) = [Ax(t) + B_1 u_1(t) + B_2 u_2(t)]dt + [D_1 u_1(t) + D_2 u_2(t)]dW(t), \\ dp(t) = -[A^T p(t) + Q_1 x(t)]dt + q(t)dW(t), \\ x(0) = \xi, \qquad p(T) = G_1 x(T). \end{cases} \tag{5.11}$$

We note that without looking at the third constraint, (5.9) looks like a standard Riccati equation for deterministic LQ problems. However, we should keep in mind that there are no positive semidefiniteness conditions imposed on either Q_1 and R_1. Also, the third constraint is not obviously to be satisfied if there is no positivity of Q_1 and R_1.

Now, let $(x(\cdot), u_1(\cdot))$ be an optimal pair of Problem $(\mathrm{LQ})_1$ and let $(p(\cdot), q(\cdot))$ be the adapted solution of BSDE in (5.11). Since R_1^{-1} exists, the optimal control $\overline{u}_1(\cdot)$ has the form (note (5.4))

$$\overline{u}_1(t) = -R_1^{-1}[B_1^T p(t) + D_1^T q(t)] = -R_1^{-1} B_1^T p(t), \qquad t \in [0,T]. \tag{5.12}$$

Plugging (5.12) into (5.11), we obtain

$$\begin{cases} dx(t) = \big[Ax(t) - B_1 R_1^{-1} B_1^T p(t) + B_2 u_2(t)\big]dt \\ \qquad\qquad + [-D_1 R_1^{-1} D_1^T q(t) + D_2 u_2(t)]dW(t), \\ dp(t) = -\big[Q_1 x(t) + A^T p(t)\big]dt + q(t)dw(t), \\ x(0) = \xi, \qquad p(T) = G_1 x(T). \end{cases} \tag{5.13}$$

This is a coupled linear FBSDE. To obtain sufficient conditions for the solvability of such an FBSDE via the result of [32], let us make some reductions. For notational

simplicity, in what follows we suppress the argument t. Let $\eta = p - G_1 x$. Then

$$
\begin{aligned}
d\eta &= dp - G_1 dx \\
&= \big[- Q_1 x - A^T p - G_1(Ax - B_1 R_1^{-1} B_1^T p + B_2 u_2)\big] dt \\
&\quad + \big[(I + G_1 D_1 R_1^{-1} D_1^T) q - G_1 D_2 u_2\big] dW \\
&= \big[(-Q_1 - G_1 A)x + (-A^T + G_1 B_1 R_1^{-1} B_1^T)(\eta + G_1 x) - G_1 B_2 u_2\big] dt \\
&\quad + \big[(I + G_1 D_1 R_1^{-1} D_1^T) q - G_1 D_2 u_2\big] dW \\
&= \big[(-Q_1 - G_1 A - A^T G_1 + G_1 B_1 R_1^{-1} B_1^T G_1) x \\
&\quad + (-A^T + G_1 B_1 R_1^{-1} B_1^T)\eta - G_1 B_2 u_2\big] dt \\
&\quad + \big[(I + G_1 D_1 R_1^{-1} D_1^T) q - G_1 D_2 u_2\big] dW.
\end{aligned} \tag{5.14}
$$

We define

$$
\zeta = (I + G_1 D_1 R^{-1} D_1^T) q. \tag{5.15}
$$

A direct computation shows that

$$
(I + G_1 D_1 R_1^{-1} D_1^T)^{-1} = I - G_1 D_1 (R_1 + D_1^T G_1 D_1)^{-1} D_1^T. \tag{5.16}
$$

This means that $(I + G_1 D_1 R_1^{-1} D_1^T)^{-1}$ exists, and from (5.15) we have

$$
q = (I + G_1 D_1 R_1^{-1} D_1^T)^{-1} \zeta. \tag{5.17}
$$

Consequently,

$$
\begin{aligned}
dx &= \big[Ax - B_1 R_1^{-1} B_1^T (\eta + G_1 x) + B_2 u_2\big] dt \\
&\qquad + \big[- D_1 R_1^{-1} D_1^T (I + G_1 D_1 R_1^{-1} D_1^T)^{-1} \zeta + D_2 u_2\big] dW \\
&= \big[(A - B_1 R_1^{-1} B_1^T G_1) x - B_1 R_1^{-1} B_1^T \eta + B_2 u_2\big] dt \\
&\qquad + [-D_1 (R_1 + D_1^T G_1 D_1)^{-1} D_1^T \zeta + D_2 u_2] dW.
\end{aligned} \tag{5.18}
$$

Also,

$$
\begin{aligned}
d\eta = \big[&(-Q_1 - G_1 A - A^T G_1 + G_1 B_1 R_1^{-1} B_1^T G_1) x \\
&+ (-A^T + G_1 B_1 R_1^{-1} B_1^T)\eta - G_1 B_2 u_2\big] dt + [\zeta - G_2 D_2 u_2] dW.
\end{aligned} \tag{5.19}
$$

Hence, (5.11) becomes

$$
\begin{cases}
dx(t) = \big[(A - B_1 R_1^{-1} B_1^T G_1) x(t) - B_1 R_1^{-1} B_1^T \eta(t) + B_2 u_2(t)\big] dt \\
\qquad\quad + \big[- D_1 (R_1 + D_1^T G_1 D_1)^{-1} D_1^T \zeta(t) + D_2 u_2(t)\big] dW(t), \\
d\eta(t) = \big[(-Q_1 - G_1 A - A^T G_1 + G_1 B_1 R_1^{-1} B_1^T G_1) x(t) \\
\qquad\quad + (-A^T + G_1 B_1 R_1^{-1} B_1^T)\eta(t) - G_1 B_2 u_2(t)\big] dt \\
\qquad\quad + [\zeta(t) - G_1 D_2 u_2(t)] dW(t), \\
x(0) = \xi, \qquad \eta(T) = 0.
\end{cases} \tag{5.20}
$$

We denote

$$
\begin{aligned}
&\mathcal{A} \triangleq \begin{pmatrix} A - B_1 R_1^{-1} B_1^T G_1 & -B_1 R_1^{-1} B_1^T \\ -Q_1 - G_1 A - A^T G_1 + G_1 B_1 R_1^{-1} B_1^T G_1 & -A^T + G_1 B_1 R_1^{-1} B_1^T \end{pmatrix}, \\
&\mathcal{C}_1 \triangleq \begin{pmatrix} -D_1 (R_1 + D_1^T G_1 D_1)^{-1} D_1^T \\ I \end{pmatrix}, \quad \mathcal{B}_2 \triangleq \begin{pmatrix} B_2 \\ -G_1 B_2 \end{pmatrix}, \quad \mathcal{D}_2 \triangleq \begin{pmatrix} D_2 \\ -G_1 D_2 \end{pmatrix}.
\end{aligned} \tag{5.21}
$$

Then (5.20) can further be written as

$$\begin{cases} d\begin{pmatrix} x(t) \\ \eta(t) \end{pmatrix} = \Big\{\mathcal{A}\begin{pmatrix} x(t) \\ \eta(t) \end{pmatrix} + \mathcal{B}_2 u_2(t)\Big\}dt + \Big\{\mathcal{C}_1\zeta(t) + \mathcal{D}_2 u_2(t)\Big\}dW(t), \\ x(0) = \xi, \quad \eta(T) = 0. \end{cases} \tag{5.22}$$

Next, we introduce the following Riccati equation for FBSDE (5.20):

$$\begin{cases} \dot\Pi(t) + \Pi(t)(A - B_1R_1^{-1}B_1^TG_1) + (A - B_1R_1^{-1}B_1^TG_1)^T\Pi(t) \\ \qquad - \Pi(t)B_1R_1^{-1}B_1^T\Pi(t) + [Q_1 + G_1A + A^TG_1 - G_1B_1R_1^{-1}B_1^TG_1] = 0, \\ \Pi(T) = 0. \end{cases} \tag{5.23}$$

It is easy to see that solution $\Pi(\cdot)$ of (5.23) and that $P(\cdot)$ of (5.9) are related by the following:

$$P(t) = G_1 + \Pi(t), \qquad t \in [0,T]. \tag{5.24}$$

The following was proved in [32].

THEOREM 5.2. *Let* (H1) *hold and let*

$$\begin{cases} \det\Big\{ \begin{pmatrix} 0 & I \end{pmatrix} e^{\mathcal{A}t}\begin{pmatrix} 0 \\ I \end{pmatrix}\Big\} > 0, \\ \det\Big\{ \begin{pmatrix} 0 & I \end{pmatrix} e^{\mathcal{A}t}\mathcal{C}_1\Big\} > 0, \end{cases} \qquad t \in [0,T]. \tag{5.25}$$

Then (5.23) *admits a unique solution* $\Pi(\cdot)$ *which has the following representation:*

$$\Pi(t) = -\Big[\begin{pmatrix} 0 & I \end{pmatrix} e^{\mathcal{A}(T-t)}\begin{pmatrix} 0 \\ I \end{pmatrix}\Big]^{-1}\begin{pmatrix} 0 & I \end{pmatrix} e^{\mathcal{A}(T-t)}\begin{pmatrix} I \\ 0 \end{pmatrix}, \qquad t \in [0,T]. \tag{5.26}$$

Moreover, (5.24) *gives the solution of* (5.9).

Combining Theorem 5.2 and (5.24), we see that when (H1) and (5.25) hold, Riccati equation (5.9) admits a unique solution $P(\cdot)$, which leads to the solvability of Problem $(\mathrm{LQ})_1$ for any given $(\xi, u_2(\cdot)) \in \mathbb{R}^n \times \mathcal{U}_2[0,T]$. It is important to note that conditions (H1) and (5.25) are checkable, in principle.

Now, under (H1) and (5.25), we have the solution $P(\cdot)$ of Riccati equation (5.9) and FBSDE (1.9) becomes

$$\begin{cases} dx(t) = \big[\widehat{A}x(t) + \widehat{F}_1\varphi(t) + B_2u_2(t)\big]dt + \big[\widehat{D}_1\theta(t) + \widehat{D}_2u_2(t)\big]dW(t), \\ d\varphi(t) = -\big[\widehat{A}^T\varphi(t) + PB_2u_2(t)\big]dt + \theta(t)dW(t), \\ x(0) = \xi, \qquad \varphi(T) = 0, \end{cases} \tag{5.27}$$

where

$$\begin{cases} \widehat{A} \triangleq A - B_1R_1^{-1}B_1^TP, \qquad \widehat{F}_1 \triangleq -B_1R_1^{-1}B_1^T, \\ \widehat{D}_1 \triangleq -D_1\widehat{R}_1^{-1}D_1^T, \qquad \widehat{D}_2 \triangleq D_2 - D_1\widehat{R}_1^{-1}D_1^TPD_2. \end{cases} \tag{5.28}$$

FBSDE (3.6) becomes (we again drop the bars)

$$\begin{cases} dy = (\widehat{A}^Ty + Q_2x)dt + zdW(t), \\ d\psi = (\widehat{A}\psi + \widehat{F}_1y)dt + \widehat{D}_1zdW(t), \\ y(T) = G_2x(T), \quad \psi(0) = 0. \end{cases} \tag{5.29}$$

Next, similarly to (H1), we introduce the following assumption.

(H2) Let R_2^{-1} exist. Moreover,

$$B_2R_2^{-1}D_2^T = 0 \tag{5.30}$$

and

$$R_2 + D_2^T G_2 D_2 > 0. \tag{5.31}$$

We note that under (H2), the following holds:

$$B_2R_2^{-1}\widehat{D}_2^T = B_2R_2^{-1}D_2^T(I - D_1\widehat{R}_1^{-1}D_1^T) = 0. \tag{5.32}$$

Further, under (H1)–(H2), (3.7) implies

$$\overline{u}_2 = -R_2^{-1}\big(B_2^T y + \widehat{D}_2^T z + B_2^T P\psi\big). \tag{5.33}$$

Thus, by setting $X = \begin{pmatrix} x \\ \psi \end{pmatrix}$, $Y = \begin{pmatrix} y \\ \varphi \end{pmatrix}$, and $Z = \begin{pmatrix} z \\ \theta \end{pmatrix}$, we have the following:

$$\begin{cases} dX = \left\{ \begin{pmatrix} \widehat{A} & -B_2R_2^{-1}B_2^TP \\ 0 & \widehat{A} \end{pmatrix} X + \begin{pmatrix} -B_2R_2^{-1}B_2^T & \widehat{F}_1 \\ \widehat{F}_1 & 0 \end{pmatrix} Y \right\} dt \\ \qquad\quad + \begin{pmatrix} -\widehat{D}_2R_2^{-1}\widehat{D}_2 & \widehat{D}_1 \\ \widehat{D}_1 & 0 \end{pmatrix} Z dW(t), \\ dY = -\left\{ \begin{pmatrix} Q_2 & 0 \\ 0 & -PB_2R_2^{-1}B_2^TP \end{pmatrix} X + \begin{pmatrix} \widehat{A}^T & 0 \\ -PB_2R_2^{-1}B_2^T & \widehat{A}^T \end{pmatrix} Y \right\} dt \\ \qquad\quad + Z dW(t), \\ X(0) = \begin{pmatrix} \xi \\ 0 \end{pmatrix} \triangleq \widehat{\xi}, \qquad Y(T) = \begin{pmatrix} G_2 & 0 \\ 0 & 0 \end{pmatrix} X(T) \triangleq \widehat{G}_2 X(T). \end{cases} \tag{5.34}$$

Now, we set

$$\widehat{\mathcal{A}} = \begin{pmatrix} \widehat{A} & -B_2R_2^{-1}B_2^TP & -B_2R_2^{-1}B_2^T & \widehat{F}_1 \\ 0 & \widehat{A} & \widehat{F}_1 & 0 \\ -Q_2 & 0 & -\widehat{A}^T & 0 \\ 0 & PB_2R_2^{-1}B_2^TP & PB_2R_2^{-1}B_2^T & -\widehat{A}^T \end{pmatrix}, \quad \widehat{\mathcal{C}}_1 = \begin{pmatrix} -\widehat{D}_2R_2^{-1}\widehat{D}_2 & \widehat{D}_1 \\ \widehat{D}_1 & 0 \\ I & 0 \\ 0 & I \end{pmatrix}.$$

Then (5.34) becomes

$$\begin{cases} d\begin{pmatrix} X(t) \\ Y(t) \end{pmatrix} = \widehat{\mathcal{A}}(t)\begin{pmatrix} X(t) \\ Y(t) \end{pmatrix} dt + \widehat{\mathcal{C}}_1(t)Z(t)dW(t), \\ X(0) = \widehat{\xi}, \qquad Y(T) = \widehat{G}_2X(T). \end{cases} \tag{5.35}$$

Note that (5.35) is an FBSDE with time-dependent and deterministic coefficients. Thus, we need only take $Z(\cdot) = 0$ and $(X(\cdot), Y(\cdot))$ to be a solution of the following ODE:

$$\begin{cases} d\begin{pmatrix} X(t) \\ Y(t) \end{pmatrix} = \widehat{\mathcal{A}}(t)\begin{pmatrix} X(t) \\ Y(t) \end{pmatrix} dt, \\ X(0) = \widehat{\xi}, \qquad Y(T) = \widehat{G}_2X(T). \end{cases} \tag{5.36}$$

Hence, by letting $\Theta(\cdot)$ be the solution of

$$\begin{cases} \dfrac{d}{dt}\Theta(t,s) = \widehat{\mathcal{A}}(t)\Theta(t,s), & t \in [s,T], \\ \Theta(s,s) = I, \end{cases} \tag{5.37}$$

one knows that any solution $(X(\cdot), Y(\cdot))$ is given by

$$\begin{pmatrix} X(t) \\ Y(t) \end{pmatrix} = \Theta(t,0) \begin{pmatrix} \widehat{\xi} \\ \widehat{\eta} \end{pmatrix}, \qquad t \in [0,T], \tag{5.38}$$

with $\widehat{\eta}$ being in $\mathbb{R}^m$ such that

$$0 = \begin{pmatrix} -\widehat{G}_2 & I \end{pmatrix} \Theta(T,0) \begin{pmatrix} \widehat{\xi} \\ \widehat{\eta} \end{pmatrix}. \tag{5.39}$$

This is equivalent to

$$\begin{pmatrix} \widehat{G}_2 & -I \end{pmatrix} \Theta(T,0) \begin{pmatrix} \widehat{\xi} \\ 0 \end{pmatrix} = \begin{pmatrix} -\widehat{G}_2 & I \end{pmatrix} \Theta(T,0) \begin{pmatrix} 0 \\ I \end{pmatrix} \widehat{\eta}. \tag{5.40}$$

This proves the following.

THEOREM 5.3. *Let* (DI), (H1), (H2), *and* (5.25) *hold. Suppose further that*

$$\det \left\{ \begin{pmatrix} -\widehat{G}_2 & I \end{pmatrix} \Theta(T,0) \begin{pmatrix} 0 \\ I \end{pmatrix} \right\} \neq 0. \tag{5.41}$$

Then (5.35) *is solvable. Consequently, if the conclusions of Proposition* 2.2(i) *and of Theorem* 3.2(i) *hold, in particular, if* Q_i, R_i, *and* G_i *are nonnegative, then the leader-follower differentiable game admits an open-loop solution which admits a state feedback representation.*

REFERENCES

[1] T. BASAR AND G. J. OLSDER, *Dynamic Noncooperative Game Theory,* SIAM, Philadelphia, 1995.

[2] A. BENSOUSSAN, *Points de Nash dans le cas de fonctionnelles quadratiques et jeux differentiels lineaires a N personnes,* SIAM J. Control, 12 (1974), pp. 460–499.

[3] L. D. BERKOVITZ, *A differential game with no pure strategy*, in Advances in Game Theory, Ann. of Math. Stud. 52, Princeton University Press, Princeton, NJ, 1964, pp. 175–194.

[4] L. D. BERKOVITZ, *The existence of value and saddle point in games of fixed duration,* SIAM J. Control Optim., 23 (1985), pp. 172–196.

[5] J.-M. BISMUT, *Linear quadratic optimal stochastic control with random coefficients,* SIAM J. Control Optim., 14 (1976), pp. 419–444.

[6] S. CHEN, X. LI, AND X. Y. ZHOU, *Stochastic linear quadratic regulators with indefinite control weight costs,* SIAM J. Control Optim., 36 (1998), pp. 1685–1702.

[7] S. CHEN AND J. YONG, *Stochastic linear quadratic optimal control problems with random coefficients,* Chinese Ann. Math. Ser. B, 21 (2000), pp. 323–338.

[8] S. CHEN AND J. YONG, *Stochastic linear quadratic optimal control problems,* Appl. Math. Optim., 43 (2001), pp. 21–45.

[9] T. EISELE, *Nonexistence and nonuniqueness of open-loop equilibrium in linear-quadratic differential games,* J. Optim. Theory Appl., 37 (1982), pp. 443–468.

[10] R. J. ELLIOT AND N. J. KALTON, *The existence of value in differential games,* Mem. Amer. Math. Soc., 126 (1972).

[11] W. H. FLEMING AND P. E. SOUGANIDIS, *On the existence of value functions of two-player, zero-sum stochastic differential games,* Indiana Univ. Math. J., 38 (1989), pp. 293–314.

[12] A. Friedman, *Differential Games,* John Wiley & Sons, New York, 1971.
[13] A. Friedman, *Stochastic differential games*, J. Differential Equations, 11 (1972), pp. 79–108.
[14] S. Hamadene, *Nonzero sum linear-quadratic stochastic differential games and backward-forward equations,* Stochastic Anal. Appl., 17 (1999), pp. 117–130.
[15] S. Hamadene and J.-P. Lepeltier, *Backward equations, stochastic control and zero-sum stochastic differential games,* Stochastics Stochastics Rep., 54 (1995), pp. 221–231.
[16] Y. Hu, *N-person differential games governed by semilinear stochastic evolution systems,* Appl. Math. Optim., 24 (1991), pp. 257–271.
[17] A. Ichikawa, *Linear quadratic differential games in a Hilbert space,* SIAM J. Control Optim., 14 (1976), pp. 120–136.
[18] R. Isaacs, *Differential Games,* John Wiley & Sons, New York, 1965.
[19] R. E. Kalman, *Contributions to the theory of optimal control,* Bol. Soc. Mat. Mexicana, 5 (1960), pp. 102–119.
[20] N. N. Krasovskii, *Stabilization of systems in which noise is dependent on the value of the control signal,* Eng. Cybern. (USSR), 3 (1965), pp. 94–102.
[21] N. N. Krasovskii and A. I. Subbotin, *Game-Theoretical Control Problems,* Springer-Verlag, New York, 1988.
[22] X. Li, *N-person differential games governed by infinite dimensional systems,* J. Optim. Theory Appl., 50 (1986), pp. 431–450.
[23] A. Lim and X. Y. Zhou, *Linear-Quadratic Control of Backward Stochastic Differential Equations,* preprint.
[24] J. Ma and J. Yong, *Forward-Backward Stochastic Differential Equations and Their Applications,* Springer-Verlag, Berlin, 1999.
[25] P. J. McLane, *Optimal stochastic control of linear systems with state- and control-dependent disturbances,* IEEE Trans. Automat. Control, 16 (1971), pp. 793–798.
[26] L. Pan and J. Yong, *A differential game with multi-level of hierarchy,* J. Math. Anal. Appl., 161 (1991), pp. 522–544.
[27] E. Pardoux and S. Peng, *Adapted solutions of backward stochastic equations,* Systems Control Lett., 14 (1990), pp. 55–61.
[28] W. E. Schmitendorf, *Existence of optimal open-loop strategies for a class of differential games,* J. Optim. Theory Appl., 5 (1970), pp. 363–375.
[29] K. Uchida, *On the existence of Nash equilibrium points in N-person nonzero sum stochastic differential games,* SIAM J. Control Optim., 16 (1978), pp. 142–149.
[30] P. Varaiya, *N-player stochastic differential games,* SIAM J. Control Optim., 14 (1976), pp. 538–545.
[31] W. M. Wonham, *On a matrix Riccati equation of stochastic control,* SIAM J. Control, 6 (1968), pp. 681–697.
[32] J. Yong, *Linear forward-backward stochastic differential equations,* Appl. Math. Optim., 39 (1999), pp. 93–119.
[33] J. Yong, *Forward-backward stochastic differential equation—a useful tool for mathematical finance and other related fields,* Survey Industry Math., to appear.
[34] J. Yong and X. Y. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations,* Springer-Verlag, New York, 1999.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 4, pp. 1042–1072

THE TOPOLOGICAL ASYMPTOTIC EXPANSION FOR THE DIRICHLET PROBLEM*

PH. GUILLAUME† AND K. SID IDRIS†

Abstract. The topological sensitivity analysis provides an asymptotic expansion of a shape function when creating a small hole inside a domain. This expansion yields a descent direction which can be used for shape optimization if one wishes to keep a classical domain throughout the optimization process. In this paper, such an expansion is obtained for the Poisson equation for a large class of cost functions and arbitrarily shaped holes. In the three-dimensional case, this expansion depends on the shape of the hole but not on its orientation if the cost function involves only the solution u to the underlying partial differential equation, whereas it may also depend on its orientation if the cost function involves the gradient ∇u. In contrast, the asymptotic expansion is independent of the shape in the two-dimensional case. A numerical example illustrates the use of the asymptotic expansion, which yields a minimizing sequence of classical domains in a case where no classical solution exists.

Key words. topological sensitivity, topological derivative, shape optimization, design sensitivity, Poisson's equation

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

PII. S0363012901384193

1. Introduction.

In many shape optimization problems, there is no "classical solution"; that is, a minimizing sequence of domains does not converge to a domain. In such cases, relaxed formulations or homogenization are often involved (see, e.g., [5, 6, 1, 25, 3, 18]), leading to the introduction of some "intermediate material" or micro-structures. The drawback is precisely that the optimal solution is not a classical design: it is a distribution of composite materials. Then penalization methods must be applied in order to retrieve a "feasible" shape. Hence, keeping a minimizing sequence of classical domains might be preferred (with a stopping criterion or additional constraints defined by the user). In that direction, global optimization techniques like genetic algorithms or simulated annealing have been proposed (see, e.g., [26]), but these methods have a high computational cost and can hardly be applied to industrial problems. Another approach, introduced by the works of Schumacher [27] and Sokołowski and Żochowski [28], is presented and analyzed in this paper in the case of the Poisson equation with Dirichlet boundary conditions for a large class of cost functions.

The shape optimization problem consists of minimizing a function $j(\Omega) := J(\Omega, u_\Omega)$ where the solution u_Ω to the Poisson equation is defined on a variable open and bounded subset Ω of $\mathbb{R}^n$. For $\varepsilon > 0$, let $\Omega_\varepsilon = \Omega \backslash (\overline{x_0 + \varepsilon\omega})$ be the subset obtained by removing a small part $\overline{x_0 + \varepsilon\omega}$ from Ω, where $x_0 \in \Omega$ and $\omega \subset \mathbb{R}^n$ is a fixed open and bounded subset containing the origin. Then, an asymptotic expansion of the function j is obtained in the following form:

$$j(\Omega_\varepsilon) = j(\Omega) + \rho(\varepsilon)\delta j(x_0) + o(\rho(\varepsilon)),$$
$$\lim_{\varepsilon \to 0} \rho(\varepsilon) = 0, \quad \rho(\varepsilon) > 0.$$

*Received by the editors January 26, 2001; accepted for publication (in revised form) April 1, 2002; published electronically October 8, 2002.
http://www.siam.org/journals/sicon/41-4/38419.html
†INSA, Département de Mathématiques, UMR MIP 5640, 135 Avenue de Rangueil, 31077 Toulouse Cedex 4, France (guillaum@gmm.insa-tlse.fr).

The "topological sensitivity" $\delta j(x_0)$ provides information for creating a small hole located at x_0: if $\delta j(x_0) < 0$, then $j(\Omega_\varepsilon) < j(\Omega)$ for small ε. For example, in the case of a circular hole and $n = 3$, and for a class of cost functions involving u_Ω but not ∇u_Ω, the first variation of the function j reads

$$j(\Omega_\varepsilon) = j(\Omega) + 4\pi\varepsilon\, u_\Omega(x_0)v_\Omega(x_0) + o(\varepsilon),$$

where v_Ω is the adjoint state. It is interesting to observe that the first order optimality condition $u_\Omega v_\Omega \geq 0$ given by Buttazzo and Dal Maso [5] follows straightforwardly from this expansion. More generally, the function δj can be used like a descent direction in an optimization process. The step length then consists of choosing the proportion of the domain which is "removed," located where $\delta j(x)$ is the most negative. The idea of this algorithm goes back to Céa, Gioan, and Michel [8] and was presented in the topological optimization context in [9]. For the sake of completeness, the principle of this algorithm is recalled in section 6. The main motivation of this approach is to provide an optimization method in which the iterated domains may have a varying topology but still remain classical domains, which may be required for feasibility when no mixture of materials is wanted. Of course, there is a drawback: there may be no solution to the optimization problem, and, in particular, the solution to the discretized problem may be mesh dependent. This is illustrated by an example in section 6, where a minimizing sequence of classical domains is obtained. Nevertheless, in contrast to classical shape optimization, which allows only moving the boundary (thus keeping fixed the topology, at least for small variations), the topological asymptotic expansion allows one to modify the topology of the domain during the optimization process, and can naturally be coupled with classical shape optimization.

A topological sensitivity framework using an adaptation of the adjoint method [7] and a truncation technique was introduced by Masmoudi [23] in the case of the Laplace equation with a circular hole. In the present paper, we analyze the case of the Poisson equation with noncircular holes. For this purpose, the technique used in [15, 16] for the elasticity equations with homogeneous Neumann conditions imposed on the boundary of the hole is adapted to Dirichlet boundary conditions, nonzero right-hand sides, and a large class of cost functions involving u_Ω or ∇u_Ω. In the three-dimensional case, it will be shown that the topological sensitivity $\delta j(x_0)$ depends on the shape of the hole but not on its orientation if the cost function involves only u_Ω, whereas it may also depend on its orientation if the cost function involves ∇u_Ω. In contrast, it will be shown that in the two-dimensional case the topological sensitivity is independent of the shape of the hole. Apart from the theoretical aspect, the consideration of noncircular holes may have some interesting applications, for example, in the study of cracks, which are beyond the scope of this paper.

First, the adaptation of the adjoint method to the topology shape optimization is recalled in section 2. Next, the formulation of the problem is presented in section 3. The truncation technique, which provides an efficient and general theoretical frame, is then applied to the problem in section 4. Section 5 presents the main results, whose proofs are reported in section 7. In the case of a circular hole, explicit expressions of the topological sensitivity are given for Dirichlet boundary conditions and for dimensions $n = 2$ or 3. Finally a numerical example in section 6 illustrates the use of the topological sensitivity in a shape optimization problem.

2. The generalized adjoint method. In this section, we recall the framework introduced in [23, 16] which extends the adjoint method [7] to the topology shape optimization. Let $\mathcal{V}$ be a Hilbert space. For $\varepsilon \geq 0$, let a_ε be a bilinear and symmetric

form on $\mathcal{V}$ and let l_ε be a linear form on $\mathcal{V}$ such that for all $\varepsilon \geq 0$,

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

where the constants $\alpha > 0$, $M_1 > 0$, and $M_2 \geq 0$ are independent of ε.

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

$$\|a_\varepsilon - a_0 - \rho(\varepsilon)\delta a\|_{\mathcal{L}_2(\mathcal{V})} = o(\rho(\varepsilon)), \tag{1}$$

$$\|l_\varepsilon - l_0 - \rho(\varepsilon)\delta l\|_{\mathcal{L}(\mathcal{V})} = o(\rho(\varepsilon)), \tag{2}$$

$$\lim_{\varepsilon \to 0} \rho(\varepsilon) = 0,$$

where $\mathcal{L}(\mathcal{V})$ (respectively, $\mathcal{L}_2(\mathcal{V})$) denotes the space of continuous and linear (respectively, bilinear) forms on $\mathcal{V}$. The same function ρ is used here for both asymptotic expansions (1)–(2). It does not exclude the case where $a_\varepsilon - a_0$ and $l_\varepsilon - l_0$ have different behaviors in $O(\rho_1(\varepsilon))$ and $O(\rho_2(\varepsilon))$, in which case ρ is chosen the "slowest" between ρ_1 and ρ_2, that is, $\rho_i(\varepsilon) = O(\rho(\varepsilon))$, $i = 1, 2$.

For $\varepsilon \geq 0$, let u_ε be the solution to the following problem: find $u_\varepsilon \in \mathcal{V}$ such that

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

LEMMA 2.1 (see [16]). *For $\varepsilon \geq 0$, this problem has a unique solution u_ε, and*

$$\|u_\varepsilon - u_0\| = O(\rho(\varepsilon)).$$

Consider now a function $j(\varepsilon) := J_\varepsilon(u_\varepsilon)$, where J_ε is defined on $\mathcal{V}$ for $\varepsilon \geq 0$, and J_0 is differentiable with respect to u, its derivative being denoted $DJ_0(u)$. Moreover, suppose that there exists a function δJ defined on $\mathcal{V}$ such that

$$J_\varepsilon(v) - J_0(u) = DJ_0(u)(v-u) + \rho(\varepsilon)\delta J(u) + o(\|v-u\| + \rho(\varepsilon)).$$

This expression looks like a first order (total) derivative and would be in fact the first order derivative of the function $\mathcal{J}(s,u)$ defined by $\mathcal{J}(s,u) := J_{\rho^{-1}(s)}(v) - J_0(u)$, with the change of variable $s = \rho(\varepsilon)$. Next, the Lagrangian L_ε is defined by

$$L_\varepsilon(u,v) = J_\varepsilon(u) + a_\varepsilon(u,v) - l_\varepsilon(v) \qquad \forall u, v \in \mathcal{V}.$$

Its variation with respect to ε is given by

$$\begin{aligned} \delta L(u,v) &= \delta J(u) + \delta a(u,v) - \delta l(v),\\ L_\varepsilon(u,v) - L_0(u,v) &= \rho(\varepsilon)\delta L(u,v) + o(\rho(\varepsilon)). \end{aligned}$$

THEOREM 2.2 (see [16]). *The function j has the following asymptotic expansion:*

$$j(\varepsilon) = j(0) + \rho(\varepsilon)\delta L(u_0, v_0) + o(\rho(\varepsilon)),$$

where v_0 is the solution to the following adjoint problem: find $v_0 \in \mathcal{V}$ such that

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

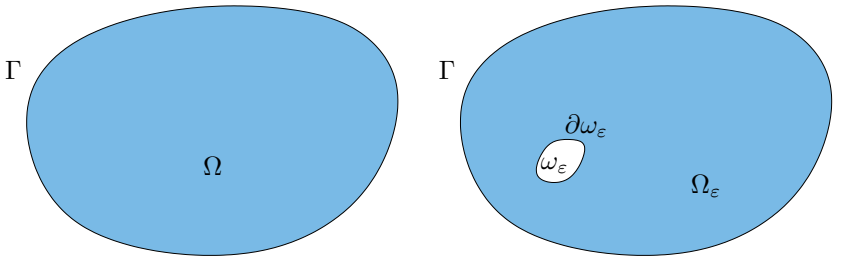


FIG. 1. *The initial domain shown before and after inclusion of the hole.*

3. Formulation of the problem. Let Ω be an open and bounded subset of $\mathbb{R}^n$ with boundary Γ, $n = 2$ or 3. The Poisson equation with homogeneous Dirichlet boundary conditions reads

$$\left\{ \begin{array}{rl} -\Delta u_\Omega = f & \text{in } \Omega, \\ u_\Omega = 0 & \text{on } \Gamma. \end{array} \right. \tag{3}$$

Some regularity is required from f, at least $f \in L^q(\Omega)$, $q > n/2$. (In fact, that will be needed only around x_0.) This equation has a unique solution in $H_0^1(\Omega)$, and due to the regularity of f, this solution is continuous in Ω [12]. The case of a nonhomogeneous boundary condition on Γ can be treated similarly.

For a given $x_0 \in \Omega$, consider the modified open subset $\Omega_\varepsilon = \Omega \backslash \overline{\omega_\varepsilon}$, $\omega_\varepsilon = x_0 + \varepsilon\omega$, where ω is a fixed open and bounded subset of $\mathbb{R}^n$ containing the origin ($\omega_\varepsilon = \emptyset$ if $\varepsilon = 0$), whose boundary $\partial\omega$ is connected and piecewise of class $\mathcal{C}^1$ (see Figure 1). The modified solution u_{Ω_ε} satisfies

$$\left\{ \begin{array}{rl} -\Delta u_{\Omega_\varepsilon} = f & \text{in } \Omega_\varepsilon, \\ u_{\Omega_\varepsilon} = 0 & \text{on } \Gamma, \\ u_{\Omega_\varepsilon} = 0 & \text{on } \partial\omega_\varepsilon. \end{array} \right. \tag{4}$$

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

In the context of identification of conductivity imperfections, the asymptotic behavior of the voltage potential $u_{\Omega_\varepsilon} - u_\Omega$ was studied in [13, 10] for the Laplace equation. Here we consider a right-hand side and a cost function which can be defined on Ω_ε. The adjoint technique is used for computing the variation of the cost function. It will be shown that the adjoint state is independent of the location of the hole.

The function u_{Ω_ε} is defined on the variable open subset Ω_ε, and thus it belongs to a functional space which depends on ε. Hence, if we want to derive the asymptotic expansion of a function of the form

$$j(\varepsilon) := \widetilde{J}_\varepsilon(u_{\Omega_\varepsilon}) \tag{5}$$

with $\widetilde{J}_\varepsilon$ being defined on $H^1(\Omega_\varepsilon)$ for $\varepsilon \geq 0$, we cannot apply directly the tools of section 2, which require a fixed functional space. In classical shape optimization, this requirement can be satisfied with the help of a domain parameterization technique [24, 22, 17]. This technique involves a fixed domain and a bi-Lipschitz map between this domain and the modified one. In the topology optimization context, such a map does not exist between Ω and Ω_ε. However, a functional space independent of ε can be constructed by using the following domain truncation technique. Let $R > 0$ be such that the closed ball $\overline{B}(x_0, R)$ is included in Ω. It is supposed throughout this

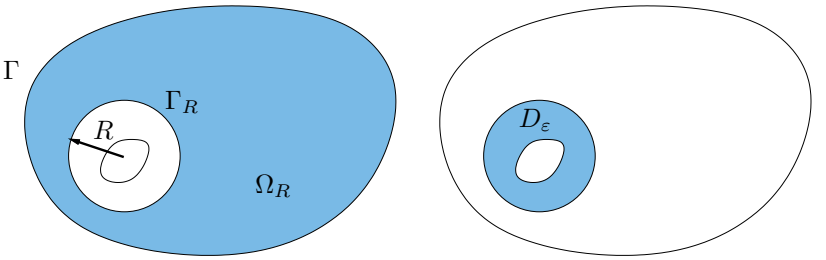


FIG. 2. *The truncated domain.*

paper that ε remains small enough so that $\overline{\omega_\varepsilon} \subset B(x_0, R)$. The truncated open subset is defined by (see Figure 2)

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

In section 4 the following are defined:

- a Hilbert space $\mathcal{V}_R$ independent of ε;
- a $\mathcal{V}_R$-elliptic bilinear and continuous form a_ε;
- a linear and continuous form l_ε

such that the solution u_ε to the equation

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

is equal to the restriction of u_{Ω_ε} to Ω_R. A bilinear form δa satisfying (1) and a linear form δl satisfying (2) will be obtained in section 7, from which the asymptotic expansion of the cost function will be derived by using the framework described in section 2. The main results and the particular case of a spherical hole are detailed in section 5.

A natural question is, Why do we need such a truncation technique? For $\varepsilon \geq 0$, one could simply define a_ε on $H_0^1(\Omega) \times H_0^1(\Omega)$ by

$$a_\varepsilon(u, v) = \int_{\Omega_\varepsilon} \nabla u . \nabla v \, dx$$

and work with functions of $H_0^1(\Omega_\varepsilon)$ extended by 0 on ω_ε. The main difficulty with this approach is that we cannot apply the Lagrangian technique described in section 2, because there is no bilinear and continuous form δa such that $\|a_\varepsilon - a_0 - \rho(\varepsilon)\delta a\|_{\mathcal{L}_2(H_0^1(\Omega))} = o(\rho(\varepsilon))$ for some adequate positive function ρ. Indeed, we have

$$a_\varepsilon(u, v) - a_0(u, v) = -\int_{\omega_\varepsilon} \nabla u . \nabla v \, dx,$$

and for smooth functions u, v, and $n = 3$ (for example) we have

$$a_\varepsilon(u, v) - a_0(u, v) = -\varepsilon^3 \nabla u(x_0) . \nabla v(x_0) \int_\omega dx + o(\varepsilon^3).$$

But $\delta a(u, v) := \nabla u(x_0) . \nabla v(x_0)$ cannot be continuously extended on $H_0^1(\Omega) \times H_0^1(\Omega)$. Besides, if u_{Ω_ε} is extended by 0 on ω_ε, the behavior of $\|u_{\Omega_\varepsilon} - u_\Omega\|_{H^1(\Omega)}$ is not of order ε^3 but only of order $\varepsilon^{1/2}$ (see Lemma 7.3). This change of order comes from the lack

of continuity of the above bilinear form δa. In contrast, the bilinear form a_ε defined in the next section will be associated to a bilinear and *continuous* form δa which will satisfy $a_\varepsilon - a_0 = \varepsilon\delta a + o(\varepsilon)$ (see Proposition 7.6), and the associated solution u_ε will yield the same order: $\|u_\varepsilon - u_0\|_{H^1(\Omega_R)} = O(\varepsilon)$ (consequence of Lemma 2.1). Moreover, it will be seen that $\delta a(u,v)$ involves $u(x_0)$ and $v(x_0)$, and not $\nabla u(x_0)$ and $\nabla v(x_0)$. Another point is that the truncation technique can be applied to the case of a Neumann boundary condition on the hole [16], or even to more general boundary conditions.

4. The truncated problem. The open subset $B(x_0,R)\setminus\overline{\omega_\varepsilon}$ is denoted by D_ε (see Figure 2). For $\varphi\in H^{1/2}(\Gamma_R)$ and $\varepsilon>0$, let $u_\varepsilon^{f,\varphi}\in H^1(D_\varepsilon)$ be the solution to the following problem: find $u_\varepsilon^{f,\varphi}$ such that

$$\left\{\begin{array}{rl} -\Delta u_\varepsilon^{f,\varphi} = f & \text{in } D_\varepsilon,\\ u_\varepsilon^{f,\varphi} = \varphi & \text{on } \Gamma_R,\\ u_\varepsilon^{f,\varphi} = 0 & \text{on } \partial\omega_\varepsilon,\end{array}\right. \tag{6}$$

where Γ_R is the boundary of the ball $B(x_0,R)$. For $\varepsilon=0$, $u_0^{f,\varphi}$ is the solution to

$$\left\{\begin{array}{rl} -\Delta u_0^{f,\varphi} = f & \text{in } B(x_0,R),\\ u_0^{f,\varphi} = \varphi & \text{on } \Gamma_R.\end{array}\right. \tag{7}$$

Clearly we have

$$u_\varepsilon^{f,\varphi} = u_\varepsilon^{f,0} + u_\varepsilon^{0,\varphi}. \tag{8}$$

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

$$\begin{array}{rcl} T_\varepsilon : H^{1/2}(\Gamma_R) & \longrightarrow & H^{-1/2}(\Gamma_R),\\ \varphi & \longmapsto & T_\varepsilon\varphi = \nabla u_\varepsilon^{0,\varphi}.\mathbf{n},\end{array}$$

and the function $f_\varepsilon\in H^{-1/2}(\Gamma_R)$ is defined by

$$f_\varepsilon = -\nabla u_\varepsilon^{f,0}.\mathbf{n},$$

where the normal $\mathbf{n}$ is chosen outward to D_ε on Γ_R and $\partial\omega_\varepsilon$. Thus we have

$$\nabla u_\varepsilon^{f,\varphi}.\mathbf{n} = T_\varepsilon\varphi - f_\varepsilon.$$

Finally, we define for $\varepsilon\geq 0$ the solution u_ε to the truncated problem

$$\left\{\begin{array}{rl} -\Delta u_\varepsilon = f & \text{in } \Omega_R,\\ u_\varepsilon = 0 & \text{on } \Gamma,\\ -\nabla u_\varepsilon.\mathbf{n} + T_\varepsilon u_\varepsilon = f_\varepsilon & \text{on } \Gamma_R.\end{array}\right. \tag{9}$$

The variational formulation associated to (9) is the following: find $u_\varepsilon\in\mathcal{V}_R$ such that

$$a_\varepsilon(u_\varepsilon,v) = l_\varepsilon(v) \qquad \forall v\in\mathcal{V}_R, \tag{10}$$

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

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

$$a_\varepsilon(u,v) = \int_{\Omega_R}\nabla u.\nabla v\,dx + \int_{\Gamma_R} T_\varepsilon uv\,d\gamma(x), \tag{11}$$

$$l_\varepsilon(v) = \int_{\Omega_R} fv\,dx + \int_{\Gamma_R} f_\varepsilon v\,d\gamma(x). \tag{12}$$

Here $x.y$ denotes the usual dot product of $\mathbb{R}^n$ and $d\gamma(x)$ is the Lebesgue measure on the boundary. Symmetry, continuity, and coercivity of a_ε, and continuity of l_ε follow directly from

$$\int_{\Gamma_R} T_\varepsilon \varphi\psi \, d\gamma(x) = \int_{D_\varepsilon} \nabla u_\varepsilon^{0,\varphi}.\nabla u_\varepsilon^{0,\psi} \, dx,$$
$$\int_{\Gamma_R} f_\varepsilon \psi \, d\gamma(x) = \int_{D_\varepsilon} f u_\varepsilon^{0,\psi} \, dx.$$

Notice that $\int_{D_\varepsilon} \nabla u_\varepsilon^{f,0}.\nabla u_\varepsilon^{0,\psi} \, dx = 0$. The proof of the following result is standard.

PROPOSITION 4.1. *Let $\varepsilon \geq 0$. Problems* (4) *and* (9) *have a unique solution. Moreover, the restriction to Ω_R of the solution u_{Ω_ε} to* (4) *is the solution u_ε to* (9)*, and on D_ε we have*

$$(u_{\Omega_\varepsilon})_{|D_\varepsilon} = u_\varepsilon^{f,0} + u_\varepsilon^{0,\varphi}, \tag{13}$$

where φ is the trace of u_ε on Γ_R.

We now have at our disposal the fixed Hilbert space $\mathcal{V}_R$ required by section 2. The cost function (5) can be redefined in the following way: for $u \in \mathcal{V}_R$, let $\widetilde{u}_\varepsilon \in H^1(\Omega_\varepsilon)$ be the extension of u which coincides with u on Ω_R and with $u_\varepsilon^{f,\varphi}$ on D_ε for $\varphi = u_{|\Gamma_R}$. Then a function J_ε can be defined on $\mathcal{V}_R$ by

$$J_\varepsilon(u) := \widetilde{J}_\varepsilon(\widetilde{u}_\varepsilon). \tag{14}$$

Particularly, it follows from the previous proposition that

$$j(\varepsilon) = \widetilde{J}_\varepsilon(u_{\Omega_\varepsilon}) = J_\varepsilon(u_\varepsilon). \tag{15}$$

Notice that $J_\varepsilon(u_\varepsilon)$ is independent of the choice of R. For example, for a given target function u_{d} defined on Ω, if

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

then we have

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

5. Asymptotic expansion of the cost function. This section contains the main results of this paper. All the proofs are reported in section 7. Henceforth we have to distinguish the cases $n = 2$ and $n = 3$. This is due to the fact that the fundamental solutions to the Laplace equation in $\mathbb{R}^2$ and $\mathbb{R}^3$ have an essentially different asymptotic expansion at infinity, and Problem (16) has generally no solution if $n = 2$.

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

$$\begin{cases} -\Delta v_\omega = 0 & \text{in } \mathbb{R}^3 \backslash \overline{\omega}, \\ v_\omega = 0 & \text{at } \infty, \\ v_\omega = u_\Omega(x_0) & \text{on } \partial\omega, \end{cases} \tag{16}$$

where u_Ω is the solution to (3). Recall that $f \in L^q(\Omega)$, $q > n/2$, so that u_Ω is continuous inside Ω and the above boundary condition is well defined. The function v_ω can be expressed by a single layer potential on $\partial\omega$. Let

$$E(y) = \frac{1}{4\pi r} \tag{17}$$

with $r = \|y\|$. It is a fundamental solution for the Laplace equation in $\mathbb{R}^3$. Then the function v_ω reads

$$v_\omega(y) = \int_{\partial\omega} E(y-x) p_\omega(x)\, d\gamma(x), \qquad y \in \mathbb{R}^3 \backslash \overline{\omega}, \tag{18}$$

where $p_\omega \in H^{-1/2}(\partial\omega)$ is the solution to boundary integral equation [12]

$$\int_{\partial\omega} E(y-x) p_\omega(x)\, d\gamma(x) = u_\Omega(x_0) \qquad \forall y \in \partial\omega.$$

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

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

and the asymptotic expansion at infinity of the function v_ω is given by

$$v_\omega(y) = P_\omega(y) + W_\omega(y), \tag{19}$$

$$P_\omega(y) = A_\omega(u_\Omega(x_0)) E(y), \tag{20}$$

$$A_\omega(u_\Omega(x_0)) = \int_{\partial\omega} p_\omega(x)\, d\gamma(x), \tag{21}$$

$$W_\omega(y) = O\left(\frac{1}{r^2}\right).$$

Notice that $P_\omega \in L^m_{\mathrm{loc}}(\mathbb{R}^3)$ for all $m < 3$. Clearly, the function $\alpha \longmapsto A_\omega(\alpha)$ is linear on $\mathbb{R}$, and the number $A_\omega(\alpha)$ depends on the shape of ω. For example, if ω is changed in $k\omega$, $k > 0$, then $v_{k\omega}(ky) = v_\omega(y)$ in (16), and it follows from (18) that $k p_{k\omega}(kx) = p_\omega(x)$ for $x \in \partial\omega$. Then using (21) we obtain $A_{k\omega}(u_\Omega(x_0)) = k A_\omega(u_\Omega(x_0))$. However, it is interesting to observe that $A_\omega(\alpha)$ is independent of the orientation of the hole ω: if R is a rotation, one obtains in a similar way $A_{R\omega}(u_\Omega(x_0)) = A_\omega(u_\Omega(x_0))$. Next we consider the constant $Q_\omega \in \mathbb{R}$ defined by

$$Q_\omega = \frac{A_\omega(u_\Omega(x_0))}{4\pi R} = (P_\omega)_{|\Gamma_R}. \tag{22}$$

The main result is the following, which will be proved in section 7. It is based on the fact that

$$\varepsilon(Q_\omega - P_\omega)_{|D_\varepsilon} \tag{23}$$

is the *first order approximation* of $(u_\varepsilon^{f,\varphi} - u_0^{f,\varphi})_{|D_\varepsilon}$ with $\varphi = (u_\Omega)_{|\Gamma_R}$, in a sense which will be stated precisely in section 7. Observe that it depends on the shape of ω through the term $A_\omega(u_\Omega(x_0))$ involved in (20). The stronger hypothesis $f \in L^q(\Omega)$, $q > n$, is used in the study of the variation of the linear form l_ε (12); cf. Proposition 7.7, which

involves the $\mathcal{C}^1$ norm of u_0 around x_0. If l_ε does not depend on ε (which happens, for example, if f vanishes on D_0), then $f \in L^q(\Omega)$, $q > n/2$, is sufficient.

THEOREM 5.1. *Let $f \in L^q(\Omega)$, $q > n$, and let J_ε be a function defined on $\mathcal{V}_R$ for all $\varepsilon \geq 0$. Suppose that for all $v \in \mathcal{V}_R$ and $\varepsilon > 0$, one has*

$$J_\varepsilon(v) - J_0(u_0) = DJ_0(u_0)(v - u_0) + \varepsilon\,\delta J(u_0) + o(\varepsilon + \|v - u_0\|_{\mathcal{V}_R}), \tag{24}$$

where $DJ_0(u_0)$ is linear and continuous on $\mathcal{V}_R$, and u_ε, $\varepsilon \geq 0$, is the solution to (10). *Let $v_0 \in \mathcal{V}_R$ be the solution to the adjoint equation*

$$a_0(w, v_0) = -DJ_0(u_0)w \qquad \forall w \in \mathcal{V}_R. \tag{25}$$

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

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

with

(26)

$$\delta j(x_0) = -\int_{\Gamma_R} \nabla P_\omega.\mathbf{n}\, v_0\, d\gamma(x) + \delta J(u_0) = \frac{A_\omega(u_\Omega(x_0))}{4\pi R^2} \int_{\Gamma_R} v_0(x)\, d\gamma(x) + \delta J(u_0).$$

The function $\delta j(x_0)$ is called the *topological sensitivity* or the *topological gradient*. Moreover, as j is usually independent of R (at least when it is of the form (15), which is the "natural" way of posing the problem) and $\delta j(x_0)$ is independent of ε, it follows from the uniqueness of an asymptotic expansion that $\delta j(x_0)$ is also independent of R. This is not necessarily true for the terms $\delta a(u_0, v_0)$, $\delta l(v_0)$ (see section 7), or $\delta J(u_0)$ considered separately, because a, l, and J do depend on R.

Practically, what is computed is the solution u_Ω to (3) and the solution v_Ω to

$$\int_\Omega \nabla w.\nabla v_\Omega\, dx = -D\widetilde{J}_0(u_\Omega)w \qquad \forall w \in H_0^1(\Omega). \tag{27}$$

As observed in Proposition 4.1, u_0 is the restriction to Ω_R of u_Ω. The same property holds for v_0 and v_Ω. This can easily be seen by observing that for $w \in H_0^1(\Omega)$ such that $\Delta w = 0$ in D_0, and denoting by v_R and w_R the restrictions of v_Ω and w to Ω_R, on the one hand we have

$$\begin{aligned} a_0(w_R, v_R) &= \int_{\Omega_R} \nabla w_R.\nabla v_R\, dx + \int_{\Gamma_R} T_0 w_R v_R\, d\gamma(x) \\ &= \int_\Omega \nabla w.\nabla v_\Omega\, dx, \end{aligned} \tag{28}$$

and on the other hand, due to (14), we have $\widetilde{J}_0(u) = J_0(u_R)$ for all $u \in H_0^1(\Omega)$ such that $-\Delta u = f$ in D_0 (with $u_R = u_{|\Omega_R}$); hence

$$D\widetilde{J}_0(u_\Omega)w = DJ_0(u_0)w_R. \tag{29}$$

Then, gathering (28), (27), and (29), we obtain

$$a_0(w_R, v_R) = -DJ_0(u_0)w_R \qquad \forall w_R \in \mathcal{V}_R,$$

which proves that v_R is the solution to (25), that is, v_0 is the restriction to Ω_R of v_Ω. The basic property of an adjoint technique is here also satisfied, in that the function

u_Ω and the adjoint state v_Ω do not depend on x_0. Hence *only two systems have to be solved* in order to compute the topological sensitivity $\delta j(x)$ for all $x \in \Omega$.

Thanks to Green's formula, $P_\omega = Q_\omega$ on Γ_R and $\nabla Q_\omega = 0$, the integral in (26) also reads

$$-\int_{\Gamma_R} \nabla P_\omega.\mathbf{n}\, v_0\, d\gamma(x) = \int_{\Gamma_R} \nabla v_\Omega.\mathbf{n}\, P_\omega\, d\gamma(x) - \nabla P_\omega.\mathbf{n}\, v_\Omega\, d\gamma(x) - \int_{D_0} \Delta v_\Omega Q_\omega\, dx.$$

When v_Ω is smooth enough, it follows from Newton's potential theory [12] that

$$v_\Omega(x_0) = \int_{\Gamma_R} \nabla v_\Omega.\mathbf{n}\, E\, d\gamma(x) - \nabla E.\mathbf{n}\, v_\Omega\, d\gamma(x) - \int_{D_0} \Delta v_\Omega E\, dx.$$

Multiplying by $A_\omega(u_\Omega(x_0))$ and using $P_\omega = A_\omega(u_\Omega(x_0))E$ yields

$$\int_{\Gamma_R} \nabla v_\Omega.\mathbf{n}\, P_\omega\, d\gamma(x) - \nabla P_\omega.\mathbf{n}\, v_\Omega\, d\gamma(x) = \int_{D_0} \Delta v_\Omega P_\omega\, dx + A_\omega(u_\Omega(x_0))v_\Omega(x_0).$$

Then using (26) leads to the following result.

COROLLARY 5.2. *Under the assumptions of Theorem* 5.1, *if* $\Delta v_\Omega \in L^q(D_0)$, $q > n/2$, *then*

$$\delta j(x_0) = A_\omega(u_\Omega(x_0))v_\Omega(x_0) + \int_{D_0} \Delta v_\Omega(P_\omega - Q_\omega)\, dx + \delta J(u_0). \tag{30}$$

Proposition 5.3 will show that, in fact, the two last terms in the right-hand side of (30) cancel each other for a class of cost functions which do not involve ∇u_Ω. In that case, the dependence on the shape of ω occurs only through the term $A_\omega(u_\Omega(x_0))$ (21). For example, if ω is changed in $k\omega$, $k > 0$, we have observed previously that $A_{k\omega}(u_\Omega(x_0)) = kA_\omega(u_\Omega(x_0))$. Hence ε is changed in $k\varepsilon$, which is not surprising. We have also observed that $A_\omega(u_\Omega(x_0))$ was independent of the orientation of ω; hence it remains true for $\delta j(x_0)$. However, $\delta j(x_0)$ can depend on the orientation of ω if the cost function involves ∇u_Ω; see Proposition 5.4. What will be more surprising is that in the two-dimensional case, $\delta j(x_0)$ is independent of ω (size, shape, orientation); see Propositions 5.5 and 5.6.

When ω is the unit ball $B(0,1)$, then $v_\omega(y)$, $P_\omega(y)$, $W(y)$, and Q_ω can be computed explicitly:

$$v_\omega(y) = \frac{u_\Omega(x_0)}{r} = P_\omega(y), \qquad W(y) = 0, \quad 0 \neq y \in \mathbb{R}^3,$$
$$Q_\omega = u_\Omega(x_0).$$

Then it follows from (17) and (20) that

$$A_\omega(u_\Omega(x_0)) = 4\pi u_\Omega(x_0).$$

It can also be easily checked that

$$p_\omega(y) = u_\Omega(x_0) \qquad \forall y \in \partial\omega.$$

We examine now two particular cases of cost functions.

5.1.1. First example. The first example consists of functions of the form

$$\widetilde{J}_\varepsilon(u) = \int_{\Omega_\varepsilon} g(x, u(x))\,dx, \qquad u \in H^1(\Omega_\varepsilon). \tag{31}$$

The hypotheses on g are the following:

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

$$|g_s(x,t) - g_s(x,s)| \le M\,|t-s| \qquad \forall (x,s,t) \in \Omega \times \mathbb{R} \times \mathbb{R}; \tag{32}$$

- the function $x \mapsto g_s(x,0)$ belongs to $L^2(\Omega)$ and $x \mapsto g(x,0)$ belongs to $L^{3/2}(\Omega)$ (or to $L^p(\Omega)$, $p > 1$, if $n = 2$; cf. section 5.2).

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

$$|g(x,s)| \le |g(x,0)| + |g_s(x,0)s| + \frac{M}{2}s^2, \tag{33}$$

$$|g_s(x,s)| \le |g_s(x,0)| + M\,|s|\,, \tag{34}$$

and the functions $x \mapsto g(x,u(x))$ and $x \mapsto g_s(x,u(x))^2$ are integrable on Ω for all $u \in L^2(\Omega)$. If $u \in L^6(\mathcal{O})$ (or $L^m(\mathcal{O})$, $m > 2$, if $n = 2$), then the function $x \mapsto g(x,u(x))$ belongs to $L^{3/2}(\mathcal{O})$ (or to $L^{p'}(\mathcal{O})$, $1 < p' = \min(p, 2m/(m+2), m/2)$, if $n = 2$). The usual example

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

satisfies these hypotheses if u_{d} belongs to $L^3(\Omega)$ (or to $L^{2p}(\Omega)$ if $n = 2$).

REMARK 5.1. *These assumptions are standard in shape optimization (see, for example,* [6]*), with the difference that $x \mapsto g(x,0)$ is usually supposed to be in $L^1(\Omega)$ only. Equation* (60) *is the only place where $g(.,0) \in L^{3/2}(\Omega)$ is used. That comes from the choice made on the function $\widetilde{J}_\varepsilon$. When $\widetilde{J}_\varepsilon$ is of the form*

$$\widetilde{J}_\varepsilon(u) = \int_{\Omega} g(x, u(x))\,dx, \quad u \in H^1(\Omega_\varepsilon) \text{ extended by } 0 \text{ on } \omega_\varepsilon,$$

then the term (60) *disappears, $g(.,0) \in L^1(\Omega)$ is sufficient, and the result remains the same.*

PROPOSITION 5.3. *If these hypotheses are satisfied and if $f \in L^q(\Omega)$, $q > n$, then*

$$\delta J(u_0) = \int_{D_0} g_s(x, u_\Omega)(Q_\omega - P_\omega)\,dx,$$
$$\Delta v_\Omega = g_s(x, u_\Omega)$$

with $v_\Omega \in H_0^1(\Omega)$, and the function j has the following asymptotic expansion:

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

If ω is the unit ball $B(0,1)$, then

$$j(\varepsilon) = j(0) + 4\pi\varepsilon\, u_\Omega(x_0)v_\Omega(x_0) + o(\varepsilon).$$

5.1.2. Second example. The second example consists of functions of the form

$$\widetilde{J}_\varepsilon(u) = \frac{1}{2}\int_{\Omega_\varepsilon} B(x)\nabla(u-u_{\rm d}).\nabla(u-u_{\rm d})\,dx, \qquad u \in H^1(\Omega_\varepsilon), \tag{35}$$

where $B \in W^{1,\infty}(\Omega, \mathbb{R}^{3\times 3})$, $B(x)$ is a symmetric matrix for all $x \in \Omega$, and $u_{\rm d} \in H^1(\Omega)$. Here tr $B(x)$ denotes the trace of the matrix $B(x)$, div denotes the divergence operator, and $W^{m,\infty}(\Omega)$ is the Sobolev space of distributions whose derivatives up to the order m are in $L^\infty(\Omega)$.

PROPOSITION 5.4. *If these hypotheses are satisfied and if f and $\Delta u_{\rm d}$ belong to $L^q(\Omega)$, $q > n$, then*

$$\delta J(u_0) = \frac{1}{2}\int_{\mathbb{R}^3\setminus\overline{\omega}} B(x_0)\nabla v_\omega(y).\nabla v_\omega(y)\,dy + \int_{D_0} \Delta v_\Omega(Q_\omega - P_\omega)\,dx,$$
$$-\Delta v_\Omega = \operatorname{div}\left(B(\nabla u_\Omega - \nabla u_{\rm d})\right)$$

with $v_\Omega \in H^1_0(\Omega)$, and the function j has the following asymptotic expansion:

$$j(\varepsilon) = j(0) + \varepsilon A_\omega(u_\Omega(x_0)).v_\Omega(x_0) + \frac{\varepsilon}{2}\int_{\mathbb{R}^3\setminus\overline{\omega}} B(x_0)\nabla v_\omega(y).\nabla v_\omega(y)\,dy + o(\varepsilon). \tag{36}$$

If ω is the unit ball, then

$$v_\omega = \frac{u_\Omega(x_0)}{r}$$

and the integral can be computed explicitly:

$$j(\varepsilon) = j(0) + \varepsilon\left(4\pi u_\Omega(x_0)v_\Omega(x_0) + \frac{2\pi u_\Omega(x_0)^2}{3}\operatorname{tr} B(x_0)\right) + o(\varepsilon).$$

Here, due to the form of the integral in (36) and the definition of v_ω, one can observe that the topological sensitivity will usually depend on the orientation of the hole ω, unless, for example, the matrix $B(x_0)$ is proportional to the identity.

5.2. The two-dimensional case. We briefly describe the transposition of the previous results to the two-dimensional case. As before, u_Ω and the adjoint state v_Ω are, respectively, the solutions to (3) and (27). A fundamental solution for the Laplace equation in $\mathbb{R}^2$ is given by

$$E(y) = \frac{-1}{2\pi}\log r.$$

The exterior problem must now be defined differently than in (16). Let v_ω be the solution to

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

The function v_ω has the form

$$v_\omega(y) = u_\Omega(x_0)\log\|y\| + P_\omega + W_\omega(y),$$

where P_ω is constant and $W_\omega(y) = o(1)$ at infinity [12]. The *first order approximation* of $(u_\varepsilon^{f,\varphi} - u_0^{f,\varphi})_{|D_\varepsilon}$ with $\varphi = (u_\Omega)_{|\Gamma_R}$ is now (compare with (23))

$$\frac{1}{\log(R/\varepsilon)}\left(u_\Omega(x_0)\left(\log\frac{\|x\|}{\varepsilon} - \log\frac{R}{\varepsilon}\right)\right)_{|D_\varepsilon} \simeq -\frac{u_\Omega(x_0)\log(\|x\|/R)_{|D_\varepsilon}}{\log\varepsilon}.$$

In the following propositions (where ω is not supposed to be a ball), one can observe that in the two-dimensional case the topological sensitivity does not depend on the shape of the hole ω, in contrast to the three-dimensional case.

PROPOSITION 5.5. *The assumptions are the same as in Proposition* 5.3, *with* $\widetilde{J}_\varepsilon$ *of the form*

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

Then the function j *has the following asymptotic expansion:*

$$j(\varepsilon) = j(0) - \frac{2\pi u_\Omega(x_0)v_\Omega(x_0)}{\log\varepsilon} + o\left(\frac{-1}{\log\varepsilon}\right).$$

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

PROPOSITION 5.6. *The assumptions are the same as in Proposition* 5.4, *with* $\widetilde{J}_\varepsilon$ *of the form*

$$\widetilde{J}_\varepsilon(u) = \frac{1}{2}\int_{\Omega_\varepsilon} B(x)\nabla(u - u_{\mathrm{d}}).\nabla(u - u_{\mathrm{d}})\,dx, \qquad u \in H^1(\Omega_\varepsilon).$$

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

$$\begin{aligned} j(\varepsilon) &= j(0) - \frac{2\pi}{\log\varepsilon}u_\Omega(x_0)v_\Omega(x_0) + \frac{1}{2\log^2\varepsilon}\int_{D_\varepsilon/\varepsilon} B(x_0)\nabla v_\omega(y).\nabla v_\omega(y)\,dy + o\left(\frac{-1}{\log\varepsilon}\right) \\ &= j(0) - \frac{1}{\log\varepsilon}\left(2\pi u_\Omega(x_0)v_\Omega(x_0) + \frac{\pi}{2}u_\Omega(x_0)^2\operatorname{tr} B(x_0)\right) + o\left(\frac{-1}{\log\varepsilon}\right). \end{aligned}$$

The proofs use the same tools as for the three-dimensional case (see section 7) and will not be repeated for the two-dimensional case.

6. A numerical example. We illustrate the use of the asymptotic expansion given by Proposition 5.3 on an example taken from [6], to which we refer the reader for more details on its construction. It consists of minimizing

$$j(\Omega) = \int_{B(0,1)} (u_\Omega - u_d)^2\,dx,$$

where the solution $u_\Omega \in H_0^1(\Omega)$ to

$$-\Delta u_\Omega = 1 \quad \text{in } \Omega \subset \overline{B}(0,1) \subset \mathbb{R}^2 \tag{37}$$

is extended by 0 on $\overline{B}(0,1)\backslash\overline{\Omega}$ (see, however, Remark 5.1) and

$$u_d(x) = \begin{cases} \frac{r_1^2 - r^2}{4} + a & \text{if } r \le r_1, \\ a & \text{if } r_1 \le r \le 1. \end{cases}$$

Here $0 < a < 3/16$, $r = \|x\|$, and r_1 is the first minimum of the function

$$q(r_1) = 2\pi \int_{r_1}^{1} \left(\left(a - \frac{1 - r_1^2}{4} \right) \frac{\log r}{\log r_1} + \frac{1 - r^2}{4} - a \right)^2 r\, dr.$$

We use the value $a = 1/19$, which gives $r_1 \simeq 0.503$. This problem has no classical solution. We seek a minimizing sequence of classical solutions, when obtaining a classical approximate solution is a constraint imposed on the optimization process.

The relaxed formulation reads

$$\min_{\mu} \int_{B(0,1)} (u_\mu - u_d)^2\, dx, \tag{38}$$

where $u_\mu \in H_0^1(B(0,1))$ is the solution to

$$-\Delta u_\mu + \mu u_\mu = 1 \quad \text{in } B(0,1)$$

and μ is a nonnegative Borel measure on $B(0,1)$ which vanishes on all sets of capacity zero. For this example, the solution to (38) is known: it is given by

$$\mu = \gamma H^1_{\partial B(0,r_1)}, \quad \gamma = \frac{4a - 1 + r_1^2}{4 a r_1 \log r_1},$$

$$u_\mu(x) = \begin{cases} u_d(x) & \text{if } r \le r_1, \\ \left(a - \frac{1 - r_1^2}{4} \right) \frac{\log r}{\log r_1} + \frac{1 - r^2}{4} & \text{if } r_1 \le r \le 1, \end{cases}$$

where $H^1_{\partial B(0,r_1)}$ denotes the Hausdorff measure on the circle $\partial B(0, r_1)$. The minimum of the (relaxed) cost function is

$$\int_{B(0,1)} (u_\mu - u_d)^2\, dx = q(r_1) \simeq 1.12\, 10^{-3}. \tag{39}$$

A minimizing sequence $(\Omega_N)_{N \ge 1}$ for the optimal design problem (38) is given by ($\mathbb{R}^2$ is identified with the complex plane)

$$\Omega_N = B(0,1) \backslash \bigcup_{k=1}^{N} \left\{ z \in \mathbb{C}\,;\, \left| z - \mathrm{e}^{2i\pi k/N} \right| \le \mathrm{e}^{-N/\gamma r_1} \right\}. \tag{40}$$

Using the topological asymptotic expansion of j in a way similar to that described in [16], we retrieve the above minimizing sequence in less than 20 iterations for different values of the mesh size h. One can observe in Figure 3 that the number of "holes" is approximately proportional to $\log h$. This agrees with the fact that if the size of the holes in (40) is set to $h = \mathrm{e}^{-N/\gamma r_1}$, then $N = -\gamma r_1 \log h$. The cost function at each step is illustrated by Figure 4. The obtained minimum is not far from the exact (relaxed) minimum (39).

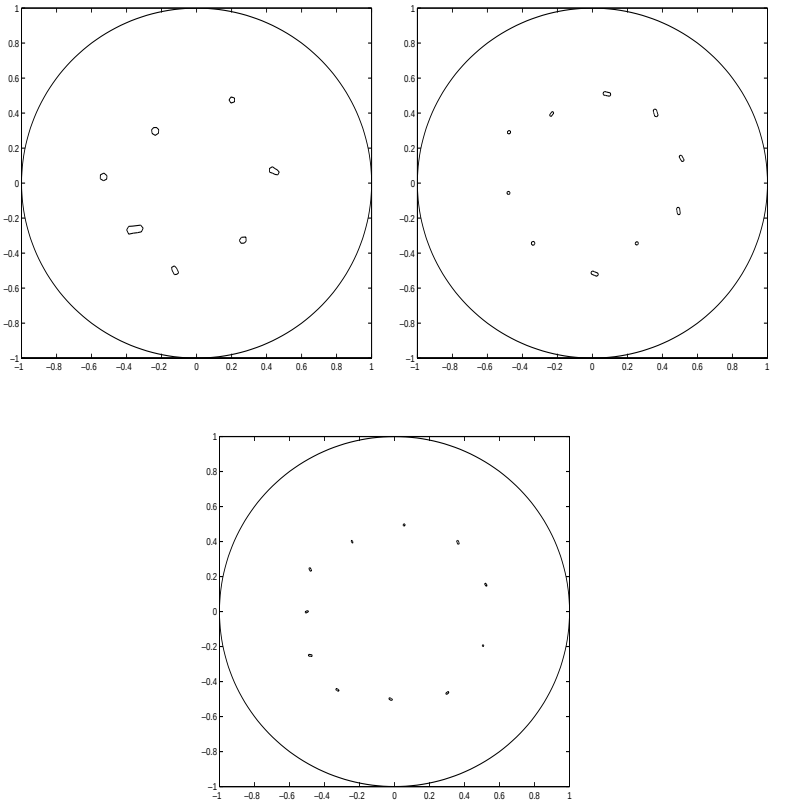


FIG. 3. *Solution for* $h = .05$, $h = .025$ *(top) and* $h = .0125$ *(bottom).*

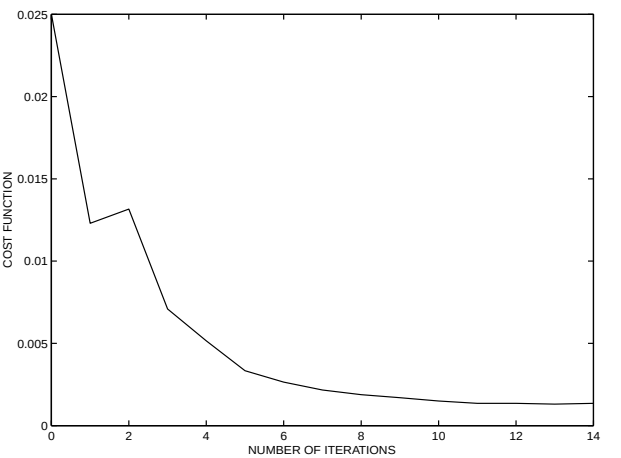


FIG. 4. $j(\Omega_k)$ *for* $h = 0.025$.

For completeness, we recall here the topology optimization algorithm [16]. Let $(m_k)_{k\geq 0}$ be a decreasing sequence of volume constraints, with $m_0 = \text{meas}(B(0,1))$. For example, a geometrical sequence may be chosen. At the kth iteration, the topological sensitivity is denoted by $\delta j_k(x)$. The algorithm is as follows:

- Initialization: chose $\Omega_0 = B(0,1)$, and set $k = 0$.
- Repeat until target is reached:
  1. solve (37) in Ω_k,
  2. compute the topological sensitivity δj_k,
  3. set $\Omega_{k+1} = \{x \in \Omega_k,\ \delta j_k(x) \geq c_{k+1}\}$, where c_{k+1} is chosen such that $\text{meas}(\Omega_{k+1}) = m_{k+1}$,
  4. $k \leftarrow k+1$.

This algorithm can be seen as a descent method where the descent direction is determined by the topological sensitivity δj_k, and the step length is given by the volume variation $m_{k+1} - m_k$. One possible stopping criterion is when no more improvement can be done, or simply when the optimality condition $\delta j_k(x) \geq 0$ for all $x \in \Omega_k$ is satisfied.

7. Proofs. This section consists of the proofs of Theorem 5.1 and Propositions 5.3 and 5.4. The variations of the bilinear form a_ε and the linear form l_ε (see (11) and (12)) read

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

Hence, the problem reduces to the analysis of $(T_\varepsilon - T_0)\varphi$ for $\varphi \in H^{1/2}(\Gamma_R)$ and of $f_\varepsilon - f_0$ in $H^{-1/2}(\Gamma_R)$. More precisely, it will be shown in sections 7.3 and 7.4 that there exist an operator $\delta T \in \mathcal{L}(H^{1/2}(\Gamma_R); H^{-1/2}(\Gamma_R))$ and a function $\delta f \in H^{-1/2}(\Gamma_R)$ such that

$$\|T_\varepsilon - T_0 - \varepsilon\delta T\|_{\mathcal{L}(H^{1/2}(\Gamma_R);\, H^{-1/2}(\Gamma_R))} = O(\varepsilon^2), \tag{41}$$
$$\|f_\varepsilon - f_0 - \varepsilon\delta f\|_{H^{-1/2}(\Gamma_R)} = O(\varepsilon^2). \tag{42}$$

Consequently, defining δa and δl by

$$\delta a(u,v) = \int_{\Gamma_R} \delta T uv\, d\gamma(x), \qquad u,\, v \in \mathcal{V}_R,$$
$$\delta l(v) = \int_{\Gamma_R} \delta f\, v\, d\gamma(x), \qquad v \in \mathcal{V}_R,$$

will yield straightforwardly

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

In order to derive (41)–(42), we need some definitions and preliminary lemmas.

7.1. Definitions. For convenience, the following norms and seminorms are chosen for the functional spaces which will be used.

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

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

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

$$|u|^2_{k,\mathcal{O}} := \sum_{|\alpha|=k} \int_{\mathcal{O}} |\partial_\alpha u|^2\, dx. \tag{43}$$

- For a given $\varepsilon > 0$, the space $H^{1/2}(\Gamma_{R/\varepsilon})$ is equipped with the following norm:

$$\|v\|_{1/2,\Gamma_{R/\varepsilon}} = \inf\left\{\|u\|_{1,C(R/(2\varepsilon),R/\varepsilon)};\quad u = v \quad \text{on} \quad \Gamma_{R/\varepsilon}\right\},$$

where $C(r,r') := \left\{x \in \mathbb{R}^3;\quad r < \|x\| < r'\right\}$.

- The dual space $H^{-1/2}(\Gamma_{R/\varepsilon})$ is equipped with the natural norm

$$\|w\|_{-1/2,\Gamma_{R/\varepsilon}} = \sup\left\{\langle w, v\rangle_{-1/2,1/2}\,;\quad v \in H^{1/2}(\Gamma_{R/\varepsilon}),\, \|v\|_{1/2,\Gamma_{R/\varepsilon}} = 1\right\}.$$

It can easily be checked that if $\psi \in H^1(C(R/2,R))$ with $\Delta\psi = 0$ in $C(R/2,R)$, then

$$\|\nabla\psi.\mathbf{n}\|_{-1/2,\Gamma_R} \leq c\ |\psi|_{1,C(R/2,R)}\,. \tag{44}$$

Here and in what follows, c is a positive constant independent of the data (e.g., on ε).

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

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

Due to $\nabla u(x) = \nabla\widetilde{u}(y)/\varepsilon$ and to definition (43), we have

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

hence

$$|u|_{1,\mathcal{O}} = \varepsilon^{1/2}\,|\widetilde{u}|_{1,\widetilde{\mathcal{O}}}\,. \tag{45}$$

Similarly, we have

$$\|u\|_{0,\mathcal{O}} = \varepsilon^{3/2}\,\|\widetilde{u}\|_{0,\widetilde{\mathcal{O}}}\,. \tag{46}$$

LEMMA 7.1. *For $\varphi \in H^{1/2}(\partial\omega)$ let v be the solution to the problem*

$$\begin{cases} -\Delta v = 0 & \text{in } \mathbb{R}^3\backslash\overline{\omega}, \\ v = 0 & \text{at } \infty, \\ v = \varphi & \text{on } \partial\omega. \end{cases} \tag{47}$$

The function v is split into

$$v(y) = V(y) + W(y),$$
$$V(y) = E(y) \int_{\partial\omega} p(x)\, d\gamma(x),$$

where $E(y) = 1/4\pi \|y\|$, and $p \in H^{-1/2}(\partial\omega)$ is the unique solution to

$$\int_{\partial\omega} E(y-x)p(x)d\gamma(x) = \varphi(y) \qquad \forall y \in \partial\omega. \tag{48}$$

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

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

Proof. The function v reads

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

Using a Taylor expansion of E computed at the point y and the well-posedness of (48) we have for large $\|y\|$

$$\begin{aligned}
|V(y)| &\le \frac{c}{r}\|\varphi\|_{1/2,\partial\omega}, \qquad |W(y)| \le \frac{c}{r^2}\|\varphi\|_{1/2,\partial\omega},\\
|\nabla V(y)| &\le \frac{c}{r^2}\|\varphi\|_{1/2,\partial\omega}, \qquad |\nabla W(y)| \le \frac{c}{r^3}\|\varphi\|_{1/2,\partial\omega},
\end{aligned}$$

from which the above inequalities follow straightforwardly. ∎

LEMMA 7.2. *For $\varphi \in H^{1/2}(\Gamma_R)$, let v_ε be the solution to the problem*

$$\begin{cases} -\Delta v_\varepsilon = 0 & \text{in } D_\varepsilon,\\ v_\varepsilon = \varphi & \text{on } \Gamma_R,\\ v_\varepsilon = 0 & \text{on } \partial\omega_\varepsilon. \end{cases} \tag{49}$$

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

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

Proof. Let $\varepsilon_0 > 0$. Problem (49) is well-posed; hence there exists a constant c such that

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

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

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

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

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

We also have

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

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

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

LEMMA 7.3. *For $\varepsilon > 0$ and $\psi \in H^1(D_0)$, let u_ε be the solution to the problem*

$$\left\{\begin{array}{rl} -\Delta u_\varepsilon = 0 & \text{in } D_\varepsilon,\\ u_\varepsilon = 0 & \text{on } \Gamma_R,\\ u_\varepsilon = \psi & \text{on } \partial\omega_\varepsilon.\end{array}\right. \tag{50}$$

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

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

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

$$\left\{\begin{array}{ll} -\Delta\widetilde{v}_\varepsilon = 0 & \text{in } \mathbb{R}^3\backslash\overline{\omega},\\ \widetilde{v}_\varepsilon = 0 & \text{at } \infty,\\ \widetilde{v}_\varepsilon = \psi(\varepsilon y) & \text{on } \partial\omega.\end{array}\right.$$

The function u_ε can be written

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

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

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

It follows from (45), (46), and Lemmas 7.1 and 7.2 that there exist $c > 0$ and $\varepsilon_1 > 0$ such that for all $0 < \varepsilon < \varepsilon_1$,

$$\begin{aligned}
|v_\varepsilon|_{1,C(R/2,R)} &\le c\,\varepsilon^{1/2}|\widetilde{v}_\varepsilon|_{1,C(R/2\varepsilon,R/\varepsilon)} \le c\,\varepsilon\,\|\psi(\varepsilon y)\|_{1/2,\partial\omega}\,,\\
\|w_\varepsilon\|_{1,D_\varepsilon} &\le c\,\|v_\varepsilon\|_{1/2,\Gamma_R}\\
&\le c\,\|v_\varepsilon\|_{1,C(R/2,R)}\\
&\le c\,(|v_\varepsilon|_{0,C(R/2,R)} + |v_\varepsilon|_{1,C(R/2,R)})\\
&= c\,(\varepsilon^{3/2}|\widetilde{v}_\varepsilon|_{0,C(R/2\varepsilon,R/\varepsilon)} + \varepsilon^{1/2}|\widetilde{v}_\varepsilon|_{1,C(R/2\varepsilon,R/\varepsilon)})\\
&\le c\,\varepsilon\,\|\psi(\varepsilon y)\|_{1/2,\partial\omega}\,. \qquad (51)
\end{aligned}$$

Hence

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

Similarly we have

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

and

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

Lemmas 7.2 and 7.3 are summarized in the following lemma.

LEMMA 7.4. *Let v_ε be the solution to the problem*

$$\begin{cases} -\Delta v_\varepsilon = 0 & \text{in } D_\varepsilon,\\ v_\varepsilon = \varphi & \text{on } \Gamma_R,\\ v_\varepsilon = \psi & \text{on } \partial\omega_\varepsilon, \end{cases}$$

where $\varphi \in H^{1/2}(\Gamma_R)$ and $\psi \in H^1(D_0)$. There exist a constant $c > 0$ (independent of φ, ψ, and ε) and $\varepsilon_1 > 0$ such that for all $0 < \varepsilon < \varepsilon_1$,

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

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

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

For $\varphi \in H^{1/2}(\Gamma_R)$, recall that $u_\varepsilon^{0,\varphi}$ is the solution to (6) or (7) if $\varepsilon = 0$. Let $v_\omega^{0,\varphi}$ be the solution to

$$\begin{cases} -\Delta v_\omega^{0,\varphi} = 0 & \text{in } \mathbb{R}^3\backslash\overline{\omega},\\ v_\omega^{0,\varphi} = 0 & \text{at } \infty,\\ v_\omega^{0,\varphi} = u_0^{0,\varphi}(x_0) & \text{on } \partial\omega. \end{cases} \qquad (52)$$

As (19)–(22), let $P_\omega^{0,\varphi}(y) = A_\omega(u_0^{0,\varphi}(x_0))E(y)$ be the dominant part of $v_\omega^{0,\varphi}$, and let $Q_\omega^{0,\varphi} = A_\omega(u_0^{0,\varphi}(x_0))/(4\pi R)$, $P_\omega^{0,\varphi}(x) = Q_\omega^{0,\varphi}$ on Γ_R. The linear operator δT (independent of ε) is defined as follows:

$$\begin{array}{rcl} \delta T : H^{1/2}(\Gamma_R) & \longrightarrow & H^{-1/2}(\Gamma_R), \\ \varphi & \longmapsto & \delta T\varphi := -\nabla P_\omega^{0,\varphi}.\mathbf{n}. \end{array} \tag{53}$$

PROPOSITION 7.5. *The asymptotic expansion of T_ε is*

$$\|T_\varepsilon - T_0 - \varepsilon\delta T\|_{\mathcal{L}(H^{1/2}(\Gamma_R);\, H^{-1/2}(\Gamma_R))} = O(\varepsilon^2). \tag{54}$$

Proof. Let $\varphi \in H^{1/2}(\Gamma_R)$. For simplicity we drop the subscripts $(.)^{0,\varphi}$. For $y = x/\varepsilon$, we have $v_\omega(y) = P_\omega(y) + W_\omega(y)$ with $P_\omega(x/\varepsilon) = \varepsilon P_\omega(x)$ and $W_\omega(y) = O(1/\|y\|^2)$. Let

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

We have

$$\begin{aligned} \psi_\varepsilon(x) &= \left(\nabla u_\varepsilon - \nabla u_0 + \varepsilon\nabla\left(P_\omega - Q_\omega\right)\right).\mathbf{n} \\ &= \nabla(w_\varepsilon(x) - W_\omega(x/\varepsilon)).\mathbf{n}, \end{aligned}$$

where w_ε is defined by

$$w_\varepsilon(x) = u_\varepsilon(x) - u_0(x) + v_\omega(x/\varepsilon) - \varepsilon Q_\omega.$$

The function w_ε is the solution to

$$\left\{ \begin{array}{ll} -\Delta w_\varepsilon = 0 & \text{in } D_\varepsilon, \\ w_\varepsilon = v_\omega(x/\varepsilon) - \varepsilon Q_\omega & \text{on } \Gamma_R, \\ w_\varepsilon = -u_0(x) + u_0(x_0) - \varepsilon Q_\omega & \text{on } \partial\omega_\varepsilon. \end{array} \right.$$

In order to apply Lemma 7.4, we have to estimate the two right-hand sides.

On Γ_R, due to $P_\omega(x) = Q_\omega$, we have

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

Using (45), (46), Lemma 7.1, and elliptic regularity we obtain

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

On ω_ε, putting $\theta_\varepsilon(x) := (-u_0(x) + u_0(x_0) - \varepsilon Q_\omega)/\varepsilon$, we have for small ε

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

We can now apply Lemma 7.4, which gives

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

Finally it follows from (44), (45), and Lemma 7.1 that

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

Hence

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

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

PROPOSITION 7.6. *Let*

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

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

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

7.4. Variation of the linear form. The technique is the same as in section 7.3. The difference comes from the boundary condition imposed on $\partial\omega$ to the solution to the exterior problem: $v_\omega^{0,\varphi} = u_0^{0,\varphi}(x_0)$ in (52) for the study of the bilinear form, and $v_\omega^{f,0} = u_0^{f,0}(x_0)$ in (55) for the study of the linear form. Hence estimations involving $\|\varphi\|_{1/2,\Gamma_R}$ are replaced by estimations involving $\|f\|_{L^q}$.

The variation of the linear form l_ε reads

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

Recall that $u_\varepsilon^{f,0}$ is the solution to (6) or (7) if $\varepsilon = 0$. Let $v_\omega^{f,0}$ be the solution to

$$\left\{\begin{array}{ll} -\Delta v_\omega^{f,0} = 0 & \text{in } \mathbb{R}^3\backslash\overline{\omega},\\ v_\omega^{f,0} = 0 & \text{at } \infty,\\ v_\omega^{f,0} = u_0^{f,0}(x_0) & \text{on } \partial\omega. \end{array}\right. \tag{55}$$

As with (19)–(22), let $P_\omega^{f,0}(y) = A_\omega(u_0^{f,0}(x_0))E(y)$ be the dominant part of $v_\omega^{f,0}$, and let $Q_\omega^{f,0} = A_\omega(u_0^{f,0}(x_0))/(4\pi R)$, $P_\omega^{f,0}(x) = Q_\omega^{f,0}$ on Γ_R. The function $\delta f \in H^{-1/2}(\Gamma_R)$ (independent of ε) is defined by

$$\delta f = \nabla P_\omega^{f,0}.\mathbf{n}. \tag{56}$$

PROPOSITION 7.7. *Let $f \in L^q(\Omega)$, $q > n$. The asymptotic expansion of f_ε with respect to ε is*

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

Proof. The proof runs as in Proposition 7.5 (we drop the subscripts $(.)^{f,0}$) with w_ε and θ_ε defined by

$$w_\varepsilon(x) = u_\varepsilon(x) - u_0(x) + v_\omega(x/\varepsilon) - \varepsilon Q_\omega,$$
$$\theta_\varepsilon(x) = (-u_0(x) + u_0(x_0) - \varepsilon Q_\omega)/\varepsilon.$$

The only difference lies in the elliptic regularity estimate [20, 21]

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

and for small ε,

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

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

PROPOSITION 7.8. *Let*

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

Then the asymptotic expansion of linear form l_ε is given by

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

7.5. Proof of Theorem 5.1. The fundamental hypotheses (1) and (2) are satisfied; hence we can apply Theorem 2.2:

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

It follows from (13), (16), (52), and (55) that

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

which implies

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

Then, using (53), (56), and Propositions 7.6 and 7.8, we obtain

$$\begin{aligned}
\delta a(u,v) - \delta l(v) &= \int_{\Gamma_R} -\left(\nabla P_\omega^{0,\varphi} + \nabla P_\omega^{f,0}\right).\mathbf{n}\, v\, d\gamma(x) \\
&= -\int_{\Gamma_R} \nabla P_\omega.\mathbf{n}\, v\, d\gamma(x),
\end{aligned}$$

which achieves the proof of Theorem 5.1.

7.6. Proof of Proposition 5.3. This section describes the variations of $J_\varepsilon(u) = \widetilde{J}_\varepsilon(\widetilde{u}_\varepsilon)$ (see (14)) when $\widetilde{J}_\varepsilon$ is of the form (31)

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

The hypotheses on g (32)–(34) described in section 5 are supposed to be satisfied. Throughout this and the next subsection, $\widetilde{u}_\varepsilon \in H^1(\Omega_\varepsilon)$, $\varepsilon \geq 0$, denotes the extension of $u \in \mathcal{V}_R$ which coincides with u on Ω_R and with $u_\varepsilon^{f,\varphi}$ on D_ε for $\varphi = u_{|\Gamma_R}$.

LEMMA 7.9. *Let $\varphi \in H^{1/2}(\Gamma_R)$ and $f \in L^q(\Omega)$, $q > n$. Let $u_\varepsilon^{f,\varphi}$ and $u_0^{f,\varphi}$ be, respectively, the solutions to* (6) *and* (7). *Then*

$$\left\| u_\varepsilon^{f,\varphi} - u_0^{f,\varphi} - \varepsilon(Q_\omega^{f,\varphi} - P_\omega^{f,\varphi}) \right\|_{0,D_\varepsilon} = O(\varepsilon^{3/2}), \tag{57}$$

$$\left\| u_\varepsilon^{f,\varphi} - u_0^{f,\varphi} - \varepsilon Q_\omega^{f,\varphi} + v_\omega^{f,\varphi}(x/\varepsilon) \right\|_{1,D_\varepsilon} = O(\varepsilon^{3/2}), \tag{58}$$

where $P_\omega^{f,\varphi}$ is the dominant part (19) *of the solution $v_\omega^{f,\varphi}$ to the exterior problem* (16) *with $u_0^{f,\varphi}(x_0)$ substituted for $u_\Omega(x_0)$ and $Q_\omega^{f,\varphi}$ is the associated constant* (22).

Proof. Recall that $v_\omega = P_\omega + W_\omega$ (19) with $P_\omega(x/\varepsilon) = \varepsilon P_\omega(x)$ and $W_\omega(y) = O(1/\|y\|^2)$ (we drop the subscripts $(.)^{f,\varphi}$). Let

$$\begin{aligned} w_\varepsilon(x) &= (u_\varepsilon - u_0 - \varepsilon(Q_\omega - P_\omega))(x) + W_\omega(x/\varepsilon) \\ &= u_\varepsilon(x) - u_0(x) + v_\omega(x/\varepsilon) - \varepsilon Q_\omega. \end{aligned} \tag{59}$$

The function w_ε is the solution to

$$\begin{cases} -\Delta w_\varepsilon = 0 & \text{in } D_\varepsilon, \\ w_\varepsilon = v_\omega(x/\varepsilon) - \varepsilon Q_\omega & \text{on } \Gamma_R, \\ w_\varepsilon = -u_0(x) + u_0(x_0) - \varepsilon Q_\omega & \text{on } \partial\omega_\varepsilon. \end{cases}$$

Using the same arguments as in the proofs of Propositions 7.5 and 7.7 we obtain

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

It follows from Lemma 7.4 that

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

The second equation proves (58). Due to (46), Lemma 7.1, and elliptic regularity we also have

$$\begin{aligned} \|W_\omega(x/\varepsilon)\|_{0,D_\varepsilon} &= \varepsilon^{3/2} \|W_\omega\|_{0,D_\varepsilon/\varepsilon} \\ &\leq c\,\varepsilon^{3/2} \|u_0(x_0)\|_{1/2,\partial\omega} \\ &\leq c\,\varepsilon^{3/2} \left(\|\varphi\|_{1/2,\Gamma_R} + \|f\|_{L^q} \right). \end{aligned}$$

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

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

LEMMA 7.10. *For $u \in \mathcal{V}_R$ we have*

$$J_\varepsilon(u) = J_0(u) + \varepsilon \delta J(u) + o(\varepsilon),$$
$$\delta J(u) = \int_{D_0} g_s(x, \widetilde{u}_0(x))(Q_\omega - P_\omega)\, dx,$$

where Q_ω and P_ω are defined as in Lemma 7.9 with $\varphi = u_{|\Gamma_R}$.

Proof. Let

$$I_\varepsilon = J_\varepsilon(u) - J_0(u) - \varepsilon \int_{D_0} g_s(x, \widetilde{u}_0)(Q_\omega - P_\omega)\, dx.$$

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

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

Due to the hypotheses on g (32) and (33), we have for all $(x, s, t) \in \Omega \times \mathbb{R} \times \mathbb{R}$

$$g(x,t) - g(x,s) = g_s(x,s)(t-s) + \theta(x,s,t)(t-s)^2,$$
$$|\theta(x,s,t)| \leq \frac{M}{2}.$$

Then

$$\begin{aligned} I_\varepsilon = &\int_{D_\varepsilon} g_s(x, u_0^{f,\varphi}) \left(u_\varepsilon^{f,\varphi} - u_0^{f,\varphi} - \varepsilon(Q_\omega - P_\omega) \right) dx \\ &- \varepsilon \int_{\omega_\varepsilon} g_s(x, u_0^{f,\varphi})(Q_\omega - P_\omega)\, dx - \int_{\omega_\varepsilon} g(x, u_0^{f,\varphi})\, dx \\ &+ \int_{D_\varepsilon} \theta(x, u_0^{f,\varphi}, u_\varepsilon^{f,\varphi})(u_\varepsilon^{f,\varphi} - u_0^{f,\varphi})^2\, dx \end{aligned}$$

and

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

It follows from the hypotheses on g (32)–(34), Lemma 7.9, the regularity of $u_0^{f,\varphi}$ (which implies that $x \mapsto g(x, u_0^{f,\varphi}(x))$ is in $L^{3/2}(B(0, R/2))$), $\|P_\omega\|_{0,\omega_\varepsilon} = c\,\varepsilon^{1/2}$, and $\|g_s(., u_0^{f,\varphi}(.))\|_{0,\omega_\varepsilon} = o(1)$ that

$$\begin{aligned} \int_{D_\varepsilon} \left| g_s(x, u_0^{f,\varphi}) \left(u_\varepsilon^{f,\varphi} - u_0^{f,\varphi} - \varepsilon(Q_\omega - P_\omega) \right) \right| dx &\leq c \left\| u_\varepsilon^{f,\varphi} - u_0^{f,\varphi} - \varepsilon(Q_\omega - P_\omega) \right\|_{0,D_\varepsilon} \\ &= O(\varepsilon^{3/2}), \\ \int_{D_\varepsilon} \frac{M}{2} (u_\varepsilon^{f,\varphi} - u_0^{f,\varphi})^2\, dx &= O(\varepsilon^2), \end{aligned}$$

$$\varepsilon \int_{\omega_\varepsilon} \left| g_s(x, u_0^{f,\varphi})(Q_\omega - P_\omega) \right| \, dx \le \varepsilon \left\| g_s(., u_0^{f,\varphi}(.)) \right\|_{0,\omega_\varepsilon} \|Q_\omega - P_\omega\|_{0,\omega_\varepsilon}$$
$$= o(\varepsilon^{3/2}),$$

$$\int_{\omega_\varepsilon} \left| g(x, u_0^{f,\varphi}) \right| \, dx \le \left(\int_{\omega_\varepsilon} \left| g(x, u_0^{f,\varphi}) \right|^{3/2} \, dx \right)^{2/3} \left(\int_{\omega_\varepsilon} dx \right)^{1/3} \tag{60}$$
$$= o(\varepsilon).$$

Hence

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

We can now check hypothesis (24) involved by Theorem 5.1.

PROPOSITION 7.11. *The function J_0 is differentiable on $\mathcal{V}_R$, and we have for all $u, v \in \mathcal{V}_R$*

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

Proof. We have

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

It follows from the hypotheses on g (32) that the function $\widetilde{J}_0$ is differentiable on $H^1(\Omega)$ with

$$D\widetilde{J}_0(\widetilde{u}_0)w = \int_\Omega g_s(x, \widetilde{u}_0) w \, dx, \qquad w \in H^1(\Omega).$$

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

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

Hence, applying Lemma 7.10 yields

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

It remains to prove that $\varepsilon(\delta J(v) - \delta J(u)) = o(\varepsilon + \|v - u\|_{\mathcal{V}_R})$. For this it is sufficient to prove that $\delta J(v) - \delta J(u) = O(\|v - u\|_{\mathcal{V}_R})$. With the notation defined below in (61) and (62), it follows from Lemma 7.10 that

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

Hence, using the hypotheses on g (32)–(33) we obtain

$$|\delta J(v) - \delta J(u)| \leq \int_{D_0} M\,|\widetilde{v}_0 - \widetilde{u}_0|\,|Q_\omega^v - P_\omega^v|\,dx \\ + \int_{D_0} (|g_s(x,0)| + M\,|\widetilde{u}_0|)\,(|Q_\omega^v - Q_\omega^u| + |P_\omega^v - P_\omega^u|)\,dx.$$

We conclude by using linearity and continuity of

$$\begin{array}{ccccccc} \mathcal{V}_R & \to & H^{1/2}(\Gamma_R) & \to & H^1(D_0) & \to & L^2(D_0), \\ u & \mapsto & \varphi := u_{|\Gamma_R} & \mapsto & u_0^{f,\varphi} & \mapsto & P_\omega^u := A_\omega(u_0^{f,\varphi})(x_0))E \end{array} \tag{61}$$

and

$$\begin{array}{ccccc} \mathcal{V}_R & \to & H^{1/2}(\Gamma_R) & \to & \mathbb{R}, \\ u & \mapsto & (P_\omega^u)_{|\Gamma_R} & \mapsto & Q_\omega^u. \end{array} \tag{62}$$

□

Hence, hypothesis (24) is fulfilled and we can apply Theorem 5.1. The adjoint equation (27) reads

$$\int_\Omega \nabla v_\Omega.\nabla w\,dx = -\int_\Omega g_s(x,u_\Omega)w\,dx,$$

and hence

$$\Delta v_\Omega = g_s(x,u_\Omega). \tag{63}$$

It follows from (30), Lemma 7.10, and (63) that

$$\begin{aligned} \delta j(x_0) &= A_\omega(u_\Omega(x_0))v_\Omega(x_0) + \int_{D_0} \Delta v_\Omega(P_\omega - Q_\omega)\,dx + \delta J(u_0) \\ &= A_\omega(u_\Omega(x_0))v_\Omega(x_0) + \int_{D_0} \Delta v_\Omega(P_\omega - Q_\omega)\,dx + \int_{D_0} g_s(x,u_\Omega)(Q_\omega - P_\omega)\,dx \\ &= A_\omega(u_\Omega(x_0))v_\Omega(x_0), \end{aligned}$$

which achieves the proof of Proposition 5.3.

7.7. Proof of Proposition 5.4. Here $\widetilde{J}_\varepsilon$ is of the form (35):

$$\widetilde{J}_\varepsilon(v) = \frac{1}{2}\int_{\Omega_\varepsilon} B(x)\nabla(v - u_{\mathrm{d}}).\nabla(v - u_{\mathrm{d}})\,dx, \qquad v \in H^1(\Omega_\varepsilon).$$

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

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

Due to the assumption on u_{d} and to $f \in L^q(\Omega)$, $q > n$, we have ∇u_{d}, $\nabla\,\widetilde{u}_0 \in \mathcal{C}^0(\overline{B}(0,R/2))^3$ [12], and hence

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

This and the fact that $b_{ij}(x) = b_{ji}(x)$ yields

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

Equation (59) reads here

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

with Q_ω constant, and

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

Recall that $v_\omega = P_\omega + W_\omega$ (19) with $P_\omega(x/\varepsilon) = \varepsilon P_\omega(x)$ and $W_\omega(y) = O(1/\|y\|^2)$. Then

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

Here ∇_x denotes the derivative with respect to x, and particularly $\nabla(v(x/\varepsilon)) = \nabla_x v(x/\varepsilon) = \nabla v(x/\varepsilon)/\varepsilon$. It follows from Lemmas 7.1 and 7.9 that $\|v_\omega(x/\varepsilon)\|_{1,D_\varepsilon} = O(\varepsilon^{1/2})$ and $\|w_\varepsilon\|_{1,D_\varepsilon} = O(\varepsilon^{3/2})$; hence

$$\begin{aligned} \int_{D_\varepsilon} B\nabla(\widetilde{u}_0 - u_{\mathrm{d}}).\nabla w_\varepsilon(x)\, dx &= O(\varepsilon^{3/2}), \\ \int_{D_\varepsilon} B\nabla w_\varepsilon.\nabla w_\varepsilon\, dx &= O(\varepsilon^3), \\ -\int_{D_\varepsilon} B\nabla_x v_\omega(x/\varepsilon).\nabla w_\varepsilon\, dx &= O(\varepsilon^2). \end{aligned}$$

We have $\|\nabla_x W_\omega(x/\varepsilon)\|_{L^1(D_\varepsilon)} = \varepsilon^2 \|\nabla W_\omega\|\,|_{L^1(D_\varepsilon/\varepsilon)} = O(\varepsilon^2|\log\varepsilon|)$, and thus

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

Using $\nabla P_\omega = O(1/r^2)$, which implies that $\int_{\omega_\varepsilon} B\nabla(\widetilde{u}_0 - u_\mathrm{d}).\nabla P_\omega\,dx = O(\varepsilon)$, we obtain (with $\nabla Q_\omega = 0$)

$$J_\varepsilon(u) - J_0(u) = -\varepsilon \int_{D_0} B\nabla(\widetilde{u}_0 - u_\mathrm{d}).\nabla(P_\omega - Q_\omega)\,dx + \frac{1}{2}\int_{D_\varepsilon} B\nabla_x v_\omega(x/\varepsilon).\nabla_x v_\omega(x/\varepsilon)\,dx + o(\varepsilon).$$

The adjoint equation implies for all $\varphi \in \mathcal{D}(D_0)$

$$\int_{D_0} -\Delta v_\Omega\,\varphi\,dx = -\int_{D_0} B\nabla(u_\Omega - u_\mathrm{d}).\nabla\varphi\,dx. \tag{64}$$

Due to $B \in W^{1,\infty}(\Omega)^{3\times 3}$ and Δu_d, $f \in L^q(\Omega)$ (thus $D^2 u_\mathrm{d}$, $D^2 u_\Omega \in L^q(D_0)$; cf. the Calderon–Zygmund theorem [12]), we have $-\Delta v_\Omega = \operatorname{div}[B\nabla(u_\Omega - u_\mathrm{d})] \in L^q(D_0)$. Moreover, $q > n/2$ and $P_\omega - Q_\omega \in L^m(D_0)$ for all $m < 3$; thus $\Delta v_\Omega(P_\omega - Q_\omega) \in L^1(D_0)$. Hence, as $P_\omega - Q_\omega$ vanishes on Γ_R, (64) still holds for $\varphi = P_\omega - Q_\omega$, and

$$J_\varepsilon(u_0) - J_0(u_0) = -\int_{D_0} \varepsilon\Delta v_\Omega.(P_\omega - Q_\omega) + \frac{1}{2}\int_{D_\varepsilon} B\nabla_x v_\omega(x/\varepsilon).\nabla_x v_\omega(x/\varepsilon)\,dx + o(\varepsilon).$$

Then the proof can be achieved as in section 7.6. It follows from (30) that

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

Using Lebesgue's convergence theorem, we deduce that

$$\begin{aligned}\int_{D_\varepsilon} B(x)\nabla_x v_\omega(x/\varepsilon).\nabla_x v_\omega(x/\varepsilon)\,dx &= \varepsilon\int_{D_\varepsilon/\varepsilon} B(\varepsilon y)\nabla v_\omega(y).\nabla v_\omega(y)\,dy \\ &= \varepsilon\int_{\mathbb{R}^3\setminus\overline{\omega}} B(x_0)\nabla v_\omega(y).\nabla v_\omega(y)\,dy + o(\varepsilon),\end{aligned}$$

which proves that

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

If ω is the unit ball, then we have the evident solution $v_\omega = u_\Omega(x_0)/r$ and

$$\begin{aligned}\int_{\mathbb{R}^3\setminus\overline{\omega}} B(x_0)\nabla v_\omega(y).\nabla v_\omega(y)\,dy &= u_\Omega(x_0)^2\int_{\mathbb{R}^3\setminus B(0,1)} \frac{1}{r^4} B(x_0)\mathbf{e}_\mathrm{r}.\mathbf{e}_\mathrm{r}\,dy \\ &= u_\Omega(x_0)^2\,\operatorname{tr} B(x_0)\frac{1}{3}\int_{\mathbb{R}^3\setminus B(0,1)}\frac{dy}{r^4} \\ &= \frac{4\pi u_\Omega(x_0)^2}{3}\operatorname{tr} B(x_0),\end{aligned}$$

where $\mathbf{e}_\mathrm{r}(y) = y/\|y\|$. This completes the proof of Proposition 5.4.

Acknowledgment. The authors are grateful to the referees for their comments and suggestions, which have greatly helped improve this work and its presentation.

REFERENCES

[1] G. Allaire and R. Kohn, *Optimal bounds on the effective behavior of a mixture of two well-ordered elastic materials,* Quart. Appl. Math., 51 (1993), pp. 643–674.
[2] M. Becker, *Optimisation topologique de structure en variables discrètes,* Thesis, Université de Lièges, 1996.
[3] M. Bendsøe, *Optimal Topology Design of Continuum Structure: An Introduction,* Technical report, Department of Mathematics, Technical University of Denmark, Lyngby, Denmark, 1996.
[4] H. Brezis, *Analyse fonctionnelle,* Masson, Paris, 1993.
[5] G. Buttazzo and G. Dal Maso, *Shape optimization for Dirichlet problems: Relaxed formulation and optimality conditions,* Appl. Math. Optim., 23 (1991), pp. 17–49.
[6] M. Chipot and G. Dal Maso, *Relaxed shape optimization: The case of nonnegative data for the Dirichlet problem,* Adv. Math. Sci. Appl., 1 (1992), pp. 47–81.
[7] J. Céa, *Conception optimale ou identification de forme, calcul rapide de la dérivée directionnelle de la fonction coût,* RAIRO Modél. Math. Anal. Numér., 20 (1986), pp. 371–402.
[8] J. Céa, A. Gioan, and J. Michel, *Quelques résultats sur l'identification de domains,* Calcolo, 10 (1973), pp. 207–232.
[9] J. Céa, S. Garreau, Ph. Guillaume, and M. Masmoudi, *The shape and topological optimizations connection,* Comput. Methods Appl. Mech. Engrg., 188 (2000), pp. 713–726.
[10] D.J. Cedio-Fengya, S. Moskow, and M. Vogelius, *Identification of Conductivity Imperfections of Small Diameter by Boundary Measurements. Continuous Dependence and Computational Reconstruction,* Preprint 1502, Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, MN, 1997.
[11] P. Ciarlet, *The Finite Element Method for Elliptic Problems,* North-Holland, Amsterdam, 1978; reprinted by SIAM, Philadelphia, PA, 2002.
[12] R. Dautray and J. Lions, *Analyse mathématique et calcul numérique pour les sciences et les techniques,* INSTN: Collection Enseignement, Masson, Paris, 1987.
[13] A. Friedman and M. Vogelius, *Identification of small inhomogeneities of extreme conductivity by boundary measurements: A theorem on continuous dependence,* Arch. Ration. Mech. Anal., 105 (1989), pp. 267–278.
[14] J. Giroire, *Formulations variationnelles par équations intégrales de problèmes aux limites extérieurs,* Thesis, Ecole Polytechnique, Palaiseau, France, 1976.
[15] S. Garreau, Ph. Guillaume, and M. Masmoudi, *The topological sensitivity for linear isotropic elasticity,* in European Conference on Computational Mechanics (ECCM99), report MIP 99.45, Université Paul Sabatia, 1999.
[16] S. Garreau, Ph. Guillaume, and M. Masmoudi, *The topological asymptotic for PDE systems: The elasticity case,* SIAM J. Control Optim., 39 (2001), pp. 1756–1778.
[17] P. Guillaume and M. Masmoudi, *Computation of high order derivatives in optimal shape design,* Numer. Math., 67 (1994), pp. 231–250.
[18] J. Jacobsen, N. Olhoff, and E. Rønholt, *Generalized Shape Optimization of Three-Dimensional Structures Using Materials with Optimum Microstructures,* Technical report, Institute of Mechanical Engineering, Aalborg University, Aalborg, Denmark, 1996.
[19] V.D. Kupradze, ed., *Three-Dimensional Problems of the Mathematical Theory of Elasticity and Thermoelasticity,* North-Holland Ser. Appl. Math. Mech. 25, North-Holland, Amsterdam, 1979.
[20] O.A. Ladyženskaja and N.N. Ural'ceva, *Équations aux dérivées partielles de type elliptique,* Dunod, Paris, 1968.
[21] J. Lions and E. Magenes, *Problèmes aux limites non homogenes et applications,* Dunod, Paris, 1968.
[22] M. Masmoudi, *Outils pour la conception optimale de formes,* thèse d'état, Université de Nice, Nice, France, 1987.
[23] M. Masmoudi, *The topological asymptotic,* in Computational Methods for Control Applications, International Series GAKUTO, H. Kawarada and J. Periaux, eds., to appear.
[24] F. Murat and J. Simon, *Sur le contrôle par un domaine géométrique,* thèse d'état, Université Pierre et Marie Curie, Paris, 1976.
[25] F. Murat and L. Tartar, *Calcul des variations et homogénéisation,* in Les méthodes de l'homogénéisation: Théorie et Applications en Physique, Eyrolles, 1985, pp. 319–369.

[26] M. Schoenauer, L. Kallel, and F. Jouve, *Mechanical inclusions identification by evolutionary computation,* Rev. Européenne Élém. Finis, 5 (1996), pp. 619–648.
[27] A. Schumacher, *Topologieoptimierung von Bauteilstrukturen unter Verwendung von Lopchpositionierungkrieterien,* Thesis, Universität-Gesamthochschule-Siegen, 1995.
[28] J. Sokołowski and A. Żochowski, *On the topological derivative in shape optimization,* SIAM J. Control Optim., 37 (1999), pp. 1251–1272.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 4, pp. 1073–1088

CONTINUOUS-TIME DYNKIN GAMES WITH MIXED STRATEGIES*

NIZAR TOUZI† AND NICOLAS VIEILLE‡

Abstract. Let (X, Y, Z) be a triple of payoff processes defining a Dynkin game

$$\tilde{R}(\sigma,\tau) = E\left[X_\sigma \mathbf{1}_{\{\tau>\sigma\}} + Y_\tau \mathbf{1}_{\{\tau<\sigma\}} + Z_\tau \mathbf{1}_{\{\tau=\sigma\}}\right],$$

where σ and τ are stopping times valued in $[0, T]$. In the case $Z = Y$, it is well known that the condition $X \leq Y$ is needed in order to establish the existence of value for the game, i.e., $\inf_\tau \sup_\sigma \tilde{R}(\sigma,\tau) = \sup_\sigma \inf_\tau \tilde{R}(\sigma,\tau)$.

In order to remove the condition $X \leq Y$, we introduce an extension of the Dynkin game by allowing for an extended set of strategies, namely, the set of mixed strategies. The main result of the paper is that the extended Dynkin game has a value when $Z \leq Y$, and the processes X and Y are restricted to be semimartingales continuous at the terminal time T.

Key words. optimal stopping, Dynkin games, stochastic analysis, minimax theorem

AMS subject classifications. Primary, 60G40, 90D15; Secondary, 60H30, 46N10

PII. S0363012900369812

1. Introduction.

Dynkin games have been introduced by Dynkin (1967) as a generalization of optimal stopping problems. Since then, many authors contributed to solve the problem both in discrete and continuous-time models; see, e.g., Dynkin and Yushkevich (1968), Bensoussan and Friedman (1974), Neveu (1975), Bismut (1977), Stettner (1982), Alario, Lepeltier, and Marchal (1982), Morimoto (1984), Lepeltier and Maingueneau (1984), Cvitanić and Karatzas (1996), and Karatzas and Wang (2001), among others.

The setting of the problem is very simple. There are two players, labeled Player 1 and Player 2, who observe two payoff processes X and Y defined on a probability space $(\Omega, \mathcal{F}, P)$. Player 1 (resp., 2) chooses a stopping time σ (resp., τ) as control for this optimal stopping problem. At (stopping) time $\sigma \wedge \tau$ the game is over, and Player 2 pays the amount $X_\sigma \mathbf{1}_{\{\tau>\sigma\}} + Y_\tau \mathbf{1}_{\{\tau<\sigma\}} + Z_\tau \mathbf{1}_{\{\tau=\sigma\}}$ to Player 1. Therefore the objective of Player 1 is to maximize this payment, while Player 2 wishes to minimize it. It is then natural to introduce the lower and upper values of the game:

$$\sup_\sigma \inf_\tau E\left[X_\sigma \mathbf{1}_{\{\tau>\sigma\}} + Y_\tau \mathbf{1}_{\{\tau<\sigma\}} + Z_\tau \mathbf{1}_{\{\tau=\sigma\}}\right],$$

$$\inf_\tau \sup_\sigma E\left[X_\sigma \mathbf{1}_{\{\tau>\sigma\}} + Y_\tau \mathbf{1}_{\{\tau<\sigma\}} + Z_\tau \mathbf{1}_{\{\tau=\sigma\}}\right].$$

If the above values are equal, then the game is said to have a value. In the previously cited literature, it is proved that the game has a value essentially under the conditions $X_\cdot \leq Y_\cdot = Z_\cdot$, P-a.s. A precise discussion of this is given in section 2.

The purpose of this paper is to remove the condition $X_\cdot \leq Y_\cdot = Z_\cdot$, P-a.s. by suitably convexifying the set of strategies of the players. This is achieved by introducing

*Received by the editors March 22, 2000; accepted for publication (in revised form) January 23, 2002; published electronically October 8, 2002.
http://www.siam.org/journals/sicon/41-4/36981.html

†Centre de Recherche, en Economie et Statistique, 15 Bd. Gabriel Péri, 92245 Malakoff, France (touzi@ensae.fr).

‡HEC, 78351 Jouy en Josas, France, and Laboratoire d'Econométrie de l'Ecole Polytechnique, Paris, France (vieille@poly.polytechnique.fr).

the notion of mixed strategies, standard in (discrete-time) game theory literature. Loosely speaking, instead of choosing a stopping time, we shall allow both players to choose a distribution on the set of stopping times. Namely, at each time, both players fix a probability of stopping and decide whether or not to stop according to this probability.

This leads us to define mixed strategies as nondecreasing right-continuous processes with zero initial data and final data less than 1. In section 7 of this paper, we provide two justifications of this definition. The first is obtained by enlarging the probability space in order to allow for an independent randomizing device for each player. The second justification consists of defining the notion of randomized stopping time by means of functional analysis arguments, as in Bismut (1979).

Section 3 reports the precise definition of the extended Dynkin game and the main result of the paper: the extended Dynkin game has a value, provided the payoff processes X and Y are semimartingales continuous at the terminal time T, and $Z_\cdot \leq Y_\cdot$, P-a.s. For ease of presentation, we split the proof as follows. Section 4 provides the main steps of the proof, which basically relies on the two following technical results. In the first one, reported in section 5, we prove that the players' strategy sets can be reduced without affecting the lower and the upper values of the game. The second one states that the game with restricted strategies has a value. The proof of the last claim, reported in section 6, is obtained by an application of Sion's min-max theorem.

Before concluding this introduction, let us set up some notation which will be extensively used in the paper.

Given a right-continuous process with left limits S, we denote $S_{t-} := \lim_{s\uparrow t} S_s$. The jumps of S are denoted by $\Delta S_t := S_t - S_{t-}$. We shall denote by ΔS the process of jumps of S, and by S_- the process of left limits of S.

We shall denote by λ the Lebesgue measure on $[0,T]$, and by E_λ the associated expectation operator. For a nondecreasing process A, we denote by m_A the positive finite measure induced by A. If S is a semimartingale, then it admits a decomposition $S = M + A$, where A is a finite variation process and M is a martingale. We shall denote by m_M the measure induced by the (nondecreasing) predictable quadratic variation process $\langle M, M\rangle$ of M, i.e., $m_M(B) = E_\lambda[\mathbf{1}_B \langle M\rangle_\infty]$. We abuse the latter notation by saying that some property holds m_S-a.s. whenever it holds both m_A-a.s. and m_M-a.s.

2. Dynkin game with pure strategies. In this section, we recall the classical formulation of a Dynkin game, as suggested by Dynkin and Yushkevich (1968), Neveu (1975), and Bismut (1977).

Let $(\Omega, \mathcal{F}, P)$ be a complete probability space, and let $T > 0$ be a fixed terminal time. Let $X = \{X_t,\, 0 \leq t \leq T\}$, $Y = \{Y_t,\, 0 \leq t \leq T\}$, and $Z = \{Z_t,\, 0 \leq t \leq T\}$ be real-valued càdlàg processes, satisfying the integrability condition

$$E\left[\sup_t |X_t| + \sup_t |Y_t| + \sup_t |Z_t|\right] < +\infty. \tag{2.1}$$

We denote by $\mathbb{F} = \{\mathcal{F}_t,\, 0 \leq t \leq T\}$ the P-augmentation of the filtration generated by (X, Y, Z), and by $\mathcal{T}$ the set of all stopping times for $\mathbb{F}$.

The structure of a Dynkin game is the following. Two players observe the triple of stochastic processes (X, Y, Z). Player 1 chooses a stopping time $\sigma \in \mathcal{T}$, and Player 2 chooses a stopping time $\tau \in \mathcal{T}$. Player 2 pays Player 1 the amount

$$X_\sigma \mathbf{1}_{\{\tau > \sigma\}} + Y_\tau \mathbf{1}_{\{\tau < \sigma\}} + Z_\tau \mathbf{1}_{\{\tau = \sigma\}}.$$

The payoff of the game is then defined by the expected value of the above payoff:

$$\tilde{R}(\sigma,\tau) := E\left[X_\sigma \mathbf{1}_{\{\tau>\sigma\}} + Y_\tau \mathbf{1}_{\{\tau<\sigma\}} + Z_\tau \mathbf{1}_{\{\tau=\sigma\}}\right].$$

Player 1 wishes to maximize $\tilde{R}(\sigma,\tau)$, while Player 2 wishes to minimize it. It is then natural to define the lower and upper values of the game:

$$\underline{V} := \sup_\sigma \inf_\tau \tilde{R}(\sigma,\tau) \qquad \text{and} \qquad \overline{V} := \inf_\tau \sup_\sigma \tilde{R}(\sigma,\tau),$$

which satisfy $\underline{V} \le \overline{V}$. If it happens that

$$\underline{V} = \overline{V},$$

then the above Dynkin game is said to have a value.

There is extensive literature providing sufficient conditions for the existence of the value for the continuous-time Dynkin game in the case $Z = Y$. Bismut (1977) proved existence of the value under the condition

$$X_\cdot \le Y_\cdot = Z_\cdot, \qquad P\text{-a.s.} \tag{2.2}$$

as well as some regularity conditions and Mokobodski's hypothesis (namely, that there exist positive bounded supermartingales Z and Z' satisfying $X \le Z - Z' \le Y$). The regularity assumption was weakened by Alario, Lepeltier, and Marchal (1982), and then Lepeltier and Maingueneau (1984) established the existence of the value without Mokobodski's hypothesis, assuming only $X_\cdot \le Y_\cdot = Z_\cdot$.

We also mention the paper by Cvitanić and Karatzas (1996), which derives the latter result in the context of a Brownian filtration by means of doubly reflected backward stochastic differential equations.

3. Dynkin game with mixed strategies. The chief goal of this paper is to remove condition (2.2) by "convexifying" the set of stopping times. A precise discussion of the problem of extending the set of strategies is provided in section 7. In this section, we give only the main intuition in order to obtain an extended version of the Dynkin game, and we state the main result of the paper.

The main idea is to identify stopping times with $\{0,1\}$-valued, nondecreasing processes. Then convexifying the set of these processes leads naturally to considering the set $\mathcal{V}^+$ of all adapted, nondecreasing, right-continuous processes A with $A_{0-} = 0$ and $A_T \le 1$.

More precisely, let $\mathcal{V}_{0,1}$ be the subset of $\{0,1\}$-valued processes of $\mathcal{V}^+$. For every stopping time τ, define the process F^τ by

$$F^\tau_t := \mathbf{1}_{\{\tau\le t\}}, \qquad 0 \le t \le T.$$

It is clear that $F^\tau \in \mathcal{V}_{0,1}$. Conversely, given $F \in \mathcal{V}_{0,1}$, let

$$\tau_F := \inf\{t \in [0,T] : \ F_t > 0\}$$

with the usual convention $\inf \emptyset = +\infty$. From the right-continuity of F, it is clear that τ_F is a stopping time for $\mathbb{F}$. This provides an identification of $\mathcal{V}_{0,1}$ and $\mathcal{T}$. Clearly, the payoff function $\tilde{R}$ can be written in terms of $F, G \in \mathcal{V}_{0,1}$ as

$$R(F,G) := \tilde{R}(\tau_F,\tau_G) = E\left[\int_0^T X(1-G)dF + \int_0^T Y(1-F)dG + \sum_{[0,T]} Z\Delta F\Delta G\right].$$

Observe that the right-hand side expression is well defined for $F, G \in \mathcal{V}^+$. Our interest is in the extended Dynkin game, in which players choose elements of $\mathcal{V}^+$, and the payoff is given by R. A rigorous justification of the set $\mathcal{V}^+$ as being the set of mixed strategies is reported in section 7, as well as the extension of the payoff function $\tilde{R}$ to $\mathcal{V}^+$.

The following is the main result of the paper.

THEOREM 3.1. *Let (X, Y, Z) be a triple of payoff processes satisfying* (2.1). *Suppose that X and Y are semimartingales with trajectories continuous at time T, P-a.s. Assume further that $Z \leq Y$. Then*

$$\sup_{F \in \mathcal{V}^+} \inf_{G \in \mathcal{V}^+} R(F, G) = \inf_{G \in \mathcal{V}^+} \sup_{F \in \mathcal{V}^+} R(F, G),$$

i.e., the extended Dynkin game has a value.

This theorem states that the Dynkin game has a value when the set of strategies $\mathcal{V}_{0,1}$ is convexified in the natural way. The only conditions required for this result are $Z \leq Y$, and X and Y are semimartingales continuous at the terminal time T. The reason for the restriction to semimartingales is explained in Remark 5.1.

An alternative way of convexifying the set $\mathcal{T}$ of stopping times is to allow the players to choose a randomized stopping time, i.e., a probability distribution over stopping times. This corresponds to the concept of mixed strategy in game theory. Although in some respect more natural, this approach is more technically demanding, as it requires an abstract construction by means of functional analysis tools (see section 7 and Bismut (1979)).

The connection between the two approaches is that any process in $\mathcal{V}^+$ can intuitively be viewed as the random distribution function of a randomized stopping time. Another interpretation is that each player chooses randomly, at each time t, whether to stop or not. This corresponds to the concept of behavioral strategy in game theory.

There is entensive literature in game theory, starting with Kuhn (1953), on the equivalence between mixed strategies and behavioral strategies. In discrete time, both notions are equivalent under fairly general assumptions (see Mertens, Sorin, and Zamir (1994)).

A by-product of section 7 is that, in the context of the simple game studied in this paper, behavioral strategies and mixed strategies are equivalent.

4. Proof of the main result. We prove the result by applying the following well-known min-max theorem.

THEOREM 4.1 (see Sion (1958)). *Let S and T be convex subsets of topological vector spaces, one of which is compact, and let $g : S \times T \longrightarrow \mathbb{R}$. Assume that for every real c, the sets $\{t : g(s_0, t) \leq c\}$ and $\{s : g(s, t_0) \geq c\}$ are closed and convex for every $(s_0, t_0) \in S \times T$. Then*

$$\sup_{s \in S} \inf_{t \in T} g(s, t) = \inf_{t \in T} \sup_{s \in S} g(s, t).$$

If S (resp., T) is compact, then sup *(resp.,* inf*) may be replaced by* max *(resp.,* min*), i.e., the corresponding player has an optimal strategy.*

The main difficulty in the proof of Theorem 3.1 is that the above min-max theorem does not apply directly to the set of strategies $\mathcal{V}^+$ (see the proofs of Lemmas 6.3 and 6.4). We therefore start by reducing the set of strategies to some subsets of $\mathcal{V}^+$ for which the min-max theorem applies.

We first restrict the strategies of the first player. Define

$$\mathcal{V}_1 := \left\{ F \in \mathcal{V}^+ : \ F \text{ is continuous, } P\text{-a.s. } \right\}.$$

As for the second player, we introduce the subset of strategies:

$$\mathcal{V}_2 := \left\{ G \in \mathcal{V}^+ : \ G_T = 1 \text{ on } \{Y_T < 0 < X_T\}, \text{ and } Y_T \Delta G_T \leq 0 \right\}.$$

We shall prove that the restriction of the strategies of Player 2 from $\mathcal{V}^+$ to $\mathcal{V}_2$ does not change the value of the game. The following is an intuitive justification of this claim. On the event set $\{Y_T < 0 < X_T\}$, it follows from the continuity of the payoff processes X and Y at T that it is optimal for Player 2 to stop the game before time T; recall that $Z_\cdot \leq Y_\cdot$, implying that the situation is even better for Player 2 if Player 1 stops at the same time. On the other hand, on the event set $\{Y_T > 0\}$, Player 2 can obtain the same value of the game by smoothing his strategy at time T, again taking advantage of the continuity at time T of the process Y.

Also, given that the strategies of Player 2 are restricted to $\mathcal{V}_2$, we shall prove that the restriction of the strategies of Player 1 to $\mathcal{V}_1$ does not change the value of the game; i.e., Player 1 can achieve the same value by means of continuous strategies.

For ease of presentation, the proof of the following two propositions will be reported in section 5.

PROPOSITION 4.1. *Let (X, Y, Z) be a triple of payoff processes satisfying* (2.1). *Then*

$$\sup_{F \in \mathcal{V}_1} \inf_{G \in \mathcal{V}_2} R(F, G) = \sup_{F \in \mathcal{V}_1} \inf_{G \in \mathcal{V}^+} R(F, G).$$

PROPOSITION 4.2. *Under the assumptions of Theorem* 3.1, *we have*

$$\inf_{G \in \mathcal{V}_2} \sup_{F \in \mathcal{V}_1} R(F, G) = \inf_{G \in \mathcal{V}_2} \sup_{F \in \mathcal{V}^+} R(F, G).$$

We then apply the min-max theorem to the strategy sets $S = \mathcal{V}_1$ and $T = \mathcal{V}_2$.

PROPOSITION 4.3. *Let (X, Y, Z) be a triple of processes satisfying* (2.1). *Assume further that X and Y are semimartingales. Then, we have*

$$\sup_{F \in \mathcal{V}_1} \inf_{G \in \mathcal{V}_2} R(F, G) = \inf_{G \in \mathcal{V}_2} \sup_{F \in \mathcal{V}_1} R(F, G).$$

The proof of the last proposition will be carried out in section 6. We now complete the proof of Theorem 3.1. By Proposition 4.2 and the fact that $\mathcal{V}_2 \subset \mathcal{V}^+$, we see that

$$\inf_{G \in \mathcal{V}_2} \sup_{F \in \mathcal{V}_1} R(F, G) = \inf_{G \in \mathcal{V}_2} \sup_{F \in \mathcal{V}^+} R(F, G) \ \geq \ \inf_{G \in \mathcal{V}^+} \sup_{F \in \mathcal{V}^+} R(F, G).$$

Similarly, it follows from Proposition 4.1 and the fact that $\mathcal{V}_1 \subset \mathcal{V}^+$ that

$$\sup_{F \in \mathcal{V}_1} \inf_{G \in \mathcal{V}_2} R(F, G) = \sup_{F \in \mathcal{V}_1} \inf_{G \in \mathcal{V}^+} R(F, G) \ \leq \ \sup_{F \in \mathcal{V}^+} \inf_{G \in \mathcal{V}^+} R(F, G).$$

In view of Proposition 4.3, this provides

$$\inf_{G \in \mathcal{V}^+} \sup_{F \in \mathcal{V}^+} R(F, G) \leq \sup_{F \in \mathcal{V}^+} \inf_{G \in \mathcal{V}^+} R(F, G),$$

which ends the proof, as the reverse inequality is trivial.

5. A priori restrictions on strategies. This section is devoted to the proofs of Propositions 4.1 and 4.2.

5.1. Proof of Proposition 4.1. Let F be a fixed strategy of Player 1 in the set $\mathcal{V}_1$. For each $G \in \mathcal{V}^+$, we define $\overline{G} \in \mathcal{V}_2$ by

$$\begin{aligned}
\overline{G}_T &= 1 \quad \text{on the event set } \{X_T > 0 > Y_T\},\\
\overline{G}_T &= G_{T-} \quad \text{on the event set } \{Y_T > 0\},\\
\overline{G} &= G \quad \text{otherwise.}
\end{aligned}$$

Then it is immediately checked that

$$\begin{aligned}
R(F,\overline{G}) - R(F,G) &= E\left[X_T(\Delta G_T - \Delta\overline{G}_T)\Delta F_T\right]\\
&\quad + E\left[Y_T(1-F_T)(\Delta\overline{G}_T - \Delta G_T)\right]\\
&\quad + E\left[Z_T\Delta F_T(\Delta\overline{G}_T - \Delta G_T)\right]\\
&= E\left[Y_T(1-F_T)(\Delta\overline{G}_T - \Delta G_T)\right]
\end{aligned}$$

since F is continuous. By definition of $\overline{G}$, we have $\Delta\overline{G}_T = 0$ on $\{Y_T > 0\}$ and $\Delta\overline{G}_T \geq \Delta G_T$ on $\{Y_T < 0\}$. It follows that $R(F,\overline{G}) - R(F,G) \leq 0$, and therefore

$$\sup_{F\in\mathcal{V}_1}\inf_{G\in\mathcal{V}_2} R(F,G) \leq \sup_{F\in\mathcal{V}_1}\inf_{G\in\mathcal{V}^+} R(F,G).$$

The required result follows from the fact that $\mathcal{V}_2 \subset \mathcal{V}^+$.

5.2. Proof of Proposition 4.2. We introduce the subset of strategies $\mathcal{W}_1$ defined by

$$\mathcal{W}_1 = \left\{F \in \mathcal{V}^+ : \quad \Delta F_T = 0 \ \text{ on } \{X_T > 0, Y_T \geq 0\}\right\}.$$

In order to prove Proposition 4.2, we first need to prove that the restriction of the strategies of Player 1 from $\mathcal{V}^+$ to $\mathcal{W}_1$ does not change the value of the game. As we shall see in the subsequent proof, this is a consequence of the continuity of the payoff processes X and Y at time T.

LEMMA 5.1. *Let (X,Y,Z) be a triple of processes satisfying (2.1). Assume further that X and Y have continuous trajectories at time T. Then, for any $G \in \mathcal{V}_2$ and $F \in \mathcal{V}^+$, there exists a sequence $(F^n)_n$ in $\mathcal{W}_1$ such that*

$$\limsup_{n\to\infty} R(F^n,G) \geq R(F,G).$$

Proof. We organize the proof in four steps.

Step 1. Let $\mathcal{T}_{[t,T]}$ denote the set of $[t,T]$-valued stopping times. We introduce the two Snell envelopes U and V defined by

$$\begin{aligned}
U_t &:= \operatorname{ess}\sup_{\zeta\in\mathcal{T}_{[t,T]}} E[X_\zeta|\mathcal{F}_t],\\
V_t &:= \operatorname{ess}\inf_{\zeta\in\mathcal{T}_{[t,T]}} E[Y_\zeta|\mathcal{F}_t].
\end{aligned}$$

In view of our assumptions on X and Y, the processes U and V can be considered in their càdlàg modifications; see, e.g., Appendix D in Karatzas and Shreve (1998). In

the rest of this step, we prove that

$$U \text{ and } V \text{ are continuous at } T, \quad P\text{-a.s.}$$

To see this, observe that

$$0 \leq U_t - E[X_T|\mathcal{F}_t] \leq E\left[\sup_{t\leq s\leq T} X_s - X_T|\mathcal{F}_t\right], \tag{5.1}$$

and, by Theorem VI.6 in Dellacherie and Meyer (1975),

$$E[X_T|\mathcal{F}_t] \longrightarrow E[X_T|\mathcal{F}_{T-}] = X_T \quad \text{as } t \nearrow T \tag{5.2}$$

by continuity of X at T. Now, notice that the process $A_t := \sup_{t\leq s\leq T} X_s - X_T$ is decreasing. Then, for fixed $s < T$, we have

$$0 \leq \limsup_{t\nearrow T} E[A_t|\mathcal{F}_t] \leq E[A_s|\mathcal{F}_{T-}].$$

By sending s to T, it follows from the dominated convergence theorem that

$$0 \leq \limsup_{t\nearrow T} E[A_t|\mathcal{F}_t] \leq E[A_T|\mathcal{F}_{T-}] = 0, \tag{5.3}$$

where we used the continuity of A at T inherited from X. The required continuity result follows from (5.1)–(5.3).

Step 2. For each $\varepsilon > 0$, define

$$\theta^\varepsilon := \inf\{t \geq T-\varepsilon : X_t \geq 0,\ U_t - \varepsilon \leq X_t \text{ and } V_t \geq -\varepsilon\} \wedge T.$$

Since X, U, and V are right-continuous, θ^ε is a stopping time. Observe that $\theta^\varepsilon \to T$, P-a.s., as $\varepsilon \to 0$.

Next, for each integer $n \geq 1$, define the sequence of stopping times

$$\theta^{\varepsilon,n} := T \wedge \left(\theta^\varepsilon + \frac{1}{n}\right).$$

We define $(F^{\varepsilon,n}) \in \mathcal{V}^+$ to be a continuous process on $(\theta^\varepsilon, \theta^{\varepsilon,n}]$ such that

$$F^{\varepsilon,n} = F \text{ on } [0,\theta^\varepsilon] \qquad \text{and} \qquad F^{\varepsilon,n} = 1 \text{ on } [\theta^{\varepsilon,n}, T].$$

Since X, Y, U, and V are continuous at T, $F^{\varepsilon,n}$ is a sequence in $\mathcal{W}_1$. We intend to prove that

$$\limsup_{\varepsilon\to 0} \limsup_{n\to\infty} R(F^{\varepsilon,n}, G) \geq R(F,G),$$

which will provide the required result.

First, since $F^{\varepsilon,n}$ is continuous on $(\theta^\varepsilon, T]$ and $F^{\varepsilon,n} = 1$ on $[\theta^{\varepsilon,n}, T]$, we have

$$\begin{aligned} R(F^{\varepsilon,n}, G) &= A + E\left[\xi^\varepsilon \int_{\theta^\varepsilon}^{T} X(1-G)dF^{\varepsilon,n} + Y(1-F^{\varepsilon,n})dG\right] \\ &= A + E\left[\xi^\varepsilon \int_{\theta^\varepsilon}^{\theta^{\varepsilon,n}} X(1-G)dF^{\varepsilon,n}\right] + E\left[\xi^\varepsilon \int_{\theta^\varepsilon}^{\theta^{\varepsilon,n}} Y(1-F^{\varepsilon,n})dG\right], \end{aligned}$$

where $\xi^\varepsilon = \mathbf{1}_{\{\theta^\varepsilon < T\}}$ and

$$A = E\left[\int_0^{\theta^\varepsilon} X(1-G)dF + Y(1-F)dG + \sum_{[0,\theta^\varepsilon]} Z\Delta F\Delta G\right].$$

Step 3. We now fix $\varepsilon > 0$ and let n go to infinity. As for the second expectation on the right-hand side of (5.4), observe that $Y_t\xi^\varepsilon(1-F_t^{\varepsilon,n})\mathbf{1}_{[\theta^\varepsilon,T]}(t)$ converges P-a.s. to zero for all $t \in (\theta^\varepsilon, T]$. Since G is right-continuous, this implies that $Y_t\xi^\varepsilon(1-F_t^{\varepsilon,n})\mathbf{1}_{[\theta^\varepsilon,T]}(t)$ converges $m_G \otimes P$-a.s. to zero. Therefore, by dominated convergence (see Theorem I.4.31 in Jacod and Shiryaev (1987)), we have

$$\lim_{n\to\infty} E\left[\xi^\varepsilon \int_{\theta^\varepsilon}^{\theta^{\varepsilon,n}} Y(1-F^{\varepsilon,n})dG\right] = 0.$$

As for the first expectation on the right-hand side of (5.4), we have

$$\begin{aligned}
\limsup_{n\to\infty} E\left[\xi^\varepsilon \int_{\theta^\varepsilon}^{\theta^{\varepsilon,n}} X(1-G)dF^{\varepsilon,n}\right] &\geq \limsup_{n\to\infty} E\left[\xi^\varepsilon \inf_{[\theta^\varepsilon,\theta^{\varepsilon,n}]}(X(1-G)) \int_{\theta^\varepsilon}^{\theta^{\varepsilon,n}} dF^{\varepsilon,n}\right] \\
&= \limsup_{n\to\infty} E\left[\xi^\varepsilon \inf_{[\theta^\varepsilon,\theta^{\varepsilon,n}]}(X(1-G))(1-F_{\theta^\varepsilon})\right] \\
&= E\left[\xi^\varepsilon X_{\theta^\varepsilon}(1-G_{\theta^\varepsilon})(1-F_{\theta^\varepsilon})\right],
\end{aligned}$$

where the last equality follows by dominated convergence and right-continuity of $X(1-G)$. This yields

$$\begin{aligned}
\limsup_{n\to\infty} R(F^{\varepsilon,n},G) - R(F,G) &\geq E\left[\xi^\varepsilon X_{\theta^\varepsilon}(1-G_{\theta^\varepsilon})(1-F_{\theta^\varepsilon})\right] \\
&\quad -E\left[\xi^\varepsilon \int_{\theta^\varepsilon}^{T} X(1-G)dF + Y(1-F)dG\right] \\
&\quad -E\left[\xi^\varepsilon \sum_{(\theta^\varepsilon,T]} Z\Delta F\Delta G\right] \\
&= E\left[\xi^\varepsilon X_{\theta^\varepsilon}(1-G_{\theta^\varepsilon})(1-F_{\theta^\varepsilon})\right] \\
&\quad -E\left[\xi^\varepsilon \int_{\theta^\varepsilon}^{T} X(1-G)dF + Y(1-F_-)dG\right] \\
&\quad +E\left[\xi^\varepsilon \sum_{(\theta^\varepsilon,T]} (Y-Z)\Delta F\Delta G\right] \\
&\geq E\left[\xi^\varepsilon X_{\theta^\varepsilon}(1-G_{\theta^\varepsilon})(1-F_{\theta^\varepsilon})\right] \\
&\quad -E\left[\xi^\varepsilon \int_{\theta^\varepsilon}^{T} X(1-G)dF + Y(1-F_-)dG\right],
\end{aligned}$$

where we used the condition $Z \leq Y$ of Theorem 3.1. Set $\widetilde{F} := F - \Delta F\mathbf{1}_{\{T\}}$ and $\widetilde{G}$

$:= G - \Delta G \mathbf{1}_{\{T\}}$. Then

$$\begin{aligned}\limsup_{n\to\infty} R(F^{\varepsilon,n}, G) - R(F,G) &\geq E\left[\xi^\varepsilon X_{\theta^\varepsilon}(1-G_{\theta^\varepsilon})(1-F_{\theta^\varepsilon}) - \xi^\varepsilon X_T(1-G_T)\Delta F_T\right] \\ &\quad - E\left[\xi^\varepsilon \int_{\theta^\varepsilon}^T X(1-G)d\widetilde{F} + Y(1-F_-)d\widetilde{G}\right] \\ &\quad - E\left[\xi^\varepsilon Y_T(1-F_T)\Delta G_T\right] \\ &\geq E\left[\xi^\varepsilon X_{\theta^\varepsilon}(1-G_{\theta^\varepsilon})(1-F_{\theta^\varepsilon}) - \xi^\varepsilon X_T(1-G_T)\Delta F_T\right] \\ &\quad - E\left[\xi^\varepsilon \int_{\theta^\varepsilon}^T X(1-G)d\widetilde{F} + Y(1-F_-)d\widetilde{G}\right] \qquad (5.4)\end{aligned}$$

since $Y_T \Delta G_T \leq 0$ by definition of $\mathcal{V}_2$.

Step 4. We now take limits as ε goes to zero. Since $\theta^\varepsilon \to T$, and both $\widetilde{F}$ and $\widetilde{G}$ are continuous at T, the second expectation on the right-hand side of (5.4) converges to zero. We now use the following claim, whose proof will be carried out later:

$$\xi^\varepsilon X_{\theta^\varepsilon} \longrightarrow \mathbf{1}_{\{0\leq X_T, Y_T\}} X_T, \quad P\text{-a.s.} \qquad (5.5)$$

Then, by dominated convergence and the fact that $G_{T-} \leq G_T$,

$$\begin{aligned}&\lim_{\varepsilon\to 0} E\left[\xi^\varepsilon X_{\theta^\varepsilon}(1-G_{\theta^\varepsilon})(1-F_{\theta^\varepsilon}) - \xi^\varepsilon X_T(1-G_T)\Delta F_T\right] \\ &\geq E\left[\mathbf{1}_{\{0\leq X_T,Y_T\}} X_T(1-G_T)(1-F_{T-}-\Delta F_T)\right] \\ &= E\left[\mathbf{1}_{\{0\leq X_T,Y_T\}} X_T(1-G_T)(1-F_T)\right] \\ &\geq 0\end{aligned}$$

by definition of F and G. Hence

$$\lim_{\varepsilon\to 0}\limsup_{n\to\infty} R(F^{\varepsilon,n}, G) - R(F,G) \geq 0.$$

It remains to prove (5.5). By definition of θ^ε, it is clear that θ^ε (hence also ξ^ε) increases as ε decreases to zero. Thus,

$$\xi^\varepsilon \longrightarrow \mathbf{1}_{\cap_{\varepsilon>0}\{\theta^\varepsilon < T\}}, \quad P\text{-a.s.}$$

Now, observe that $0 \leq X_T, 0 \leq Y_T$ on the event $\{\theta^\varepsilon < T \text{ for all } \varepsilon\}$ by continuity at T of the Snell envelopes U and V. Conversely, on the event $\{0 < X_T, 0 \leq Y_T\}$, it is clear that $\theta^\varepsilon < T$ for all ε, again by continuity of U and V. This provides claim (5.5). □

Given the result of Lemma 5.1, the statement of Proposition 4.2 follows directly from the following reduction of strategies of Player 1 from $\mathcal{W}_1$ to $\mathcal{V}_1$.

LEMMA 5.2. *Let (X,Y,Z) be a triple of processes satisfying (2.1). Assume further that X is a semimartingale and $Z \leq Y$. Then, for any $G \in \mathcal{V}_2$ and $F \in \mathcal{W}_1$, there exists a sequence (F^n) in $\mathcal{V}_1$ such that*

$$\limsup_{n\to\infty} R(F^n, G) \geq R(F,G).$$

Proof. For each integer n, define $\widetilde{F}^n \in \mathcal{V}^+$ by

$$\widetilde{F}^n_t = F_t - \sum_{s\leq t} \Delta F_s \mathbf{1}_{\{\Delta F_s \leq n^{-1}\}}$$

so that the jumps of $\widetilde{F}^n$ are of size greater than n^{-1}, and therefore $\widetilde{F}^n$ has a finite number of jumps. Clearly, we have the pointwise convergence

$$\widetilde{F}^n_t \longrightarrow F_t, \quad 0 \le t \le T, \quad P\text{-a.s.} \tag{5.6}$$

Since $\widetilde{F}^n$ has a finite number of jumps, it follows from a diagonal extraction argument that there exists a sequence of *continuous* processes $F^n \in \mathcal{V}^+$ such that $F^n - \widetilde{F}^n_- \to 0$ pointwise, P-a.s. From the pointwise convergence (5.6), this provides

$$F^n \longrightarrow F_-, \quad P\text{-a.s.}$$

In order to obtain the required result, we shall prove that

$$\lim_{n\to\infty} R(F^n, G) \ge R(F, G). \tag{5.7}$$

First, observe that by Itô's lemma (see, e.g., Theorem I.4.57 in Jacod and Shiryaev (1987)), we have

$$\begin{aligned} R(F,G) &= E\left[\int_0^T Y(1-F_-)dG\right] - E\left[\int_0^T F_- d(X(1-G))\right] \\ &\quad + E\left[\sum_{[0,T]} (Z-Y)\Delta F \Delta G\right] \\ &\quad + E\left[X_T(1-G_T)F_{T-}\right] + E\left[X_T(1-G_T)\Delta F_T\right] \\ &\le E\left[\int_0^T Y(1-F_-)dG\right] - E\left[\int_0^T F_- d(X(1-G))\right] \\ &\quad + E\left[X_T(1-G_T)F_{T-}\right] + E\left[X_T(1-G_T)\Delta F_T\right], \end{aligned} \tag{5.8}$$

where we used the condition $Z \le Y$ of Theorem 3.1. Since $F^n \to F_-$, m_G and $m_{X(1-G)}$-a.s., it follows from dominated convergence that

$$\lim_{n\to\infty} E\left[\int_0^T Y(1-F^n)dG\right] = E\left[\int_0^T Y(1-F_-)dG\right],$$

$$\lim_{n\to\infty} E\left[\int_0^T F^n d(X(1-G))\right] = E\left[\int_0^T F_- d(X(1-G))\right],$$

$$\lim_{n\to\infty} E\left[X_T(1-G_T)F^n_T\right] = E\left[X_T(1-G_T)F_{T-}\right].$$

In view of (5.8), and since F^n is continuous, this proves that

$$\lim_{n\to\infty} R(F^n, G) \ge R(F,G) - E\left[X_T(1-G_T)\Delta F_T\right].$$

Finally, observe that

$$\begin{aligned} X_T(1-G_T)\Delta F_T &\le X_T(1-G_T)\Delta F_T \mathbf{1}_{\{X_T>0\}}\mathbf{1}_{\{G_T<1\}} \\ &= X_T(1-G_T)\Delta F_T \mathbf{1}_{\{0<X_T, 0\le Y_T\}}\mathbf{1}_{\{G_T<1\}} \\ &= 0, \end{aligned}$$

where we used the fact that $G \in \mathcal{V}_2$ and $F \in \mathcal{W}_1$. This ends the proof of (5.7), and the proof of Lemma 5.2 is complete. □

Remark 5.1. In the last proof, we used for the first time the fact that X is a semimartingale. The reason is that we needed to apply integration by parts in the integral $\int_0^T X(1-G)dF$, and therefore we needed the stochastic integral with respect to process X to be well defined. Similar integration by parts are involved in the proofs of Lemmas 6.3 and 6.4, which then require the assumption that X and Y are semimartingales.

6. The value on restricted strategy spaces. This section is devoted to the proof of Proposition 4.3. As argued earlier, we shall apply Sion's theorem to the sets $S = \mathcal{V}_1$ and $T = \mathcal{V}_2$. We first define a suitable topology on $\mathcal{V}_1$ and $\mathcal{V}_2$.

Let $\mathcal{S}$ be the set of all $\mathbb{F}$-adapted processes Z satisfying $Z_{0-} = 0$ and

$$E\left[\int_0^T Z_t^2 dt + (\Delta Z_T)^2\right] < +\infty, \quad \text{where } \Delta Z_T = Z_T - \liminf_{t \nearrow T} Z_t.$$

The space $\mathcal{S}$ is a separable Hilbert space when endowed with the scalar product

$$\frac{1}{T+1} E\left[\int_0^T W_t Z_t dt + \Delta W_T \Delta Z_T\right].$$

Notice that $\mathcal{V}_1$ and $\mathcal{V}_2$ are convex subsets of $B_{\mathcal{S}}$, the unit ball of $\mathcal{S}$.

LEMMA 6.1. *The set $\mathcal{V}_2$ is compact for the weak topology $\sigma(\mathcal{S}, \mathcal{S})$.*

Proof. Since $B_{\mathcal{S}}$ is compact for the weak topology $\sigma(\mathcal{S}, \mathcal{S})$, it suffices to prove that $\mathcal{V}_2$ is closed for the weak topology or, equivalently, for the strong topology, by convexity.

Let (Z^n) be a sequence in $\mathcal{V}_2$, which converges strongly to some $Z \in \mathcal{S}$. Then, possibly along some subsequence,

$$Z^n \longrightarrow Z, \quad \lambda \otimes P\text{-a.s.}, \tag{6.1}$$

and

$$Z_T^n \longrightarrow Z_T, \quad P\text{-a.s.} \tag{6.2}$$

Clearly, this shows that Z inherits the nondecrease of (Z^n), $Z_{0-} = 0$, and $Z_T \le 1$. We now check that $\Delta Z_T^n \to \Delta Z_T$, P-a.s. By Fubini's theorem, it follows from (6.1) that, P-a.s., $Z_t^n \to Z_t$ for λ-a.e. $t \in [0,T]$. Since Z^n and Z are nondecreasing, we see that, P-a.s., $Z_{t-}^n \to Z_{t-}$ for every $t \in [0,T]$. Thus, from (6.2), this yields $\Delta Z_T^n \to \Delta Z_T$, P-a.s. The required result follows from the fact that $\Delta Z_T^n = 0$ on the event $\{Y_T > 0\}$. Observe finally that $Z_T^n = 1$ for every n implies $Z_T = 1$. □

LEMMA 6.2. *Let $(F^n)_n$ be a sequence in $\mathcal{V}_1$ converging to some $F \in \mathcal{V}_1$ in the sense of the strong topology of $\mathcal{S}$. Then*

$$\lim_{n\to\infty} F_t^n = F_t \qquad \textit{for all } \ t \in [0,T], \qquad P\textit{-a.s.}$$

after possibly passing to a subsequence.

Proof. Let (F^n) be as in the statement of the lemma. Then, by possibly passing to a subsequence, $F^n \longrightarrow F$, $\lambda \otimes P$-a.s., and $F_T^n \longrightarrow F_T$, P-a.s. By the same argument as in the previous proof, we use Fubini's theorem and the nondecrease of F_n and F

to see that $F^n_{t-} \longrightarrow F_{t-}$ for all $t \in [0,T]$, P-a.s. The required result follows from the continuity of F_n and F. □

LEMMA 6.3. *Let (X,Y,Z) be a triple of processes satisfying (2.1). Assume further that X is a semimartingale. Then, for all $G \in \mathcal{V}_2$, the function $R(.,G)$ is continuous on $\mathcal{V}_1$ in the sense of the strong topology of $\mathcal{S}$.*

Proof. By Itô's lemma,

$$\begin{aligned} X_T(1-G_T)F_T &= \int_0^T X(1-G_-)dF + \int_0^T F(1-G_-)dX - \int_0^T FXdG \\ &= \int_0^T X(1-G)dF + \int_0^T F(1-G_-)dX - \int_0^T FXdG \end{aligned}$$

since F is a continuous process. Then

$$\begin{aligned} R(F,G) = E\left[\int_0^T YdG\right] - E\left[\int_0^T F(1-G_-)dX\right] + E\left[X_T(1-G_T)F_T\right] \\ +E\left[\int_0^T (X-Y)FdG\right]. \end{aligned}$$

Let $(F^n)_n$ be a sequence in $\mathcal{V}_1$ converging to $F \in \mathcal{V}_1$. We intend to prove that

$$\lim_{n\to\infty} R(F^n,G) = R(F,G).$$

Consider any subsequence (F^{n_k}) such that $\lim_k R(F^{n_k},G)$ exists. It suffices to prove that this limit is independent of the subsequence and equal to $R(F,G)$. For ease of notation, rename the subsequence (F^n). From Lemma 6.2, by possibly passing to a subsequence, we can assume that, P-a.s.,

$$\lim_n F^n_t = F_t \quad \text{for all } t \in [0,T].$$

Then, $F^n \to F$, $m_X \otimes P$-a.s., and $m_G \otimes P$-a.s. and the result follows by dominated convergence. □

LEMMA 6.4. *Let (X,Y,Z) be a triple of processes satisfying (2.1). Assume further that Y is a semimartingale. Then, for all $F \in \mathcal{V}_1$, the function $R(F,.)$ is continuous on $\mathcal{V}_2$ in the sense of the strong topology of $\mathcal{S}$.*

Proof. As in the previous proof, let (G^n) be a sequence in $\mathcal{V}_2$ converging to $G \in \mathcal{V}_2$. We intend to prove that

$$\lim_{n\to\infty} R(F,G^n) = R(F,G).$$

Consider any subsequence (G^{n_k}) such that $\lim_k R(F,G^{n_k})$ exists. It suffices to prove that this limit is independent of the subsequence and equal to $R(F,G)$. For ease of notation, rename the subsequence (G^n). Recall that G is nondecreasing. Then, applying the same argument as in the proof of Lemma 6.1, we see that by possibly passing to a subsequence, we can assume that

$$G^n \longrightarrow G, \quad \lambda \otimes P\text{-a.s.}, \qquad G^n_- \longrightarrow G_-, \quad P\text{-a.s.}$$

and

$$G^n_T \longrightarrow G_T, \quad P\text{-a.s.}$$

Set $\widehat{Y} := Y(1-F)$. By Itô's formula and the continuity of F,

$$\int_0^T Y(1-F)dF^n = \widehat{Y}_T G_T^n - \int_0^T G^n d\hat{Y} + \sum_{0\le t\le T} \Delta Y_t (1-F_{t-})\Delta G_t^n$$
$$= \widehat{Y}_T G_T^n - \int_0^T G^n d\hat{Y}^c + \sum_{0\le t\le T} \Delta Y_t (1-F_t) G_{t-}^n.$$

Since F and $\widehat{Y}^c$ are continuous, $G^n \to G$, $m_F \otimes P$-a.s. and $m_{\widehat{Y}^c} \otimes P$-a.s, and the result follows by dominated convergence. □

Proof of Proposition 4.3. The strategy sets $S = \mathcal{V}_1$ and $T = \mathcal{V}_2$ are convex topological spaces when endowed with the weak topology $\sigma(\mathcal{S}, \mathcal{S})$. From Lemma 6.1, $\mathcal{V}_2$ is compact for $\sigma(\mathcal{S}, \mathcal{S})$.

Since $R(F, G)$ is bilinear, the sets $\{G \in \mathcal{V}_2 : R(F^0, G) \le c\}$ and $\{F \in \mathcal{V}_1 : R(F, G^0) \ge c\}$ are convex for all $F^0 \in \mathcal{V}_1$, $G^0 \in \mathcal{V}_2$, and $c \in \mathbb{R}$. Then in order to prove that they are closed for the weak topology $\sigma(\mathcal{S}, \mathcal{S})$, it suffices to prove that they are closed for the strong topology of $\mathcal{S}$. The latter is a direct consequence of Lemmas 6.3 and 6.4. We are then in the context of Sion's theorem, and the proof is complete. □

7. Extended problem and randomized stopping times. In this section, we first provide a justification of $\mathcal{V}^+$ as being the natural mixed strategy set, which has been described heuristically in section 3. Then we derive rigorously the payoff function $R(F, G)$ defined in the extended strategy set $\mathcal{V}^+ \times \mathcal{V}^+$.

For ease of exposition, we shall discuss the case $Z = Y$ only. The general case follows immediately by adding up the jump term induced by Z.

In game theory, mixed strategies are defined as probability distributions over pure strategies. In the context of Dynkin games, pure strategies are stopping times. At this stage, the main problem is to define a measurable structure on the set of stopping times. There are two ways to avoid this difficulty. Following Aumann (1964), one may define mixed strategies by enlarging the probability space; this viewpoint is discussed in section 7.1. An alternative approach consists of defining the notion of randomized stopping time by means of functional analysis arguments; this is discussed in section 7.2. We shall (essentially) show that $\mathcal{V}^+$ is in one-to-one correspondence with the set of mixed strategies and with the set of randomized stopping times. Therefore, both approaches are equivalent.

7.1. Mixed strategies. We enlarge the probability space from (Ω, P) to $([0,1] \times \Omega, \lambda_1 \otimes P)$, where λ_1 is the Lebesgue measure. A mixed strategy (for Player 1) is then defined as a $\lambda_1 \otimes P$ measurable function ϕ mapping $[0,1] \times \Omega$ into $[0,T]$ such that

$$\text{for } \lambda_1\text{-a.e., } r \in [0,1], \qquad \sigma_r := \phi(r, \cdot) \text{ is a stopping time.}$$

We denote by Φ the space of mixed strategies. Loosely speaking, $([0,1], \lambda_1)$ is a randomizing device for Player 1. In order to introduce an independent randomizing device for Player 2, we need to have an independent copy $([0,1], \lambda_2)$ of the probability space $([0,1], \lambda_1)$. The corresponding set of mixed strategies is denoted by Ψ; a generic element of Ψ will be denoted by ψ, and, for $r \in [0,1]$, we set $\tau_r := \psi(r, \cdot)$.

Hence, the underlying probability space for the extended Dynkin game is $([0,1] \times [0,1] \times \Omega, \lambda_1 \otimes \lambda_2 \otimes P)$.

Recall that the payoff function on the stopping times is denoted by $\widetilde{R}$, and its extension to $\mathcal{V}^+$ is denoted by R. The following result provides a justification of the

definition of $\mathcal{V}^+$ as the set of mixed strategies, and R as the payoff function on the extended strategy sets.

PROPOSITION 7.1. (i) *There exists a mapping H from Φ (or Ψ) onto $\mathcal{V}^+$.*

(ii) *For every $(\phi,\psi)\in\Phi\times\Psi$, we have*

$$E_{\lambda_1\otimes\lambda_2}\left[\widetilde{R}(\sigma,\tau)\right]=R(H(\phi),H(\psi)).$$

Proof. We only prove (i) for the set Φ. For $\phi\in\Phi$, define the process $H(\phi)$ by

$$H(\phi)_t=\int\mathbf{1}_{\{\sigma_r\le t\}}\lambda_1(dr)\;=\;E_{\lambda_1}[\mathbf{1}_{\{\sigma\le t\}}]\quad\text{for } t\in[0,T].$$

Clearly, $H(\phi)_{0-}=0$, $H(\phi)$ is nondecreasing, right-continuous and $H(\phi)_T\le 1$. Since σ_r is a stopping time for λ_1-a.e., $r\in[0,1]$, the process $H(\phi)$ is $\mathbb{F}$-adapted. This proves that $H(\phi)\in\mathcal{V}^+$. To see that H is onto, define

$$\phi^F(r,\omega):=\inf\{s\ge 0:\;F_s(\omega)>r\}\quad\text{for } F\in\mathcal{V}^+.$$

Observe that $\phi^F\in\Phi$, since F is $\mathbb{F}$-adapted and right-continuous. Set $\sigma_r:=\phi^F(r,\cdot)$. For $t\in[0,T]$, we compute

$$H(\phi^F)_t=\int\mathbf{1}_{\{\sigma_r\le t\}}\lambda_1(dr)\;=\;\int\mathbf{1}_{\{F_t\ge r\}}\lambda_1(dr)\;=\;F_t,$$

which concludes the proof of (i).

Let $(\phi,\psi)\in\Phi\times\Psi$, and set $F_t=\mathbf{1}_{\{\sigma\le t\}}$ and $G_t=\mathbf{1}_{\{\tau\le t\}}$. By Fubini's theorem,

$$E_{\lambda_1\otimes\lambda_2\otimes P}\left[\int_0^T X(1-G)dF\right]=E_{\lambda_1\otimes P}\left[\int_0^T X(1-H(\psi))dF\right].$$

By Itô's lemma, this provides

$$\begin{aligned}E_{\lambda_1\otimes\lambda_2\otimes P}\left[\int_0^T X(1-G)dF\right]&=E_{\lambda_1\otimes P}\left[X_T(1-H(\psi)_T)F_T-\int_0^T F_-d(X(1-H(\psi))\right]\\&=E_P\left[X_T(1-H(\psi)_T)H(\phi)_T-\int_0^T H(\phi)_-d(X(1-H(\psi))\right],\end{aligned}$$

where we again used Fubini's theorem. By another application of Itô's lemma, we get

$$E_{\lambda_1\otimes\lambda_2\otimes P}\left[\int_0^T X(1-G)dF\right]=E_P\left[\int_0^T X(1-H(\psi))dH(\phi)\right].$$

The same argument applies to the second integral $\int_0^T Y(1-F_-)dG$. Hence,

$$\begin{aligned}E_{\lambda_1\otimes\lambda_2}\left[\widetilde{R}(\sigma,\tau)\right]&=E_P\left[\int_0^T X(1-H(\psi))dH(\phi)+Y(1-H(\phi)_-)dH(\psi)\right]\\&=R(H(\phi),H(\psi)).\qquad\square\end{aligned}$$

7.2. Randomized stopping times. In this section, we describe briefly the functional analysis approach in order to define the notion of randomized stopping times introduced by Bismut (1979). We shall recall a representation theorem which connects randomized stopping times to our set $\mathcal{V}^+$.

Let $\mathcal{Y}$ be the space of càdlàg optional processes Y defined on $[0,T]$ such that

$$E\left[\sup_{t\in[0,T]}|Y_t|\right] < +\infty. \tag{7.1}$$

Observe that $\mathcal{Y}$ is a Banach space when endowed with the norm defined by (7.1). We denote by $\mathcal{Y}'$ the dual space of $\mathcal{Y}$. Then we have the following representation result of elements of $\mathcal{Y}'$.

PROPOSITION 7.2 (see Bismut (1979)). *For any $\mu \in \mathcal{Y}'$, there exist two right-continuous adapted processes with finite variation A and B valued in $\mathbb{R}\cup\{+\infty\}$ such that*

$$\langle \mu, Y\rangle \;=\; E\left[\int_0^T Y dA + Y_- dB\right] \quad \textit{for all } Y \in \mathcal{Y}.$$

Proposition 1.3 in Bismut (1979) provides a uniqueness result for such a representation under further restrictions on A and B.

DEFINITION 7.1. *A randomized stopping time is an element $\mu \in \mathcal{Y}'$, for which there exists a representation with $B = 0$, A nondecreasing and $A_T \le 1$.*

The following easy consequence establishes the connection between our set of extended strategies $\mathcal{V}^+$ and the set of randomized stopping times.

COROLLARY 7.1. *There is a bijection between $\mathcal{V}^+$ and the set of randomized stopping times.*

Proof. To every randomized stopping time μ, we can associate $A \in \mathcal{V}^+$ by the above representation. Conversely, given $A \in \mathcal{V}^+$, it is easy to check that $Y \mapsto E\int_0^T Y dA$ belongs to $\mathcal{Y}'$. □

Acknowledgments. The authors wish to thank Eran Shmaya and Eilon Solan for helpful comments.

REFERENCES

M. ALARIO-NAZARET, J.P. LEPELTIER, AND B. MARCHAL (1982), *Dynkin games*, in Stochastic Differential Systems (Bad Honnef, 1982), Lecture Notes in Control Inform. Sci. 43, Springer-Verlag, Berlin, New York, pp. 23–32.

R.J. AUMANN (1964), *Mixed and behavior strategies in infinite extensive games*, in Advances in Game Theory, Ann. Math. Stud. 52, M. Dresher, L.S. Shapley, and A.W. Tucker, eds., Princeton University Press, Princeton, NJ.

A. BENSOUSSAN AND A. FRIEDMAN (1974), *Non-linear variational inequalities and differential games with stopping times*, J. Funct. Anal., 16, pp. 305–352.

J.-M. BISMUT (1977), *Sur un problème de Dynkin*, Z. Warsch. V. Geb., 39, pp. 31–53.

J.-M. BISMUT (1979), *Temps d'arrêt optimal, quasi-temps d'arrêt et retournement du temps*, Ann. Probab., 6, pp. 933–964.

J. CVITANIĆ AND I. KARATZAS (1996), *Backward stochastic differential equation with reflection and Dynkin games*, Ann. Probab., 24, pp. 2024–2056.

C. DELLACHERIE AND P.A. MEYER (1975), *Probabilités et potentiel*, Hermann, Paris.

E.B. DYNKIN (1967), *Game variant of a problem on optimal stopping*, Soviet Math. Dokl., 10, pp. 270–274.

E.B. DYNKIN AND A.A. YUSHKEVICH (1968), *Theorems and Problems in Markov Processes*, Plenum Press, New York.

J. Jacod and A.N. Shiryaev (1987), *Limit Theorems for Stochastic Processes*, Springer-Verlag, Berlin.

I. Karatzas and S. Shreve (1998), *Methods of Mathematical Finance*, Springer-Verlag, New York.

I. Karatzas and H. Wang (2001), *Connections between bounded variation control and Dynkin games*, in Optimal Control and Partial Differential Equations (Volume in honor of A. Bensoussan), J.L. Menaldi, E. Rofman, and A. Sulem, eds., IOS Press, Amsterdam, pp. 363–373.

H.W. Kuhn (1953), *Extensive games and the problem of information*, in Contributions to the Theory of Games, Ann. Math. Stud. 28, H.W. Kuhn and A.W. Tucker, eds., Princeton University Press, Princeton, NJ.

J.P. Lepeltier and M.A. Maingueneau (1984), *Le jeu de Dynkin en théorie générale sans l'hypothèse de Mokobodsky*, Stochastics, 13, pp. 25–44.

J.-F. Mertens, S. Sorin, and S. Zamir (1994), *Repeated Games, Part A*, Core Discussion Papers 9420, Université Catholique de Louvain.

H. Morimoto (1984), *Dynkin games and martingale methods*, Stochastics, 13, pp. 213–228.

J. Neveu (1975), *Discrete Parameter Martingales*, North Holland, Amsterdam.

M. Sion (1958), *On general minimax theorems*, Pacific J. Math., 8, pp. 171–176.

L. Stettner (1982), *Zero-sum Markov games with stopping and impulsive strategies*, Appl. Math. Optim., 9, pp. 1–24.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 4, pp. 1089–1117

THE SOLUTION OF THE H^2/H^∞ PROBLEM BY DIRECT METHODS*

M. A. DA SILVEIRA† AND R. ADES‡

Abstract. The H^2/H^∞ problem is formulated in a Hilbertian context. It has a unique solution, which is the strong limit of sequences generated by Galerkin methods based on convenient, but not necessarily orthogonal, generator sets. Using these results, a methodology to solve the problem by a Galerkin method is proposed, and an example is solved and compared to other approaches.

Key words. optimal control, robust control, H^2/H^∞ problem, linear control systems, weighted Hardy spaces

AMS subject classifications. 49J02, 34H05, 41A20, 65D02

PII. S0363012900367618

1. Introduction. The simplest H^2/H^∞ problem is to find a function $K(.)$ in the Hardy class H_+^∞ minimizing the quadratic criterion

$$J[K(.)] = \int_{-\infty}^{\infty} \{K(-i\omega)\Gamma(i\omega)K(i\omega) + K(-i\omega)\gamma(i\omega)\}d\omega,$$

under the H^∞ constraint

$$\operatorname{ess\,sup}|A(i\omega)K(i\omega) + B(i\omega)| \leq \lambda,$$

where $\Gamma(.)$, $\gamma(.)$, $A(.)$, and $B(.)$ are known rational functions, λ is a given positive real number, and the essential supremum is taken on the set of real numbers ω. This problem arises in quadratic optimal control theory for linear systems when robustness conditions or filtering constraints are imposed on the controller. This paper shows that, in spite of the H^∞ constraint, the above optimal control problem is well-posed in a larger space, $H_+^{2,-1}$, a Hilbert space to be defined here. It means that the optimal control problem has a unique solution in this space with desired regularity properties, under suitable conditions on the functional $J[.]$. Moreover, this functional setting leads to the definition of generator sets such that Galerkin methods converge to the optimal control problem solution. A significant remark is that it is possible to measure the approached solution quality when the proposed method is coupled with the dual method presented in [1]. The design of a pitch optimal control of a fighter airplane is presented as an example to show the numerical viability and to allow for comparison with other design methods.

The crucial point in this paper is the construction of a Hilbert space containing the usual Hardy spaces H_+^2, H_+^∞ such that bounded and closed sets in both spaces are also bounded closed in this new space. The embedding of the original problem in this new setting allows the use of Hilbert space convex optimization tools to solve

*Received by the editors February 11, 2000; accepted for publication (in revised form) February 28, 2002; published electronically October 8, 2002. This research was partially supported by CNPq (Brazilian Research Council).

http://www.siam.org/journals/sicon/41-4/36761.html

†Electrical Engineering Department, PUC-Rio, R. Marquês de São Vicente, 225, 22453-900, Rio de Janeiro, RJ, Brazil (marcos@ele.puc-rio.br).

‡Electrical Engineering Department, IME, Praça General Tibúrcio, 80, 22290-270, Rio de Janeiro, RJ, Brazil (rades@epq.ime.eb.br).

the problem. The H_+^∞ constraint carries the optimal solution into this last space with no further considerations about the H_+^∞ non-Hilbertian topology. Besides, this construction will be necessary to build a chain of Hilbert spaces needed to represent the optimal solution regularity, which is essential information for the Galerkin method convergence properties.

In the remainder of this section, a survey of the H^2/H^∞ problem and the notation to be used will be presented. The geometry of the Hilbert spaces $H_+^{2,-1}$ and $H_+^{2,-k}$ is presented in section 2. The unconstrained H^2 optimal control problem is rewritten in section 3 as a minimum norm problem in a suitable space $H_+^{2,-k}$ according to the data; this clarifies its existence and regularity properties. The constrained H^2/H^∞ optimal control problem is solved in Theorem 7 of section 4, which states the existence and uniqueness results cited above. The convergence of Galerkin methods is the subject of section 5, and a numerical example is presented and discussed in section 6. Some extensions of those results are shown in the last section, particularly to multivariable problems. All proofs which are not in the main text can be found in Appendix A.

After the introduction of the Youla–Kučera parameterization [2], [3], quadratic criteria for Wiener–Hopf linear-quadratic optimal control problems have been considered, allowing the manipulation of well-defined technical or physical optimal solution characteristics, as rms transient error, plant saturation and closed-loop sensitivity [2], transient specifications against shape-deterministic exogenous inputs [4], performance measures [5], [6], [7], servomechanism specifications [8], and transient specifications [9]. The work in [9] presents a heuristic procedure to choose the criteria weighting filters in such a way that a trade-off between overshoot and time constant can be obtained. All these papers consider the controller set as an optimization variable, with the set of controllers being parameterized by real-rational proper stable rational matrices. Explicit expressions for the optimal solutions are derived.

These linear-quadratic criteria have been enriched by quadratic of H^∞ constraints to consider performance or robustness conditions in [6], [7], [10], and [11]. In particular, H^∞ constraints have been used to impose a prespecified robustness degree to the optimal solution (see [12]), but they can be used also to impose other specifications, such as filter constraints (see [13]).

Other methods have been proposed to solve the H^2/H^∞ problem as well. Some of these methods modify the original optimization criteria to obtain new mathematical problems but obscuring the original physical interpretation. Examples are the methods described in [5], [10], and [11]. Direct methods, using expansions in series, do not modify the original criteria; they were proposed in [14] in a different context, and in [15], [16], [17], and [18]. Reference [17] considers discrete-time linear systems, presenting an algorithm to solve an approximate version such that the optimal parameter is exponentially stable. The other references consider the continuous-time case by using Laguerre functions as a generator set but do not prove the existence of an optimal solution. Reference [18] addresses the existence proof discussed in [19]. However, [19] does not contain such a proof but simply states that "it is easy to show that [the quadratic functional] has a unique minimum h_* on [the constraint set]". It is worth noting that these papers assume the Youla parameter in the usual quadratic Hardy space, a functional space not containing biproper rational functions. Moreover, [18] presents an extension of the algorithm already proposed in [17]. These assumptions are addressed in the last section. In [20], the use of linear matrix inequalities (LMIs) is proposed to solve the H^2/H^∞ problem but under assumptions too restrictive and unnatural.

A recent methodology was presented in [1] where a sequence of H^2 constraints approaching the original H^∞ constraint was built. In this method, each H^2 constraint defines a pure H^2 problem solved by a dual problem whose solution is explicitly given. The present paper shows that this solution defines a lower bound to the original optimal cost, with the sequence of these solutions monotonically approaching the optimal solution, when they exist. Such an algorithm will be used here as a part of a methodology to establish lower bounds to the optimal criterion value.

Actually, it is possible to obtain only approximate solutions. Indeed, paper [19] presents a theorem stating that the optimal solution is infinite-dimensional when the H^∞ constraint is active, which forces the designer to find rational approximations to the optimal solution. Moreover, [19] shows that the optimal parameter cannot be exponentially stable.

A first explicit proof for the existence and uniqueness of the H^2/H^∞ problem solution was given by the authors in [21], searching for the solution in the space generated by completing the set of real-rational proper stable functions under the norm defined by the quadratic criterion term. This result was further developed in [22], allowing a complete methodology to solve the H^2/H^∞ problem without changes in the criteria and in the constraints other than projections on finite-dimensional spaces. This methodology will, in part, be shown here. The present paper develops a more complete mathematical theory for the problem, determining the existence, uniqueness, and regularity to the solutions under natural assumptions and proving the convergence of the Galerkin approximating sequence to the optimal solution.

Notations. Let $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}$, and $\mathbb{C}$ denote the natural numbers (i.e., the positive integers), the integers, the real, and the complex numbers, respectively. Also, let $|s|$, $\bar{s}$, and $\mathrm{Re}\{s\}$ denote the modulus, the conjugate, and the real part of a complex number s, respectively.

With i denoting $\sqrt{-1}$, let $i\mathbb{R} = \{i\omega, \omega \in \mathbb{R}\}$, $C_+^0 = \{s \in \mathbb{C}\colon \mathrm{Re}\{s\} > 0\}$, and $C_-^0 = \{s \in \mathbb{C}\colon \mathrm{Re}\{s\} < 0\}$. The functions $f : A \to B$ are denoted as f, $f(.)$, or $f(s)$, with $f(s)$ also denoting its value at $s \in A$. A function $f(.)$ is real if it maps real numbers in real numbers.

If $f(.) = n(.)/d(.)$ is rational, with $n(.)$ and $d(.)$ being polynomials, $\partial_r(f)$ denotes its relative degree, defined as the integer "degree of $d(.)$ − degree of $n(.)$." Also $f^*(s) = f(-s)$, $|f(i\omega)|^2 = \bar{f}(i\omega)f(i\omega)$ (or $f^*(i\omega)f(i\omega)$ if $f(.)$ is real).

The usual inner product and the usual quadratic norm are defined by

$$\langle f, g\rangle_2 = \int_{-\infty}^{\infty} \bar{f}(i\omega)f(i\omega)d\omega \qquad \text{and} \qquad \|f\|_2 = [\langle f, f\rangle_2]^{1/2} = \left[\int_{-\infty}^{\infty} |f(i\omega)|^2 d\omega\right]^{1/2}.$$

The H^∞ norm is defined by

$$\|f\|_\infty = \mathrm{ess\,sup}|f(i\omega)|,$$

the supremum taken on $\omega \in \mathbb{R}$. The symbols R_m, R_m^+, and R_m^- denote the classes of rational functions with relative degree greater than or equal to m, without poles in $i\mathbb{R}$, in $i\mathbb{R} \cup C_+^0$ (stable functions), and in $i\mathbb{R} \cup C_-^0$ (completely unstable functions), respectively.

The symbols H_+^2, H_-^2, H_+^∞, and H_-^∞ represent the usual Hardy classes studied in [23], [24]. The principal features of these spaces to be used here are given in what follows. The two spaces of stable functions are defined by

$$H_+^2 = \{f\colon C_+^0 \to \mathbb{C} \text{ analytic in } C_+^0 : \exists\, M < \infty \text{ with } \|f(a + i\omega\|_2 < M\ \forall\ a > 0\},$$

$$H_+^\infty = \{f\colon C_+^0 \to \mathbb{C} \text{ analytic and bounded in } \mathbb{C}_+^0\},$$

H_-^2 and H_-^∞ defined analogously by changing the symbol "+" by "−" and assuming $a < 0$.

Also, if

$$L^2(i\mathbb{R}) = \{f : i\mathbb{R} \to \mathbb{C} : \|f\|_2 < \infty\},$$

it can be proved that $\langle f, g\rangle_2$ is an inner product in $L^2(i\mathbb{R})$, H_+^2, and H_-^2, with all these spaces being Hilbert spaces under this inner product. The functional spaces H_+^2 and H_-^2 can be identified to be orthogonal subspaces of $L^2(i\mathbb{R})$ so that $L^2(i\mathbb{R}) = H_+^2 \oplus H_-^2$ (an orthogonal sum of subspaces). The symbols $[f]_+$ and $[f]_-$ denote the orthogonal projection of $f \in L^2(i\mathbb{R})$ in H_+^2 and H_-^2, respectively, and H_+^∞, H_-^∞ are Banach spaces under the norm $\|.\|_\infty$.

The symbol A_0 denotes the subset of H_+^∞-functions continuous on the completed imaginary axis. The symbol $\hat{A}(\beta_1)$ denotes the class of Laplace transforms of distributions in the Callier–Desoer algebra $A(\beta_1)$, and $\hat{A}_-(\beta)$ denotes the set of functions belonging to $\hat{A}(\beta_1)$ for some $\beta_1 < \beta$.

If H represents a locally convex topological vector space [25], H' denotes its topological dual endowed with its strong topology. If H and V are such spaces, $H + V$ and $H \oplus V$ denote their sums and their direct sums, respectively. The latter means $H \cap V = \{0\}$, the trivial subspace. Further information about these concepts can be found in [25], [26], or [27].

2. A functional setting for the optimal control problem. This section presents the functional setting to formulate the H^2/H^∞ problem as a well-posed problem. The basic idea is to define spaces containing H_+^2 and H_+^∞ such that the quadratic functional to be minimized is continuous and the constraints convex, closed, and bounded. Actually, a chain of spaces like H_+^2 will be defined to utilize the problem regularity.

DEFINITION 1. *Let* $\Phi_{-k} = (s+1)^{-k}$, $k \in \mathbb{Z}$, *let*

$$\langle f, g\rangle_{2,-k} = \int_{-\infty}^{\infty} f^*(i\omega)\Phi_{-k}^*(i\omega)\Phi_{-k}(i\omega)f(i\omega)d\omega,$$

and let

$$\|f\|_{2,-k} = [\langle f, f\rangle_{2,-k}]^{1/2} = \left[\int_{-\infty}^{\infty} |\Phi_{-k}(i\omega)f(i\omega)|^2 d\omega\right]^{1/2}.$$

Set

$$L_{-k}^2(i\mathbb{R}) = \{f : i\mathbb{R} \to \mathbb{C} : \|f\|_{2,-k} < \infty\},$$
$$H_+^{2,-k} = \{f : C_+^0 \to \mathbb{C} \text{ is analytic in } C_+^0 : \exists\, M < \infty \text{ such that } \|f(a+i\omega)\|_{2,-k} < M \ \forall\ a > 0\},$$

$H_-^{2,-k}$ *the analogous space using* $a < 0$ *and* C_-^0 *in its definition.*

As $\langle f, g\rangle_{2,-k} = \langle \Phi_{-k}f, \Phi_{-k}g\rangle_2$ and $\|f\|_{2,-k} = \|\Phi_{-k}f\|_2$, it is easy to prove that $\langle f, g\rangle_{2,-k}$ defines an inner product and $\|f\|_{2,-k}$ is the associated norm in the spaces defined above (see Appendix A). Moreover, the usual $L^2(i\mathbb{R})$, H_+^2, and H_-^2 spaces are the special cases where $k = 0$. The next theorem presents the geometrical properties of the spaces defined here.

THEOREM 1. *Let the spaces* $L_{-k}^2(i\mathbb{R})$, $H_+^{2,-k}$, $H_-^{2,-k}$ *be as in Definition* 1.

(a) $L^2_{-k}(i\mathbb{R})$, $H^{2,-k}_+$, $H^{2,-k}_-$ *are the completion of the sets* R_{1-k}, R^+_{1-k}, *and* R^-_{1-k} *in the norm* $\|.\|_{2,-k}$, *respectively. Moreover, they are Hilbert spaces with respect to the corresponding inner product.*

(b) $H^{2,-k}_+$ *and* $H^{2,-k}_-$ *are closed subspaces of* $L^2_{-k}(i\mathbb{R})$.

(c) $L^2_{-k}(i\mathbb{R}) = H^{2,-k}_+ + H^{2,-k}_-$. *If* $k \le 0$, *then* $H^{2,-k}_+ \cap H^{2,-k}_-$ *is empty. If* $k \ge 1$, *then* $H^{2,-k}_+ \cap H^{2,-k}_-$ *contains the polynomials in* s *with degree less than or equal to* $k-1$ *and the functions defined by* $\sum_{m=1}^{\infty} \alpha_m e^{-st_m}$, *where* $\sum_{m=1}^{\infty} |\alpha_m| < \infty$ *and, for any* m, $t_m > 0$.

Remark 1. It is worth noting that a rational function $f(s)$ without poles in $i\mathbb{R}$ belongs to $L^2_{-k}(i\mathbb{R})$ if and only if $\partial_r(f) \ge 1-k$. Alternatively, if $\partial_r(f) = m$ and $f(s)$ has no poles in $i\mathbb{R}$, then $f(s) \in L^2_{-k}(i\mathbb{R})$ for each $k \ge 1-m$.

The next theorem collects some results relating the topologies of H^∞_+, $L^2_{-k}(i\mathbb{R})$, and $H^{2,-k}_+$ for different indexes k.

THEOREM 2. *Let the spaces* $L^2_{-k}(i\mathbb{R})$, $H^{2,-k}_+$, $H^{2,-k}_-$ *be as in Definition* 1 *and let* $k < m$.

(a) $L^2_{-k}(i\mathbb{R}) \subset L^2_{-m}(i\mathbb{R})$. *The linear spaces* $L^2_{-k}(i\mathbb{R})$ *and* $L^2_{-m}(i\mathbb{R})$ *are isometrically isomorphic, the isometry from* $L^2_{-k}(i\mathbb{R})$ *to* $L^2_{-m}(i\mathbb{R})$ *being injective and the inverse isometry being surjective. Therefore, the* $L^2_{-k}(i\mathbb{R})$ *topology is strictly finer than the* $L^2_{-m}(i\mathbb{R})$ *topology.*

(b) $H^{2,-k}_+ \subset H^{2,-m}_+$. *The* $H^{2,-k}_+$ *topology is strictly finer than the* $H^{2,-m}_+$ *topology.*

(c) $L^2_{-k}(i\mathbb{R})$ *is dense in* $L^2_{-m}(i\mathbb{R})$, $H^{2,-k}_+$ *is dense in* $H^{2,-m}_+$. *In particular, if* $k \ge 1$, *the sets* R_0, R^+_0, *and* R^-_0 *are dense in* $L^2_{-k}(i\mathbb{R})$, $H^{2,-k}_+$, *and* $H^{2,-k}_-$, *respectively.*

(d) $H^\infty_+ \subset H^{2,-1}_+$, *the* H^∞_+ *topology being strictly finer than the one of* $H^{2,-1}_+$.

Remark 2. Property (c) says that biproper rational functions can be approached in $H^{2,-1}_+$ by strictly proper rational functions, diminishing the relative degree at the limit. As an example, $f_n(s) = n(s+n)^{-1}$ converges to the constant function $f(s) \equiv 1$ in the $H^{2,-1}_+$ topology. This explains why it is possible to find complete sets for $H^{2,-k}_+$, $k \ge 1$, formed by strictly proper real-rational stable functions (R^+_0 functions).

Remark 3. Let $S(i\mathbb{R})$ denote the space of functions going quickly to zero at infinity and $(S(i\mathbb{R}))'$ its topological dual (the space of temperate distributions) [26], [27]. Define $S_+(i\mathbb{R})$ as $S(i\mathbb{R}) \cap H^2_+$ and $(S_+(i\mathbb{R}))'$ as its closure in the $(S(i\mathbb{R}))'$ topology. With these notations it is possible to prove that, for any $k > 1$,

$$
\begin{array}{lllllll}
S(i\mathbb{R}) \subset & L^2_k(i\mathbb{R}) \subset & L^2_1(i\mathbb{R}) \subset & L^2(i\mathbb{R}) \subset & L^2_{-1}(i\mathbb{R}) \subset & L^2_{-k}(i\mathbb{R}) \subset & (S(i\mathbb{R}))' \\
\cup & \cup & \cup & \cup & \cup & \cup & \cup \\
S_+(i\mathbb{R}) \subset & H^{2,k}_+ \subset & H^{2,1}_+ \subset & H^2_+ \subset & H^{2,-1}_+ \subset & H^{2,-k}_+ \subset & (S_+(i\mathbb{R}))' \\
 & & & & \cup & & \\
 & & & & H^\infty_+. & &
\end{array}
$$

Each space is dense in the next bigger one in the chain. An analogous sequence can be built for unstable functions spaces. The original H^2/H^∞ problem will be embedded in these chains of Hilbert spaces, as will be shown in the next section.

Remark 4. Let $\mathbb{H}_k$ denote the order k Sobolev space [26], [27]. As $\mathbb{H}_k$ is the Fourier transform image of $L^2_k(i\mathbb{R})$ (by an adaptation of a construction found in [26]), it is possible to define stable Sobolev spaces $[\mathbb{H}_k]^+$ as the inverse Fourier transform image of H^k_+. Then, it is possible to build a corresponding sequence of stable Sobolev

spaces also beginning in $S_+(i\mathbb{R})$ and ending in $(S_+(i\mathbb{R}))'$. Also, from the structure theorem (see [26, p. 255]), it is possible to show that the temperate distributions in $(S_+(i\mathbb{R}))'$ are derivatives of some finite order of functions in $[\mathbb{H}_2]^+$.

Now, the crucial point for embedding the H^2/H^∞ problem in $H_+^{2,-k}$ will be considered.

THEOREM 3. *Consider the spaces presented in Definition* 1.

(a) *If $k \leq m$, the bounded subsets of $L^2_{-k}(i\mathbb{R})$ are bounded in $L^2_{-m}(i\mathbb{R})$, and the bounded closed subsets of $L^2_{-k}(i\mathbb{R})$ are bounded and closed in $L^2_{-m}(i\mathbb{R})$, the same relations existing between sets in $H_+^{2,-k}$ and $H_+^{2,-m}$.*

(b) *The bounded subsets of H_+^∞ are bounded in $H_+^{2,-1}$, and the bounded closed subsets of H_+^∞ are bounded and closed in $H_+^{2,-1}$.*

Remark 5. Here it is essential that the subset be bounded. The spaces H_+^∞ and H_+^2 are closed and unbounded in its own topologies, but they are dense in $H_+^{2,-1}$ in the coarser topology. Also, closed balls in H_+^∞ have empty interiors in relation to $H_+^{2,-1}$ topology.

The next step is to collect the properties of linear and quadratic functionals in $H_+^{2,-k}$, preparing more tools for minimizing the quadratic criteria defined in section 1.

THEOREM 4. *Let $\gamma(s)$ be a real-rational function without poles in $i\mathbb{R}$.*

(a) *The linear functional*

$$F(f) = \int_{-\infty}^{\infty} f^*(i\omega)\gamma(i\omega)d\omega$$

is continuous on $H_+^{2,-k}$ if and only if $\partial_r(\gamma) \geq k+1$.

(b) *The space of continuous linear functional on $H_+^{2,-k}$ can be identified to $H_+^{2,k}$ for any k.*

THEOREM 5. *Let $\Gamma(s)$ be a real rational para-Hermitian function in R_{2k} without poles or zeros in $i\mathbb{R}$, i.e., $\Gamma(s) = \Gamma^*(s)$ and $|\Gamma(i\omega)| > 0$ for each finite ω.*

(a) *$\Gamma(s) = \Phi^*(s)\Phi(s)$, $\Phi(s)$ being a real-rational stable function in R_k^+ with all its zeros in C_+^0 (i.e., minimum-phase).*

(b) *The quadratic functional*

$$f \mapsto \int_{-\infty}^{\infty} f^*(i\omega)\Gamma(i\omega)f(i\omega)d\omega = \langle \Phi f, \Phi f\rangle_2 = \|\Phi f\|_2^2$$

is continuous in $H_+^{2,-m}$ if and only if $m \leq k$. It is coercive in $H_+^{2,-m}$ (i.e., there is a real number $\alpha > 0$ such that $\langle \Phi f, \Phi f\rangle_2 \geq \alpha^2 \|f\|_{2,-m}^2$ for all $f \in H_+^{2,-m}$) if and only if $m = k$. Moreover, it is strictly convex, and $\|\Phi f\|_2$ defines a norm in $H_+^{2,-k}$ equivalent to $\|f\|_{2,-k}$.

3. Optimal H^2 unconstrained control problems. This section presents the mathematical extension of the usual H^2 unconstrained optimal control problem on the mathematical framework developed in the last section. New conditions about its solution will be obtained, clarifying the ones in [8], [9]. This extension will be used in the next section to solve the H^2/H^∞ optimal control problem.

The unconstrained H^2 problem can be defined as follows: Find a function $\breve{K}(s)$ solution to

$$\inf_K \left\{ \int_{-\infty}^{\infty} [K^*(i\omega)\Gamma(i\omega)K(i\omega) - 2K^*(i\omega)\gamma(i\omega)]d\omega \right\} = \inf_K J[K], \tag{3.1}$$

where the functions $K(s)$ belong to some $H_+^{2,-k}$ space, or, formally, to $(S_+)'$, a space containing $H_+^{2,-k}$ for all integer k. $\Gamma(s)$ and $\gamma(s)$ are given real-rational functions. Recall that $K(s)$ is the parameter describing the set of stabilizing controllers (or the set of controllers solving a given servomechanism problem), initially a free real-rational stable and proper function. To define the functional $J[\,.\,]$ some assumptions are needed:

(A1) $\Gamma(s) = \Phi^*(s)\Phi(s)$ is a para-Hermitian real-rational function in R_{2k} without poles or zeros in $i\mathbb{R}$, $\Phi(s)$ being a real-rational stable function in R_k^+ with all its zeros in C_+^0 (i.e., minimum-phase);

(A2) $\gamma(s)$ is a real-rational function without poles in $i\mathbb{R}$, $\partial_r(\gamma) = p$.

The functional $J[\,.\,]$ will be finite only for a meager parameter subset if $\Gamma(s)$ or $\gamma(s)$ have poles on the imaginary axis, as both are rational functions. Indeed, if such happens, $J[K]$ will be finite only for $K(s)$ with zeros on those imaginary poles. The other conditions on assumption A1 are natural for quadratic functional on $H_+^{2,-k}$ spaces, according to Theorem 5 above. Indeed, it is possible to represent all integral quadratic real functional on $H_+^{2,-k}$ spaces as an integral quadratic operator with a para-Hermitian kernel by a procedure similar to the autoadjoint representation for integral quadratic functional on L^2 spaces. Moreover, $\Gamma(s)$ is assumed with no zeros on the imaginary axis because this allows unstable solutions (see Remark 8). Finally, the Wiener–Hopf factorization $\Gamma = \Phi^*\Phi$ is a consequence of the known Youla factorization theorem cited above as Theorem 5(a) [28].

LEMMA 1. *Under assumptions* A1 *and* A2, *let* $m = \min\{k, p-1\}$. *Then the functional* $J[\,.\,]$ *is continuous in* $H_+^{2,-m}$ *but not well-defined in larger spaces, i.e., the integrals in* $J[K]$ *diverge for* $K \in (S_+)' - H_+^{2,-k}$ (*the complement of* $H_+^{2,-k}$ *in* $(S_+)'$).

Proof. The first statement follows from continuity conditions in Theorems 4 and 5(b). For the second statement, if $K \in (S_+)' - H_+^{2,-k}$ is a rational function, $J[K]$ is not defined because $\partial_r(K^*\Gamma K) \le 1$ or $\partial_r(K^*\gamma) \le 1$. □

DEFINITION 2. *The space* $H_+^{2,-m}$ *in Lemma* 1 *will be called the effective domain of the function* $J[\,.\,]$. *This terminology is inherited from convex analysis and adapted to the chain of spaces defined here.*

Now, note that, for $K \in H_+^{2,-m}$ and m as in Lemma 1, $\Phi K \in H_+^2$. Then the functional $J[K]$ can be written as

$$J[K] = \|\Phi K\|_2^2 - 2\int_{-\infty}^{\infty} \{[\Phi(i\omega)K(i\omega)]^*[\Phi^*(i\omega)]^{-1}\gamma(i\omega)\}d\omega. \tag{3.2}$$

As $(\Phi^*)^{-1}\gamma \in L_{p-k-1}^2(i\mathbb{R})$, it can be factorized as a sum of a function in $H_+^{2,p-k-1}$ with a function in $H_-^{2,p-k-1}$, according to Theorem 1(c).

If $p \le k$, this factorization is not unique because $p-k-1 \le -1$. As $(\Phi^*)^{-1}\gamma$ is rational, it is possible to choose a factorization where the polynomial part of $(\Phi^*)^{-1}\gamma$ is taken on the unstable factor. This factorization will be denoted by

$$(\Phi^*)^{-1}\gamma = [(\Phi^*)^{-1}\gamma]_+ + [(\Phi^*)^{-1}\gamma]_-,$$

with $\partial_r([(\Phi^*)^{-1}\gamma]_+) \ge 1$, $[(\Phi^*)^{-1}\gamma]_+ \in H_+^{2,p-k-1}$, $[(\Phi^*)^{-1}\gamma]_- \in H_-^{2,p-k-1}$.

Actually, $[(\Phi^*)^{-1}\gamma]_+ \in H_+^2$ because it is a stable strictly proper rational function with all its poles in C_-^0.

If $p+1 \ge k$, $L_{p-k-1}^2(i\mathbb{R}) \subset L^2(i\mathbb{R})$, $p-k-1 \ge 0$. The above factorization will be interpreted as $[(\Phi^*)^{-1}\gamma]_+ \in H_+^{2,p-k-1} \subset H_+^2$, $[(\Phi^*)^{-1}\gamma]_- \in H_-^{2,p-k-1} \subset H_-^2$, because

$\partial_r([(\Phi^*)^{-1}\gamma]_+) \geq p-k \geq 1$. Note that all the stable projections in different spaces $H_-^{2,p-k-1}$ are denoted by the same symbol $[\,.\,]_+$, but the spaces will be clear from the context.

With this notation, the linear part of $J[K]$ becomes

$$\begin{aligned} &-2\langle \Phi K, [(\Phi^*)^{-1}\gamma]_+\rangle_2 - 2\int_{-\infty}^{\infty} K^*(i\omega)\Phi^*(i\omega)[(\Phi^*(i\omega))^{-1}\gamma(i\omega)]_- d\omega \\ &= -2\langle \Phi K, [(\Phi^*)^{-1}\gamma]_+\rangle_2, \end{aligned}$$

the integral being zero since all the integrand poles are in C_+^0 and its relative degree is less than or equal to 2. (The residue theorem applied to a circuit involving C_-^0 proves the statement; see [8].) In other words, the unstable term $[(\Phi^*)^{-1}\gamma]_-$ is orthogonal to the stable function ΦK. Then, completing the square in (3.2), we get

$$\begin{aligned} J[K] &= \|\Phi K\|_2^2 - 2\langle \Phi K, [(\Phi^*)^{-1}\gamma]_+\rangle_2 + \|[(\Phi^*)^{-1}\gamma]_+\|_2^2 - \|[(\Phi^*)^{-1}\gamma]_+\|_2^2 \\ &= \|\Phi K - [(\Phi^*)^{-1}\gamma]_+\|_2^2 - \|[(\Phi^*)^{-1}\gamma]_+\|_2^2. \end{aligned}$$

Therefore, the minimum of $J[\,.\,]$ is attained at a parameter $\breve{K}$ such that $\Phi\breve{K} - [(\Phi^*)^{-1}\gamma]_+ = 0$, but only if $J[\breve{K}] < \infty$, i.e., only if $\breve{K}$ belongs to the effective domain of $J[\,.\,]$, that is, to $H_+^{2,-m}$. These conclusions are collected in the next theorem.

THEOREM 6. *Let assumptions* A1 *and* A2 *be verified, and let* $\breve{K}$ *be a rational function given by*

$$\breve{K} = \Phi^{-1}[(\Phi^*)^{-1}\gamma]_+, \tag{3.3}$$

where $[(\Phi^*)^{-1}\gamma]_+$ *denotes the stable strictly proper part of* $(\Phi^*)^{-1}\gamma$, $m = \min\{k, p-1\}$. *If* $\breve{K} \in H_+^{2,-m}$ *(the* $J[\,.\,]$ *effective domain), then* $\inf\{J[K]\} = J[\breve{K}]$ *in* $H_+^{2,-m}$.

As commented above, in common H^2/H^∞ problems, $K(s)$ is a proper stable real-rational function, which means $\partial_r(K) \geq 0$. In the mathematical framework presented here, this implies $K \in H_+^{2,-1}$. This situation is explored in the next corollary, easily proved from Theorem 6 and the calculations above. Note that the condition $H_+^{2,-m} \supset H_+^{2,-1}$ is not necessary, but only the condition $\breve{K} \in H_+^{2,-q} \subset H_+^{2,-1}$ with $H_+^{2,-q} \subset H_+^{2,-m}$ for some $q \leq 1$.

COROLLARY 1. *Under the same assumptions as in Theorem* 6, $\partial_r(\breve{K}) \geq 0$ *if and only if* $\partial_r([(\Phi^*)^{-1}\gamma]_+) \geq k$. *Sufficient conditions for this conclusion are* $p = \partial_r(\gamma) \geq 2k$ *or* $k = \partial_r(\Phi) \leq 1$.

Remark 6. The conditions presented in Corollary 1 are sufficient but not necessary. Indeed, for any Φ with $\partial_r(\Phi) = k$ and for any $q \leq k$, it is possible to find a function $\gamma(s)$ as in (3.1) such that $\partial_r(\breve{K}) = 1-q$ and $J[\breve{K}] < \infty$. For that, let $\gamma = \Phi^* B$, $B \in L^2(i\mathbb{R})$ such that $\partial_r([B]_+) = 1+k-q$, which is always possible if $q \leq k$. Note that $k \leq p-1 = \partial_r(\gamma) - 1$, which implies that $J[\,.\,]$ is well-defined in $H_+^{2,-k}$. Then

$$\partial_r(\breve{K}) = \partial_r(\Phi^{-1}[(\Phi^*)^{-1}\Phi^* B]_+) = \partial_r(\Phi^{-1}[B]_+) = 1-q.$$

Also, $\breve{K} \in H_+^{2,-k}$, then $J[\breve{K}] < \infty$.

In the control context, criteria such as (3.1) usually appear as a functional in the form

$$\begin{aligned}&\|AK+B\|_2^2\\ &=\int_{-\infty}^{\infty}\{K^*(i\omega)A^*(i\omega)A(i\omega)K(i\omega)-2K^*(i\omega)A^*(i\omega)B(i\omega)+B^*(i\omega)B(i\omega)\}d\omega,\end{aligned}\tag{3.4}$$

where $B(i\omega)\in L^2(i\mathbb{R})$ is a real-rational strictly proper function with $\partial_r(B)\geq 1$. For each simple functional, by direct verification, $\Phi(s)=A(s)$, $\gamma(s)=A^*(s)B(s)$. Therefore $\partial_r(\gamma)=\partial_r(\Phi)+\partial_r(B)$, which implies the condition:

$$\partial_r(\gamma)\geq\partial_r(\Phi)+1.\tag{3.5}$$

Condition (3.5) is inherited by sums of quadratic functionals as in (3.4) and will greatly simplify the use of Theorem 6. Indeed, under such condition, the function $[\Phi^*(i\omega)]^{-1}\gamma(i\omega)\in L^2(i\mathbb{R})$ because (3.5) corresponds to $p+1\geq k$. Then the decomposition used to prove Theorem 6 will be the usual $L^2(i\mathbb{R})=H_+^2\oplus H_-^2$. In other words, $[(\Phi^*)^{-1}\gamma]_+$ is the usual projection on H_+^2. Moreover, $m=\min\{k,p-1\}=k$. Therefore, $\breve{K}=\Phi^{-1}[(\Phi^*)^{-1}\gamma]_+$ is a rational function with $\partial_r(\breve{K})\geq\partial_r([(\Phi^*)^{-1}\gamma]_+)-\partial_r(\Phi)=1-k$, which implies $J[\breve{K}]<\infty$ and $\breve{K}\in H_+^{2,-m}=H_+^{2,-k}$ with no further condition. In the other sense, if $H_+^{2-m}=H_+^{2,-k}$, then $m=k\leq p-1$, which implies (3.5).

These remarks are collected in the next corollary.

COROLLARY 2. *Let assumptions* A1 *and* A2 *hold. Then condition* (3.5) *is equivalent to saying that the effective domain of* $J[\,.\,]$ *is* $H_+^{2,-k}$. *In this case the function* $\breve{K}(s)$ *given by* (3.3) *is such that* $\inf\{J[K]\}=J[\breve{K}]$ *in* $H_+^{2,-k}$, $[\,.\,]_+$ *denoting the usual orthogonal projection on* H_+^2.

Remark 7. The conditions found in the literature about the unconstrained problem are particular cases of assumptions in Corollaries 1 and 2 [8]. See especially [9], in which a well-motivated criterion is presented such that these conditions are naturally verified.

Remark 8. If $\Gamma(s)$ has zeros on the imaginary axis, $\Phi(s)$ will have the same zeros if the generalized Wiener–Hopf factorization is used as in [28]. Then $\breve{K}$ given in (3.2) will have these zeros as poles, being unstable. In other words, the completion of R_0 in the norm induced by the quadratic part of $J[\,.\,]$ will contain, in this case, unstable rational functions, the minimum being attained in such a function.

Remark 6 shows that $\partial_k(\breve{K})$ can be different from $1-m$, where the $J[\,.\,]$ effective domain is $H_+^{2,-m}$. This possibility will be essential to the algorithm convergence regularity; see section 5 above. Corollary 1 gives conditions for $\partial_r(\breve{K})\geq 0$ if $m\geq 1$. The same considerations used in its proof can be generalized to any relative degree for the optimal solution. Actually, much of the work found in the literature can be linked with this search for regularity, and it was essential in the existence proofs in [8], [9] and in some seminal but unclear comments in [2]. Moreover, a lot of work was needed in [9] to define a natural criterion such that $\partial_r(\breve{K})\geq 0$ for all linear systems for which the proposed servomechanism problem there is solvable. This natural criteria verify assumptions A1, A2 and condition (3.5) with $k=p=1$. Then, by Corollary 2, $m=k$, the effective domain is exactly $H_+^{2,-1}$, which eases considerably the application of the methodology proposed therein.

4. Optimal H^2/H^∞ control problems. This section presents the mathematical extension of the usual H^2/H^∞ control problem on the mathematical framework developed in section 2. The optimal solution existence and uniqueness will be proved in the following and regularity results will be presented.

In the H^2/H^∞ optimal control problem the goal is to find a function $\hat{K}(s)$ solution to

$$\inf_{K\in\Omega\cap\Theta}\left\{\int_{-\infty}^{\infty}[K^*(i\omega)\Gamma(i\omega)K(i\omega)-2K^*(i\omega)\gamma(i\omega)]d\omega\right\}\equiv\inf_{K\in\Omega\cap\Theta}J[K], \tag{4.1}$$

where Ω is a bounded closed convex subset of H_+^∞ and Θ is a bounded closed convex subset of H_+^2. The usual examples of sets Ω and Θ arising from performance, filtering, and robustness specifications are

$$\Omega=\bigcap_{m=1}^{M}\Omega_m;\qquad \Omega_m=\{K\in H_+^\infty:\|A_mK+B_m\|_\infty\le\lambda_m\},$$
$$A_m \text{ and } B_m \text{ functions in } H_+^\infty;$$
$$\Theta=\bigcap_{n=1}^{N}\Theta_n;\qquad \Theta_n=\{K\in H_+^2:\|C_nK+D_n\|_2\le\mu_n\},\quad C_n\in H_+^\infty \text{ and } D_n\in H_+^2;$$

λ_m and μ_n are positive real numbers so that the set Ω is nonempty.

Now, under the assumptions of Lemma 1, the criterion functional in (4.1) is strictly convex and continuous in its effective domain, $H_+^{2,-m}$. From Theorem 3(a), the set Θ is convex, bounded, and closed in $H_+^{2,-m}$ for $m\ge 0$ as a convex, bounded, and closed subset of H_+^2. From Theorem 3(b), Ω is convex, bounded, and closed in $H_+^{2,-1}$ as a convex, bounded, and closed subset of H_+^∞. Then Ω is convex, bounded, and closed in $H_+^{2,-m}$ for $m\ge 1$, from Theorem 3(a). Therefore, we can apply a well-known theorem [29, Theorem 2.6.1, p. 50] to show the existence and uniqueness of the optimal solution for problem (4.1).

THEOREM 7. *Let assumptions* A1 *and* A2 *with* $\partial_r(\Gamma)\ge 2$, $\partial_r(\gamma)\ge 2$ *be verified. Then*

(a) *if the constraint set* $\Omega\cap\Theta$ *is nonempty, the optimal control problem* (4.1) *has one and only one solution in* $H_+^{2,-1}$*;*

(b) *if* Ω *is nonempty, the optimal solution is in* H_+^∞*; if* Θ *is nonempty, the optimal solution is in* H_+^2*.*

Proof. (a) is proved in the above comments. The second statement is clear. ∎

Naturally, it is possible to add $H_+^{2,-1}$ closed convex subsets as new constraints without changing the above conclusions.

Remark 9. A direct consequence of this last theorem is the convergence of the approximating sequence generated by the algorithm proposed in [1] to the optimal solution of problem (4.1). In the same paper it is shown that the optimal control, if it exists, belongs to the H^∞ constraint boundary. Also, the H^2 optimal control problem is explicitly solved with only H^2 constraints by duality, a key to the method proposed therein.

Before the presentation of numerical methods to solve the optimal control problem (4.1) it will be interesting to rewrite it as a minimal norm problem, a step in the strong convergence proof. Assume that $\breve{K}\in H_+^{2,-q}\subset H_+^{2,-m}$ for some $q\le m=\inf\{k,p-1\}$.

Then $\Phi\breve{K} \in H_+^2$. Now, the calculations used to prove Theorem 6 give

$$\begin{aligned} J[K] &= \|\Phi\{K - \Phi^{-1}[(\Phi^*)^{-1}\gamma]_+\}\|_2^2 - \|\Phi\{\Phi^{-1}[(\Phi^*)^{-1}\gamma]_+\}\|_2^2 \\ &= \|\Phi(K - \breve{K})\|_2^2 - \|\Phi\breve{K}\|_2^2. \end{aligned} \tag{4.2}$$

Notation. Let $\|f\|_\Gamma = \|\Phi f\|_2$ be a norm associated to the $J[\,.\,]$ quadratic term, and let $\langle f, g\rangle_\Gamma = \langle \Phi f, \Phi g\rangle_\Gamma$ be the associated internal product.

Theorem 5(b) says that if assumption A1 is verified, $\|f\|_\Gamma$ defines a norm on $H_+^{2,-k}$ equivalent to the norm $\|\,.\,\|_{2,-k}$. Then

$$J[K] = \|K - \breve{K}\|_\Gamma^2 - \|\breve{K}\|_\Gamma^2. \tag{4.3}$$

Therefore, under the assumptions of Theorem 7, the optimal control problem (4.1) is equivalent to finding a function $\hat{K}$ solution to

$$\inf_{\Omega\cap\Theta} \|K - \breve{K}\|_\Gamma^2, \tag{4.4}$$

a best approximation problem in $H_+^{2,-k}$. Note that if condition (3.5) is verified, $H_+^{2,-k} = H_+^{2,-m}$, but here it is needed only that $\breve{K} \in H_+^{2,-q} \subset H_+^{2,-m} \subset H_+^{2,-k}$.

COROLLARY 3. *Let assumptions* A1 *and* A2 *hold with* $\partial_r(\Gamma) \geq 2$, $\partial_r(\gamma) \geq 2$. *Problems* (4.1) *and* (4.4) *are equivalent if and only if* $\breve{K} \in H_+^{2,-q} \subset H_+^{2,-m} \subset H_+^{2,-k}$, *i.e.,* $\partial_r(\breve{K}) \geq 1 - m$, $m = \min\{k, p-1\}$. *Moreover, assumptions* A1 *and* A2 *with* $\partial_r(\Gamma) \geq 2$, $\partial_r(\gamma) \geq 2$ *and condition* (3.5) *are sufficient for the same conclusion.*

Proof. According to the above comments, the first statement is a consequence of Theorem 7 and the second statement is a consequence of Theorem 7 and Corollary 2. □

The optimal control problem (4.1) can be rewritten as a minimal norm problem in $H_+^{2,-k}$ if this space is translated by $\breve{K}$. For that, redefine $G = K - \breve{K}$, $\Omega' = \Omega - \breve{K}$, $\Theta' = \Theta - \breve{K}$. Note that Ω' and Θ' are convex, bounded, and closed in $H_+^{2,-k}$ because these properties are not changed by translations in a Hilbert space. In these notations the optimization problem (4.4) can be translated as the new problem: Find a $\hat{G} \in H_+^{2,-k}$ solution to

$$\inf_{G\in\Omega'\cap\Theta'} \|G\|_\Gamma^2, \tag{4.5}$$

a minimal norm problem. Note that $\breve{K}$ could not belong to $H_+^{2,-1}$. Thus (4.5) shall be solved carefully from a numerical point of view.

Regularity now is essential: if the optimal control problem needs to be solved in some $H_+^{2,-q}$ as a minimal norm problem, beyond the existence conditions in Theorem 7, the condition $H_+^{2,-q} = H_+^{2,-k} \subset H_+^{2,-m}$ will also be needed. This means $q = k \leq m$, with $\breve{K} \in H_+^{2,-q}$, or, more exactly, $\breve{K} \in H_+^{2,-r}$ for some $r \leq q$. For that, Corollary 1 (and its extensions) and Corollary 2 are useful. The usual setting is $q = 1$ as in [8], [21], [22], or, in a more restricted way, $q = k = 1$, as in [9]. In the present paper this setting is generalized to better understand the weak and strong convergence of the algorithm proposed in the next section.

5. The Galerkin method. If $\{\beta_n, n \in N\}$ is a generator set for $H_+^{2,-1}$, not necessarily orthogonal, denote by H_n the finite-dimensional subspace generated by the n first vectors in the generator set. Let Ω_n be defined as $\Omega\cap\Theta\cap H_n$. If Ω_n is nonempty,

it is possible to project the optimal control problem (2.3) in H_n, which defines the following finite-dimensional optimization problem: Find a $\hat{K}_n$ in H_n solution to

$$\inf_{K\in\Omega_n}\{\|K\|_\Gamma^2 - 2\langle K,\gamma\rangle_2\}. \tag{5.1}$$

As Ω_n is a bounded closed convex subset of H_n and the criterion is strictly convex, this optimal control problem has one and only one solution $\hat{K}_n$ in H_n for each $n\in\mathbb{N}$ (see [29, p. 50]). The Galerkin method consists of approximating the optimal solution $\hat{K}$ to the optimal control problem (4.1) by $\hat{K}_n$ if the sequence $\{\hat{K}_n\}$ converges to the optimal solution $\hat{K}$.

We need a technical assumption to have Ω_n nonempty for n sufficiently large.

(A3) Let A_0 denote the set of H_+^∞-functions continuous in the closed right convex semiplane. Assume that $\Omega\cup A_0$ has a nonempty relative interior in A_0 with respect to the H_+^∞ topology.

This assumption is verified for the usual sets Ω_m presented in section 4 because this set has a nonempty interior in H_+^∞ and A_0 is a closed subspace of the same space. Moreover, A_0 is the closure of the rational proper stable functions in H_+^∞ [30, p. 668].

THEOREM 8. *Let the assumptions in Theorem* 7 *and assumption* A3 *be verified. Also, assume that the unconstrained optimal solution* $\breve{K}$ *does not belong to* $\Omega\cap\Theta$. *(Otherwise the optimal solution will be* $\breve{K}$.*) Then the sequence* $\{\hat{K}_n\}$ *generated by the Galerkin method converges weakly in* $H_+^{2,-1}$ *to the unique optimal solution* $\hat{K}$ *to the optimal control problem* (4.1).

Under the assumptions of Corollary 3, including (3.5), the optimal control problem (4.1) can be rewritten as minimal norm problems (4.4) and (4.5), which will allow us to show the strong convergence of the sequence $\{\hat{K}_n\}$ in suitable spaces. For that, let $\{\beta_n, n\in\mathbb{N}\}$ be a generator set for $H_+^{2,-k}$ and let $\|.\|_\Gamma$ be the norm defined in section 4. Thus we can define the projection of the minimal norm problem (4.4) in H_n as some $\hat{K}_n$ in H_n solution to

$$\inf_{K\in\Omega_n}\|K-\breve{K}\|_\Gamma^2, \tag{5.2}$$

where $\breve{K}_n$ is the projection of $\breve{K}$ in H_n. Analogously, translating H_n by $\breve{K}_n$, the minimal norm problem (4.5) can be projected to finding a $\hat{G}_n$ solution of

$$\inf_{G\in\Omega_n'}\|G\|_\Gamma^2, \tag{5.3}$$

where $\Omega_n' = \Omega_n - \breve{K}_n$. As Ω_n and Ω_n' are bounded closed convex sets, the optimization problems (5.2) and (5.3) have one and only one solution, defining sequences of functions approximating the optimal solution to optimal norm problems (4.4) and (4.5) for $n\in\mathbb{N}$.

THEOREM 9. *Let assumptions* A1, A2, A3 *and condition* (3.5) *be verified. Also, assume that* $\partial_r(\Gamma)\geq 2$, $\partial_r(\gamma)\geq 2$, *and that* $\breve{K}$ *does not belong to* $\Omega\cap\Theta$. *Then the sequences* $\{\hat{K}_n\}$ *and* $\{\hat{G}_n\}$ *of all solutions to* (5.2) *and* (5.3) *for all* $n\in\mathbb{N}$ *converge strongly in* $H_+^{2,-k}$ *to the optimal solutions to* (4.4) *and* (4.5), *respectively.*

Remark 10. Note that the strong convergence in $H_+^{2,-1}$ happens only if $k=1$ and condition (3.5) is verified, as in [9].

In the proof of Theorem 8, (5.1) and (5.2) are characterized by linear variational inequalities on $H_+^{2,-k}$. Galerkin methods are powerful for solving this type of inequality in functional spaces [31], generating linear matrix inequalities (LMIs) after the

choice of a basis for $H_+^{2,-k}$. Another approach to problems (4.4) and (4.5) is the one presented under the name of best approximation, using convex projections or proximinal maps (the mapping from $\check{K}$ to $\hat{K}$). This approach is interesting for minimum norm problems in Hilbert spaces, as in the present paper, where the proximinal map is continuous (see [32, pp. 157, 164]). The same reference shows the difficulties when the problem is considered in H_+^∞, which is not a reflexive Banach space (see [32, p. 77]).

Theorems 8 and 9 deal with convergence in $H_+^{2,-1}$, not in H_+^∞. In general, strong $H_+^{2,-1}$ convergence does not imply H_+^∞ strong convergence. It allows spikes in sequences converging to zero, as in $f_n(s) = (ns+1)^{-1}$ (see the proof of Remark 5 in Appendix A). Actually, $\hat{K}_n \to \hat{K}$ strongly in $H_+^{2,-1}$ implies $\Phi_{-1}\hat{K}_n \to \Phi_{-1}\hat{K}$ in measure on the imaginary axis and $\hat{K}_n \to \hat{K}$ in measure on any finite measure subset of the imaginary axis. (In this case the $H_+^{2,-1}$ and H_+^2 strong topologies coincide.) From [33, Theorem 7.11, p. 73], this implies the almost uniform convergence on the finite measure subset. But this result does not imply H_+^∞ strong convergence even in those subsets. In spite of these difficulties, the next theorem and remark show some relevant results in H_+^∞.

THEOREM 10. *If the sequence $\hat{K}_n$ converges to $\hat{K}$ strongly in $H_+^{2,-1}$, as in Theorem 9, then it converges to $\hat{K}$ in the weak topology of H_+^∞.*

Remark 11. If the sequence $\hat{K}_n$ converges to $\hat{K}$ weakly in $H_+^{2,-1}$, as in Theorem 8, then it is possible to prove, after some identifications, that $\hat{K}_n$ converges to $\hat{K}$ in the weak-star topology of $(H_+^\infty)'$.

To end the theoretical presentation of Galerkin methods, some generator set for $H_+^{2,-1}$ and for $H_+^{2,-k}$ must be presented. Due to the density of H_+^2 in $H_+^{2,-k}$, $k \geq 1$, any one of the bases obtained from the Runge theorem [34] for the space of analytic functions on C_+^0 can be used. Note that the topology used in the Runge theorem (the topology of the uniform convergence in all compacts in C_+^0) is finer than the $L^2(i\mathbb{R})$ topology. An example, already used in [15], is the Laguerre orthonormal basis in H_+^2:

$$\left\{ L_n = \sqrt{2a}\ \frac{(s-a)^{n-1}}{(s+a)^n}, \quad n \in \mathbb{N} \right\} \text{ for each positive real number } a.$$

The numerical experiments in [22] show the interest in the use of redundant sets of generators, as

$$\left\{ L_0 = 1, L_n = \sqrt{2a}\ \frac{(s-a)^{n-1}}{(s+a)^n}, \quad n \in \mathbb{N} \right\} \text{ for each positive real number } a,$$

which capture more quickly the asymptotic behavior of the optimal solutions. The proofs of Theorems 8 and 9 apply to these redundant sets without changes.

An orthonormal basis for $H_+^{2,-1}$, in relation to the inner product $\langle \ldots, \ldots \rangle_{\Gamma_{-1}}$, is given by

$$\left\{ M_0 = 1, M_n = \sqrt{2a}\ (1-s)\ \frac{(s-a)^{n-1}}{(s_a)^n}, \quad n \in \mathbb{N} \right\}.$$

Note that $\partial_r(M_n) = 0$, which differs from the Laguerre basis. Reference [22] presents other orthonormal bases for $H_+^{2,-k}$ built under the same principle, with the poles of Φ^* as the zeros of the basis functions. The numerical solution of (5.2) needs some mathematical programming development [22], which will be presented in a future paper. Some comments about the numerical procedure following the developments in [22] will be presented next.

After the choice of a redundant generator set, say $\{1, \beta_n, n \in \mathbb{N}\}$, and the choice of the number of poles of $K_n(s)$, say n, the functions $K_n(s)$ in the finite-dimensional space H_n can be represented as

$$K_n(s) = \sum_{m=0}^{n} \alpha_m \beta_m(s),$$

where $\beta_0(s)$ represents the constant function. By substitution of this last expression in (5.1) or (5.2) an $(n+1)$-dimensional programming problem is defined, whose variable is the $(n+1)$-vector $\vec{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_n)^T$. The integrals in the quadratic functional calculation can be performed analytically, being this functional quadratic in $\vec{\alpha}$. Quadratic constraints are differentiable and can be considered by usual methods [22], but not the H^∞ constraints. Actually, there is no need to explicitly calculate these hard constraints, but only a generalized gradient. The reason is that the finite-dimensional constrained optimization problem was solved by a penalty method coupled with the known BFGS algorithm, where the position of the H^∞ constraint gradient (which does not exist) was provided by a generalized gradient. If this constraint is represented by

$$\sup |A(i\omega)K_n(i\omega) + B(i\omega)| - \lambda \leq 0,$$

it is proved in [22] that the derivative of $|A(i\omega)K_n(i\omega) + B(i\omega)|$ for $\omega = \omega_0$, ω_0 being one of the values where this function assumes its maximum, is a generalized gradient for the constraint. The ω_0 calculation uses the tools of H^∞ theory, as shown in [6]. Note that the procedure should consider also the case where $\omega_0 = \infty$. The convergence of this procedure was proved in [22], and the authors did not find significant problems in obtaining the optimal parameters $\hat{K}_n$ after performing the functional calculations through state variable and Lyapunov equation tools.

6. Numerical example. The example shown here was developed in [22], where a more complete discussion can be found. It represents the pitch optimal control of a fighter airplane described in [35] to exemplify LQG/LQR design, and it is used in [36] to exemplify the dual method from Corrêa [1]. In this example the transfer function from the elevation angle to the attitude angle is

$$P(s) = \frac{-(948{,}12s^3 + 30325s^2 + 56482s + 1215.3)}{s^6 + 64.554s^5 + 1167s^4 + 372.86s^3 - 5495.4s^2 + 1102s + 708.1},$$

the quadratic criterion being the one defined in [9] with weighting filters and weighting coefficients given by

$$\phi_w(s) = 0, \quad \phi_d(s) = \frac{1}{s^2 + 2s + 2}, \quad \phi_v(s) = \frac{1}{s + 10}, \quad \rho_v = \rho_d^n = \rho_v^u = 1.$$

After some calculations, the optimal control problem criterion can be transformed in

$$J_2[K(s)] = \|A(s) + B(s)K(s)\|_2^2 + J_F,$$

where $A(s)$ and $B(s)$ are 14th- and 10th-order rational functions (presented in Appendix B), both with unitary relative degree, $J_F = 0.30612$, and $K(s)$ is the rational proper and stable Youla parameter. The stability margin functional for the control problem, after some transformations to put it in Nehari form [22], is given by

$$J_\infty[K(s)] = \|K(s) - F_0(s)\|_\infty,$$

TABLE 1
Characteristics of some related controllers.

$K(s)$	$J_2[K(s)]$	$J_\infty[K(s)]$	Order
$K_{H2}(s)$	0.306120137	2.07804793	17
$K_{H\infty}(s)$	3.964188309	0.61051297	1
$K_{SPQ}(s)$	2.141469573	0.67180700	29
$K_{SPQR}(s)$	2.141470588	0.67193138	14

$F_0(s)$ being a second-order unstable proper rational function (also presented in Appendix B) with unitary relative degree. The minimum value for $J_\infty[K(s)]$, i.e., the optimal stability margin, is 0.610513.

If we define the robustness constraint allowing a 10% degradation of the optimal stability margin, the H^2/H^∞ problem to be solved becomes

$$\text{Find } K(s) \text{ minimizing } J_2[K(s)] \text{ subject to } J_\infty[K(s)] \leq \gamma = 0.6715643.$$

Assumptions A1 and A2, with $k = p = 1$, condition (3.5), and the others conditions on Theorem 7 are verified. Then, by Theorem 7 this problem has one and only one solution in $H_+^{2,-1}$, belonging to H_+^∞. Also, by Theorem 9 the sequence of functions generated by the Galerkin method, as exposed in section 5, converges strongly to the H^2/H^∞ problem optimal solution for any basis or redundant generator set in $H_+^{2,-1}$.

Table 1 presents some characteristics of controllers solving related optimal control problems, where $K(s)$ is the optimization parameter used to obtain a controller by the Youla parameterization. In the first column,

- $K_{H2}(s)$ represents the Youla parameter corresponding to the controller minimizing the quadratic criterion $J_2[K(s)]$ without constraints (the H^2 optimal controller);
- $K_{H\infty}(s)$ represents the Youla parameter corresponding to the controller minimizing the stability margin (the H^∞ optimal controller);
- $K_{SPQ}(s)$ represents the Youla parameter corresponding to an infeasible controller approximating the H^2/H^∞ problem solution (with $\gamma = 0.6715643$) calculated by the dual method from Corrêa [1];
- $K_{SPQR}(s)$ represents the Youla parameter corresponding to a reduced order controller generated from $K_{SPQ}(s)$ by truncation of a balanced realization.

Note that $K_{SPQ}(s)$ and $K_{SPQR}(s)$ do not verify the stability margin constraint, as expected, i.e., they are not feasible.

Table 2 presents the same characteristics for the controllers obtained by Galerkin method, $n = 1, \ldots, 9$, using the redundant generator set based on Laguerre functions as in section 4, $\gamma = 0.6715643$.

First, all solutions are feasible, as expected. Second, the greater the order, the smaller the quadratic criterion value. Third, comparing the values of $K_9(s)$ and $K_{SPQ}(s)$ and using the dual solutions properties, we verify that

$$J_2[K_{SPQ}(s)] = 2.141469573 < J_2[\hat{K}(s)] < 2.175038 = J_2[K_9(s)],$$

$\hat{K}(s)$ being the H^2/H^∞ problem optimal solution. Therefore, the difference between the quadratic criterion value error of $K_9(s)$ and the quadratic criterion value error of the optimal solution is less than 1.54%.

Table 3 presents the same characteristics for the optimal controllers obtained by the Galerkin method using an $H_+^{2,-1}$ basis generated step by step by minimization

TABLE 2
Characteristics of optimal Galerkin controllers for extended Laguerre functions.

$K(s)$	$J_2[K(s)]$	$J_\infty[K(s)]$	Order
$K_1(s)$	2.436117	0.6715643	1
$K_2(s)$	2.367955	0.6715643	2
$K_3(s)$	2.346453	0.6715643	3
$K_4(s)$	2.250182	0.6715643	4
$K_5(s)$	2.209556	0.6715643	5
$K_6(s)$	2.207430	0.6715643	6
$K_7(s)$	2.206113	0.6715643	7
$K_8(s)$	2.191661	0.6715643	8
$K_9(s)$	2.175038	0.6715643	9

TABLE 3
Characteristics of optimal Galerkin controllers for "optimal step-by-step" basis.

$K(s)$	$J_2[K(s)]$	$J_\infty[K(s)]$	Order
$K_{0A}(s)$	2.651499	0.6715643	0
$K_{1A}(s)$	2.417010	0.6715643	1
$K_{2A}(s)$	2.412348	0.6715643	2
$K_{3A}(s)$	2.278789	0.6715643	3
$K_{4A}(s)$	2.195134	0.6715643	4
$K_{5A}(s)$	2.195134	0.6715643	4
$K_{6A}(s)$	2.164122	0.6715643	6

of the quadratic criterion (under the H^∞ constraint) as a function of both the basis coefficients and the basis poles [22]. The optimization problem to be solved for each dimension n is not convex. Then the usual optimization algorithms give only H_n locally optimal solutions, depending on the algorithm initialization. The BFGS method extended for generalized gradients was used to solve the finite-dimensional optimization problems, the constraints considered by a Lagrangian method [22]. As above, $\gamma = 0.6715643$.

Note that $K_{4A}(s)$ and $K_{5A}(s)$ are equal: the new dimension did not allow a smaller criterion value for the chosen initialization vector. The local character of the n-dimensional numerical optimization and its dependence on the initialization vector is shown by the worst behavior of $K_{2A}(s)$ in relation to $K_2(s)$. In spite of those difficulties, the 6th-order controller attains a smaller criterion value than $K_9(s)$, which allows us to find a best estimation for the criterion optimal value and a best approximation for the optimal controller (corresponding to $K_{6A}(s)$):

$$J_2[K_{SPQ}(s)] = 2.141469573 < J_2[\hat{K}(s)] < 2.164122 = J_2[K_{6A}(s)],$$

with a relative error smaller than 1.05%.

For the sake of comparison, Figures 1, 2, and 3 show the Bode diagrams for the functions $K_{SPQ}(s) - F_0(s)$, $K_9(s) - F_0(s)$, and $K_{6A}(s) - F_0(s)$, respectively. It was verified in [22] that Bode diagrams for the Galerkin approximations do not show significant changes after a sufficiently great dimension n, and they do not present "spikes" in spite of the discussion just before Theorem 10. Numerical calculations were performed on a PC using MATLAB.

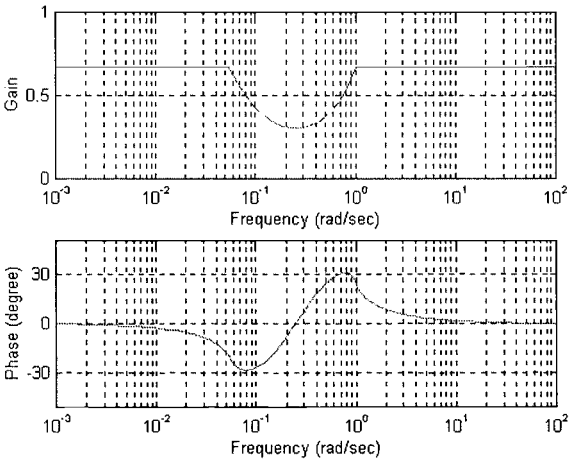


FIG. 1. *Bode diagrams for the function* $K_{SPQ}(s) - F_0(s)$.

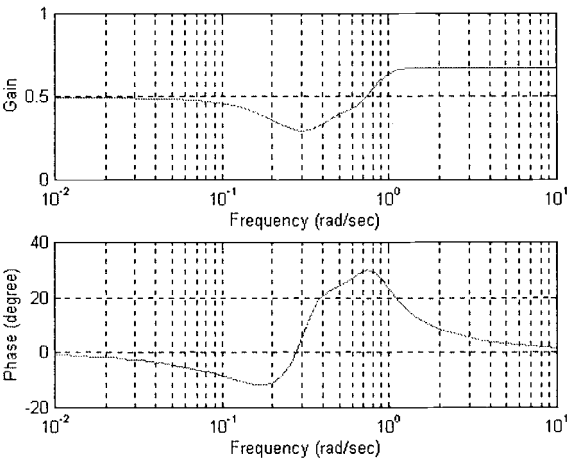


FIG. 2. *Bode diagrams for the function* $K_9(s) - F_0(s)$.

7. Conclusions and comments. In this paper the H^2/H^∞ problem was studied in the context of weighted Hardy spaces, allowing the proof of the existence and uniqueness of its solution and the proof of the convergence of the Galerkin method. Note that this method solves only one convex programming problem after the choice of the controller order.

The extension of these results to the multivariable case is straightforward but tedious, in light of the existing techniques presented, for example, in [8] and [37]. Essentially, consider the following notations: $M[A]$ denotes the set of matrices with entries in A and the dimensions established by the context, K^T denotes the transpose of the matrix K, $[K(s)]^* = [K(-s)]^T$, $\Phi(s)$ denotes a maximal rank real-rational matrix in $M[R_{1-k}]$ with all its poles and zeros in C_+^0, and $\Gamma(s) = \Phi^*(s)\Phi(s)$ denotes a maximal rank real-rational para-Hermitian matrix,

$$\langle K, G\rangle_\Gamma = \int_{-\infty}^{\infty} \mathrm{Trace}\{K^*(i\omega)\Gamma(i\omega)G(i\omega)\}d\omega,$$

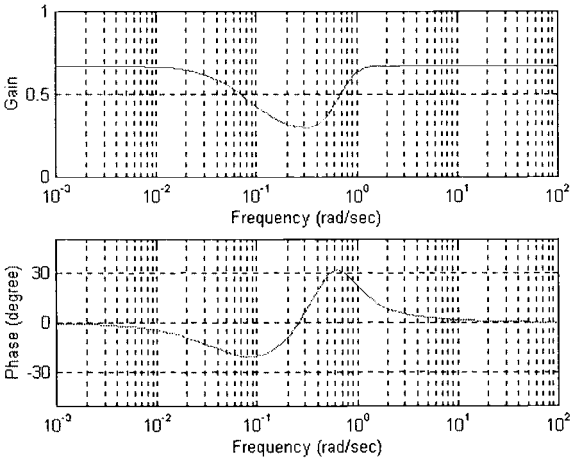


FIG. 3. *Bode diagrams for the function* $K_{6A}(s) - F_0(s)$.

$\|K\|_\Gamma = [\langle K, K\rangle_\Gamma]^{1/2}$, $\|K\|_\infty = \bar{\sigma}\{\|K_{jk}\|_\infty\}$ (the greatest singular value of the matrix whose entries are the H^∞ norm of the K-entries). With these notations the results presented in this paper can be rewritten *ipsis literis* on the spaces $M[L^2_{-k}(i\mathbb{R})]$, $M[H^{2,-k}_+]$, $M[H^{2,-k}_-]$, $M[H^\infty_+]$, and $M[H^\infty_-]$, with the use of $M[R_k]$, $M[R^+_k]$, and $M[R^-_k]$ and the obvious adaptations in notation and proofs. The conditions on zeros and poles can be written as $\det\{s^2\Gamma(i\omega)\} > \eta$ for some $\eta > 0$ and for each real number ω. This assumption is verified by the functional defined in section 2 [8], [9]. A serious problem not considered in this paper is the great number of entries in a multivariable basis, which increases dramatically the number of parameters in the optimization problems (4.1), (4.2), and (4.3). The most parsimonious basis, with the same poles in all the entries of the rational matrix $K(s)$, uses as many parameters as the product of the entry number by the number of parameters in the one-dimensional problem (i.e., for an m-dimensional problem on H_n, we have an nm-dimensional optimization problem).

The algorithms presented in sections 5 and 6 do not explore all the theoretical possibilities. The freedom in the choice of a generator set allowed by the Runge theorem linked to a convenient use of model order reduction algorithms by balanced realizations can be used to build an algorithm optimizing, in a certain sense, the generator set used in each step of the Galerkin method. The numerical behavior and the convergence of such an algorithm are better than the simpler algorithms proposed in this paper, as is shown in [22]. The presentation of the "optimized basis" methodology will be the subject of a future paper, where the mathematical programming algorithms to be used will be carefully developed.

Now, some regularity results will be considered, linking our results with [17], [18], and [19]. First, in Theorem 7 it is assumed that $\partial_r(\Gamma) \geq 2$ and $\partial_r(\gamma) \geq 2$, which imposes $H^{2,-1}_+$ as the natural space for the Youla parameter $K(s)$. Almost all the literature, in particular [14], [15], [16], [17], [18], [19], and [20], assume this parameter in H^2_+, which means $\partial_r(\Gamma) = 0$ and $\partial_r(\gamma) \geq 1$. Under these conditions, the optimal unconstrained parameter $\breve{K}$ belongs to H^2_+, $J[K] < \infty$ if and only if $K \in H^2_+$ (see section 3). Also the norm in problem (4.4) is equivalent to the H^2_+ norm. This problem can be relaxed to $H^{2,-1}_+$, where $J[\,.\,]$ is a strictly convex upper semicontinuous

functional [38]. As $\Omega \cap \Theta$ is a nonempty, convex, bounded, and closed subset of $H_+^{2,-1}$, problem (4.1) has one and only one solution $\hat{K}$ in this space [38]. But it is easy to show that there exists almost one $K \in \Omega \cap \Theta \cap H_+^2$. As $J[\hat{K}] \leq J[K] < \infty$, then $J[\hat{K}] < \infty$, which implies that $\hat{K} \in H_+^2$.

THEOREM 11. *Let assumptions* A1, A2, *and* A3 *with* $\partial_r(\Gamma) = 0$, $\partial_r(\gamma) \geq 1$ *be verified. If the constraint set* $\Omega \cap \Theta$ *is nonempty, then the optimal control problem* (4.1) *has one and only one solution in* H_+^2.

The algorithm in section 5 applies without changes if used with the original Laguerre basis or any other complete set for H_+^2. Thus the sequence $\hat{K}_n$ generated by the Galerkin method approaches the optimal solution strongly in H_+^2 and weakly in H_+^∞.

Another optimal control problem (presented in [18] to obtain rational approximations in H_+^∞) searches for solutions in A_0, the class of continuous functions in the extended imaginary axis. It is known that $A_0 \cap H_+^\infty$ is the closure of the proper rational functions in H_+^∞ [30, p. 668]. Then we can redefine problems (4.1) and (4.4) with constraint set $\Omega_0 = \Omega \cap A_0$. As A_0 is a closed linear subspace of H_+^∞, Ω_0 is also closed and all the results apply. But the sequences defined in section 5 are the same for Ω as for Ω_0. Thus the optimal solutions for both the problems are the same, which implies that $\hat{K} \in A_0$. In spite of that regularity, $\hat{K}$ does not represent an exponentially stable linear system (except for trivial cases), as a consequence of [19]. Recall that exponentially stable systems with finite numbers of inputs and outputs are characterized by their transfer functions being in the algebra $\hat{A}_-(0)$ defined in section 1 (see [30, p. 364]).

THEOREM 12. *Let assumptions* A1, A2, *and* A3 *with* $\partial_r(\Gamma) = 2k$, $\partial_r(\gamma) \geq 2k+1$, $k \geq 0$, *be verified. The optimal solution* $\hat{K}(s)$ *to* (4.1) *or* (4.4) *belongs to* A_0, *but it is not a transfer function of an exponentially stable system; i.e., it does not belong to the algebra* $\hat{A}_-(0)$ *(except for trivial cases).*

This theorem rules out functions as e^{-s}, but not $e^{-s}/(s+1)$. Actually, $A_0 \supsetneq \hat{A}_-(0)$, the latter algebra being dense in the former (because $\hat{A}_-(0)$ contains the Laguerre functions). Thus (4.1) with a constraint defined by $\Omega' = \Omega \cap \hat{A}_-(0)$ is not well-posed, the optimal solution in this case being the same as in the former problem, i.e., outside Ω', or, more precisely, in its H_+^∞-boundary.

In [18] a new optimal control problem is defined with an exponentially stable solution. The last paragraph shows that it is necessary to change the problem structure to force optimal controllers in $\hat{A}(0)$. An example of such a new problem, simpler than the one proposed in [18], is to find $\hat{K}_\varepsilon$ such that

$$J_\varepsilon(\hat{K}_\varepsilon) = \inf_{K \in \Omega_\varepsilon} \left\{ \int_{-\infty}^{\infty} [K^*(s)\Gamma(s)K(s) - 2K^*(s)\gamma(s)]_{s=i\omega-\varepsilon} d\omega \right\}, \tag{7.1}$$

where Ω_ε has the same definition as Ω but with the supremum taken in the imaginary axis translated to the left by ε, the positive real number ε sufficiently small to have $s\Gamma(s)$ bounded from zero and without poles or zeros and $\gamma(s)$ without poles in the vertical strip

$$\{s = \sigma + i\omega, \omega \in \mathbb{R}, -\varepsilon \leq \sigma \leq \varepsilon\}.$$

This new problem can be solved in the spaces $H_{+,\varepsilon}^{2,-1}$ and $H_{+,\varepsilon}^\infty$, defined on the semiplane $\{s \in \mathbb{C} : \mathrm{Re}\{s\} > -\varepsilon\}$ similarly to $H_+^{2,-1}$ and H_+^∞ but with the norms calculated on the translated imaginary axis $i\mathbb{R}_\varepsilon = \{s = -\varepsilon + i\omega, \omega \in \mathbb{R}\}$. It is straightforward to

extend calculations in sections 4 and 5 to this new problem, proving the existence and uniqueness of solutions. Applying the Galerkin method to a generator set composed by rational functions with all poles at the left of $i\mathbb{R}_\varepsilon$, we can find an approaching sequence $\hat{K}_{n\varepsilon}$ which converges to the optimal parameter $\hat{K}_\varepsilon$, strongly in $H^{2,-1}_{+,\varepsilon}$ and weakly in $H^{\infty}_{+,\varepsilon}$. Moreover, the optimal solution $\hat{K}_\varepsilon$ is analytic in an open strip containing the imaginary axis. Thus $\hat{K}_\varepsilon \in \hat{A}_-(0)$ is exponentially stable. We can show that (7.1) and the optimal control problem solved in [18] are approximations to (4.1) in the same sense. In spite of all these convergence results, conditions to $\hat{K}_n$ and $\hat{K}_\varepsilon$ converge to $\hat{K}$ strongly in H^{∞}_{+} remain an open problem.

Finally, the algorithm proposed in the present paper can be computed in polynomial time as the resolution of a finite-dimensional convex optimization problem by a quasi-Newton algorithm (in this case the generalized BFGS algorithm) after the choice of the controller order.

Appendix A. In this appendix we provide the proofs not presented in the main text.

Proof of comments after Definition 1. As $\langle f, g\rangle_{2,-k} = \langle \Phi_{-k} f, \Phi_{-k} g\rangle_2$ and $\Phi^*_{-k}(i\omega)\Phi_{-k}(i\omega) > 0$ for all ω, the announced properties are inherited from the inner product and the norm in $L^2(i\mathbb{R})$ if the integrals are finite. If $f \in R_{k-1}$, this last property is a consequence of f being a rational function without poles in $i\mathbb{R}$ and

$$\partial_r(f^*\Phi^*_{-k}\Phi_{-k}f) \geq (1-k) + 2k + (1-k) = 2$$

(then integrable on $i\mathbb{R}$). □

Proof of Theorem 1. (a) The function f belongs to $L^2_{-k}(i\mathbb{R})$ if and only if $\Phi_{-k}f$ belongs to $L^2(i\mathbb{R})$, by definition. As R_1 is dense in $L^2(i\mathbb{R})$ [26], $\Phi^{-1}_{-k}R_1$ is dense in $L^2_{-k}(i\mathbb{R})$. But $\Phi^{-1}_{-k}R_1 = R_{1-k}$. Indeed, if $f \in R_{1-k}$, $g = (\Phi_{-k})f$ belongs to R_1 because $\partial_r(\Phi_{-k}f) \geq 1$ and g has no poles in $i\mathbb{R}$. In the reverse direction, if f belongs to R_1, then $\Phi^{-1}_{-k}f$ belongs to R_{1-k} because $\partial_r(\Phi_{-k}f) \geq 1-k$ and f has no poles in $i\mathbb{R}$. Therefore, R_{1-k} is dense in $L^2_{-k}(i\mathbb{R})$. The same specified argument, when applied to R^+_1 and H^2_+, R^+_{1-k} and $H^{2,-k}_+$, R^-_1 and H^2_-, and R^-_{1-k} and $H^{2,-k}_-$, proves the stated densities. The final statement in (a) is a consequence of $L^2_{-k}(i\mathbb{R})$ being the completion of R_{1-k} in the norm $\|\,.\,\|_{2,-k}$; the same applies to H^2_+ in relation to R^+_1, to $H^{2,-k}_+$ in relation to R^+_{1-k}, etc.

(b) $H^{2,-k}_+$ and $H^{2,-k}_-$, as closures of R^+_{1-k} and R^-_{1-k} in $L^2_{-k}(i\mathbb{R})$, are closed subspaces.

(c) Theorem 1(c) follows straightforwardly from (a). Note that $e^{-s\Delta} \in H^{\infty}_+$ for Δ a positive real number because $e^{(a+\omega i)\Delta}$ is bounded on each vertical straight line in C^0_+ for each real $a < 0$. Then $e^{-s\Delta}$ belongs to $H^{2,-1}_+$.

Proof of Theorem 2. (a) As $k < m$, $\Phi_{-m} = \Phi\Phi_{-k}$ for some real-rational stable and minimum phase function with $\partial_r(\Phi) = m - k > 0$. Then a function f belongs to $L^2_{-m}(i\mathbb{R})$ if and only if Φf belongs to $L^2_{-k}(i\mathbb{R})$ as a consequence of Definition 3 and as a consequence of $\|f\|_{2,-m} = \|\Phi_{-m}f\|_2 = \|\Phi\Phi_{-k}f\|_2 = \|\Phi f\|_{2,-k}$. Therefore, the operator $f \to \Phi f$ is an isometry from $L^2_{-m}(i\mathbb{R})$ to $L^2_{-k}(i\mathbb{R})$, the inverse isometry being $g \to \Phi^{-1}_g$. By the Cauchy–Schwarz inequality applied in $L^2(i\mathbb{R})$,

$$\|f\|_{2,-m} = \|\Phi\Phi_{-k}f\|_2 \leq \|\Phi\|_2\|\Phi_k f\|_2 = \|\Phi\|_2\|f\|_{2,-k},$$

and, as $\|\Phi\|_2 < \infty$, $L^2_{-k}(i\mathbb{R}) \subset L^2_{-m}(i\mathbb{R})$. Then the isometry from $L^2_{-k}(i\mathbb{R})$ to $L^2_{-m}(i\mathbb{R})$ is an injective mapping and its inverse is a surjective mapping.

(b) Theorem 2(b) is a direct consequence of Theorems 1(b) and 2(a).

(c) Let $k \geq 0$. First, we will prove that R_1 is dense in R_0 in the $L^2_{-1}(i\mathbb{R})$ topology. Actually, we need only show that the constant function $f(s) \equiv 1$ is a limit of R_1-functions in this topology. Defining $f_n(s) = n(s+n)^{-1}$,

$$\|f_n - 1\|^2_{2,-1} = \int_{-\infty}^{\infty} \frac{\omega^2}{n^2+\omega^2} \frac{1}{1+\omega^2} d\omega = \frac{\pi}{n+1},$$

which converges to zero if n goes to ∞, showing the desired convergence and the stated density.

Second, as $R_0 \subset R_1 \subset L^2_{-1}(i\mathbb{R})$, R_1 is also dense in $L^2_{-1}(i\mathbb{R})$.

Third, as $R_1 \subset L^2(i\mathbb{R}) \subset L^2_{-1}(i\mathbb{R})$, the density of R_1 in $L^2_{-1}(i\mathbb{R})$ implies the density of $L^2(i\mathbb{R})$ in $L^2_{-1}(i\mathbb{R})$.

Fourth, more generally, let M be a total set in $L^2_{-k}(i\mathbb{R})$, $k < m$, and assume $f(s) \in L^2_{-m}(i\mathbb{R})$. Set

$$\int_{-\infty}^{\infty} f^*(i\omega)\Phi^*_{-m}(i\omega)\Phi_{-m}(i\omega)g(i\omega)d\omega = 0 \ \forall \ g(s) \in M,$$

which makes sense because $g(s) \in L^2_{-k}(i\mathbb{R}) \subset L^2_{-m}(i\mathbb{R})$. Then,

$$f\Phi_{-m}\Phi^*_{-m} \in L^2_{-m}(i\mathbb{R}) \subset L^2_k(i\mathbb{R}) \approx [L^2_{-k}(i\mathbb{R})]'$$

(where the symbol $\approx$ denotes the identification to be shown in Theorem 4(a) below, proved independently from the present theorem), implying that $f\Phi_{-m}\Phi^*_{-m}$ can be taken as the zero function. This implies that $f(s) \equiv 0$ because $\Phi_{-m}(i\omega)\Phi^*_{-m}(i\omega)$ is strictly positive for all real ω. Therefore, as $g(s)$ is any function in a total set, the set M is also total in $L^2_{-m}(i\mathbb{R})$ by a known corollary of the Hahn–Banach theorem. From $k < m$, $L^2_{-k}(i\mathbb{R}) \subset L^2_{-m}(i\mathbb{R})$, proving the density of the first in the second.

Analogous arguments can be used for $H^{2,-k}_+$ and $H^{2,-k}_-$.

(d) Assume that f_n converges to f in H^∞_+. Then

$$\begin{aligned}
\|f_n - f\|^2_{2,-1} &= \int_{-\infty}^{\infty} |f_n(i\omega) - f(i\omega)|^2 \ |\Phi_{-1}(i\omega)|^2 \, d\omega \\
&\leq \operatorname*{ess\,sup}_{\omega \in R} \ \{|f_n(i\omega) - f(i\omega)|^2\} \int_{-\infty}^{\infty} \Phi^*_{-1}(i\omega)\Phi_{-1}(i\omega)d\omega \\
&\leq \|f_n - f\|^2_\infty \|\Phi_{-1}\|^2_2,
\end{aligned}$$

and because $\|\Phi_{-1}\|^2_2$ is finite, f_n converges to f in $L^2_{-1}(i\mathbb{R})$. The stability of f is assured because $f \in H^\infty_+$. To complete the proof, let us now exhibit a function in $H^{2,-1}_+$ that does not belong to H^∞_+. First, note that there are unbounded functions in $L^2(i\mathbb{R})$, as $g(i\omega) = |i\omega|^{-1/4}\chi_{[-1,1]}$, where $\chi_{[-1,1]}$ denotes the characteristic function of the closed interval $[-1,1]$. Straightforward calculations show that $\|g\|_2 = 2$ and that $|g(i\omega)|$ diverges when ω goes to zero. As a $L^2(i\mathbb{R})$ function, $g = g_+ + g_-$, where $g_+ \in H^2_+$ and $g_- \in H^2_-$. Both functions cannot be simultaneously bounded, because g is not bounded. If g_+ is unbounded, it is the example completing the proof, because $g_+ \in H^2_+ \subset H^{2,-1}_+$ but $g_+ \notin H^\infty_+$. If g_+ is bounded, g_- is unbounded, and $g^*_-(s) = g_-(-s) \in H^2_+ \subset H^{2,-1}_+$ and is unbounded because $|g^*_-(s)| = |g_-(s)|$, $g^*_-(s)$ being the example, and completing the proof. ∎

Proof of Remark 2. Remark 2 is proved in (c) above if we note that

$$\|f_n\|_2^2 = \int_{-\infty}^{\infty} \frac{n^2}{n^2+\omega^2} d\omega = n\pi,$$

which implies that the sequence $\{f_n\}$ does not converge in $L^2(i\mathbb{R})$ when n goes to ∞, in spite of its convergence in $H_+^{2,-1}$. □

Proof of Theorem 3. (a) As $k \le m$, $\Phi_{-m} = \Phi\Phi_{-k}$ for some real-rational stable and minimum phase function with $\partial_r(\Phi) = m - k \ge 0$. First, if $\|f\|_{2,-k} \le M$, by the Cauchy–Schwarz inequality

$$\|f\|_{2,-k} \le \|f\|_{2,-m}\|\Phi\|_2 \le M\|\Phi\|_2,$$

proving the first part of the statement. Second, the closed balls of $L^2_{-k}(i\mathbb{R})$ are closed in $L^2_{-m}(i\mathbb{R})$ as an inverse image of closed sets by an isometric isomorphism (see Lemma 2).[1]

Third, if Ω is a bounded closed set in $L^2_{-k}(i\mathbb{R})$, it is within a closed ball in $L^2_{-k}(i\mathbb{R})$, which is a closed subset of $L^2_{-m}(i\mathbb{R})$. As Ω is closed in a closed subset of a metric subspace of $L^2_{-m}(i\mathbb{R})$, Ω is also closed in $L^2_{-m}(i\mathbb{R})$ (see Theorem 2,II,9,2,b in [25], page 27). Fourth, as $H_+^{2,-k}$ is a closed subspace of $L^2_{-k}(i\mathbb{R})$, the last property is inherited by $H_+^{2,-k}$.

(b) First, if $\|f\|_\infty \le M$,

$$\|f\|_{2,-1}^2 = \int_{-\infty}^{\infty} |f(i\omega)|^2\, |\Phi_{-1}(i\omega)|^2 d\omega \le \|f\|_\infty^2 \|\Phi_{-1}\|_2^2 \le M^2\|\Phi_{-1}\|_2^2 < \infty,$$

which shows that bounded subsets of H_+^∞ are bounded in the $L^2_{-1}(i\mathbb{R})$ metric.

Second, it will be shown that the closed balls in H_+^∞ are closed in $H_+^{2,-1}$. For that, let $\{f_n\}$ be a sequence in a closed ball of H_+^∞ with radius M, i.e., $\|f_n\|_\infty \le M$ for all $n \in \mathbb{N}$. Let $f \in H_+^\infty$ with $\|f\|_\infty > M$. Thus there is a positive real number ε so that $\|f\|_\infty$ is strictly greater than $M + 2\varepsilon$. The definition of "essential supremum" implies that there exists a set $E \subset \mathbb{R}$ with strictly positive measure so that $|f(i\omega)| > M + \varepsilon$ for all $\omega \in E$. Therefore, f_n does not converge to f in $H_+^{2,-1}$ because

$$\begin{aligned}
\|f_n - f\|_{2,-1}^2 &= \int_{-\infty}^{\infty} |f_n(i\omega) - f(i\omega)|^2\, |\Phi_{-1}(i\omega)|^2\, d\omega \\
&\ge \int |f_n(i\omega) - f(i\omega)|^2\, |\Phi_{-1}(i\omega)|^2\, d\omega \\
&\ge \int_E |f(i\omega)^2 |\Phi_{-1}(i\omega)|^2 d\omega - \int_E |f_n(i\omega)|^2\, |\Phi_{-1}(i\omega)|^2 d\omega \\
&> [(M+\varepsilon) - M] \int_E |\Phi_{-1}(i\omega)|^2 d\omega = \varepsilon \int_E |\Phi_{-1}(i\omega)|^2 d\omega > 0,
\end{aligned}$$

[1] A more direct proof uses the weak continuity of the multiplication by Φ. Indeed, reasoning on $L^2(i\mathbb{R})$, $L^2_{-1}(i\mathbb{R})$, if $f_n \to f$ weakly in $L^2(i\mathbb{R})$, $|\int (f_n - f)\Phi g d\omega| \le \|\Phi\|_\infty \int |(f_n - f)g| d\omega \to 0$ for all g in $L^2(i\mathbb{R})$. Then, if for all $n \in \mathbb{N}$, $\|f_n\|_2 \le M$, there is a subsequence, say $\{f_m\}$, converging weakly in $L^2(i\mathbb{R})$ to a limit f_n such that $\|f_w\|_2 \le M$ (see [28, p. 26]). The weak continuity proved above implies that Φf_m converges weakly in $L^2(i\mathbb{R})$ to Φf_w. However, as f_n converges to f strongly in $L^2_{-1}(i\mathbb{R})$, Φf_n converges strongly to Φf in $L^2(i\mathbb{R})$, and then Φf_n converges weakly to Φf in $L^2(i\mathbb{R})$. As $\{f_m\}$ represents a subsequence of $\{f_n\}$, $\Phi f_w = \Phi f$ (which implies $f_w = f$ in $L^2(i\mathbb{R})$ because $\Phi(\,.\,)$ is a continuous bounded function with no zeros on $i\mathbb{R}$), it follows that $\|f\|_2 \le M$, proving that the closed ball with radius M in $L^2(i\mathbb{R})$ is also closed in $L^2_{-1}(i\mathbb{R})$. The same reasoning applies to $L^2_{-k}(i\mathbb{R})$ for any integer k.

the last integral being strictly positive because $\Phi_{-1}(i\omega)$ is continuous and strictly positive on the real axis. The contrapositive proposition is

$$\text{If } f_n \to f \text{ in } H_+^{2,-1}, \text{ then } \|f\|_\infty \le M,$$

implying that closed balls of H_+^∞ are also closed in the $H_+^{2,-1}$ topology.

Third, if Ω is a bounded closed set in H_+^∞, it is contained in a closed ball in H_+^∞, which is a closed set in $H_+^{2,-1}$. As Ω is closed in a closed subset of a metric subspace of $H_+^{2,-1}$, Ω is also closed in $H_+^{2,-1}$ (see Theorem 2,II,9, 2,b in [25], page 27). ∎

Proof of Remark 5. Let $f_n(s) = (ns+1)^{-1}$, $n \in \mathbb{N}$. These functions belong to H_+^∞ with $\|f_n\|_\infty = f_n(0) = 1$. Also

$$\|f_n\|_2^2 = \int_{-\infty}^{\infty} \frac{(1/n)^2}{(1/n)^2 + \omega^2} d\omega = \frac{\pi}{n}.$$

Therefore, the sequence $f_n(s)$ converges to zero in H_+^2 and, a fortiori, in $H_+^{2,-1}$. But it does not converge to zero in H_+^∞. Now, let $g(s) \in H_+^\infty$ be any function such that $\|g\|_\infty \le 1$. Then $g_n(s) = g(s) + 3f_n(s)$ converges to $g(s)$ in $H_+^{2,-1}$ but does not converge in H_+^∞ because $\|g_n\|_\infty > 2$ for all n. Therefore, any function in the closed unit ball in H_+^∞ can be strongly approximated in $H_+^{2,-1}$ by functions in the exterior of this ball: all the functions in this set are in its $H_+^{2,-1}$ boundary. The H_+^∞ closed balls have an empty interior in the $H_+^{2,-1}$ topology. ∎

Proof of Theorem 4. (a) As continuous linear functionals on Hilbert spaces are uniformly continuous, we need to prove the statement only on R_{1-k}^+, a dense subset of $H_+^{2,-k}$ [38, p. 98]. In this case, as $\gamma(s)$ and $f(s)$ have no poles on the imaginary axis, the integral will be finite if and only if $\partial_r(f) + \partial_r(\gamma) \ge 2$. This occurs for all $f \in R_{1-k}^+$ if and only if $\partial_r(\gamma) \ge 2 - (1-k) = k+1$. Also, $(\Phi_{-k}^*)^{-1}\gamma \in L^2(i\mathbb{R})$ and $\Phi_{-k} f \in L^2(i\mathbb{R})$. Then, by the Cauchy–Schwarz inequality, the linear functional $F(f)$ is continuous on R_{1-k}^+ because

$$\left| \int_{-\infty}^{\infty} f^*(i\omega)\gamma(i\omega)d\omega \right| = \left| \int_{-\infty}^{\infty} (f\Phi_{-k})^*(i\omega)[(\Phi_{-k}^*)^{-1}\gamma](i\omega)d\omega \right|$$
$$\ge \|(\Phi_{-k}^*)^{-1}\gamma\|_2 \, \|f\|_{2,-k}.$$

(b) As a consequence of (a), a rational function $g(s)$ is in the dual space of $H_+^{2,-k}$ if and only if $\partial_r(g) \ge k+1$, i.e., $g \in H_+^{2,k}$. As the dual of $H_+^{2,-k}$ is a Hilbert space, the completion argument proves the statement. ∎

Proof of Theorem 5. (a) The statement is an adaptation of the known Youla Theorem; see [28].

(b) We need to prove the statement only on R_{1-k}^+, a dense subset of $H_+^{2,-k}$ (see [38, p. 100]). Now, if $f(s)$ is a rational function without poles on $i\mathbb{R}$,

$$\int_{-\infty}^{\infty} f^*(i\omega)\Gamma(i\omega)f(i\omega)d\omega = \int_{-\infty}^{\infty} [\Phi f(i\omega)]^*[\Phi f(i\omega)]d\omega = \|\Phi f\|_2^2 < \infty$$

if and only if $\partial_r(f\Phi) \ge 1$, i.e., $\partial_r(f) \ge 1-k$ or $f \in H_+^{2,-k}$.

Also, as $\Phi(s)$ and $\Phi_{-k}(s)$ are rational functions with no poles or zeros on $i\mathbb{R}$, and those functions have the same relative degree, there are real numbers α and β such that

$$0 < \alpha \le |\Phi(i\omega)\Phi_{-k}^{-1}(i\omega)| \le \beta < \infty.$$

This implies that if $f \in R^+_{1-k}$, then

$$\alpha\|f\|^2_{2,-k} \le \int_{-\infty}^{\infty} f^*(i\omega)\Gamma(i\omega)f(i\omega)d\omega = \|\Phi f\|_2^2 \le \beta\|f\|^2_{2,-k}.$$

Thus $\|\Phi f\|_2$ defines a norm equivalent to $\|f\|_{2,-k}$, the quadratic functional being continuous on R^+_{1-k} as the square of an equivalent norm. Finally, if $m < k$ and $f \in H_+^{2,-m}$, $\|f\Phi_{-m}(i\omega)|^2$ and $|\Phi\Phi^{-1}_{-m}(i\omega)|^2$ belong to $L^2(i\mathbb{R})$, then

$$\int_{-\infty}^{\infty} f^*(i\omega)\Gamma(i\omega)f(i\omega)d\omega = \int_{-\infty}^{\infty} [\Phi_{-m}f(i\omega)]^*[\Phi_{-m}f(i\omega)][\Phi\Phi^{-1}_{-m}(i\omega)]^*[\Phi\Phi^{-1}_{-m}(i\omega)]d\omega$$

$$\le \|\Phi\Phi^{-1}_{-m}\|_2^2\|\Phi_{-m}f\|_2^2 = \|\Phi\Phi^{-1}_{-m}\|_2^2\|f\|^2_{2,-m}$$

by the Cauchy–Schwarz inequality. Therefore, the quadratic functional is continuous in $H_+^{2,-m}$ at the origin, then continuous in $H_+^{2,-m}$ for $m < k$.

The coerciveness on $H_+^{2,-k}$ was shown above, where α is the coerciveness constant. For $m < k$ and $f \in H_+^{2,-m}$, let

$$f_n(s) = \Phi^{-1}_{-m}(s)[s\sqrt{n}(s+n)^{-2}], \qquad \Phi(s)\Phi^{-1}_{-m}(s) = g(s)(s+1)^{-1},$$

where $|g(i\omega)|^2 \le \beta^2 < \infty$ for some real number β because $g(s)$ is a proper rational function without poles on the imaginary axis. Straightforward calculations show that

$$0 \le \|\Phi f_n\|_2^2 \le \beta^2 \left\| \frac{1}{s+1}\frac{s\sqrt{n}}{(s+n)^2}\right\|_2^2 = \beta^2\frac{(2n^2-2n+1)\pi}{2(n^2-1)^2},$$

which converges to zero when n goes to infinity. But $\|f_n\|^2_{-m} = \pi/2$ for all n. Then there is no real number α such that $\alpha^2\|f_n\|^2_{-m} \le \|\Phi f_n\|_2^2$ for all n, which shows that the quadratic functional is not coercive on $H_+^{2,-m}$ for $m < k$.

The proof of the strictly convexity is straightforward. □

Proof of Corollary 1. As $\partial_r(\breve{K}) = \partial_r([(\Phi^*)^{-1}\gamma]_+) - k$, then $\partial_r(\breve{K}) \ge 0$ if and only if $\partial_r([(\Phi^*)^{-1}\gamma]_+) \ge k$. If $\partial_r(\gamma) \ge 2k$, then $\partial_r((\Phi^*)^{-1}\gamma) \ge k$, implying that $\partial_r([(\Phi^*)^{-1}\gamma]_+) \ge k$, which proves the sufficiency of the condition. If $k = \partial_r(\Phi) \le 1$, as $\partial_r([(\Phi^*)^{-1}\gamma]_+) \ge 1$, then $\partial_r(\breve{K}) \ge 0$. □

Proof of comments about condition (3.5). We need to prove that condition (3.5) is inherited by a finite sum of quadratic functional as in (3.4). To do that, denote the functional as

$$J[K] = \sum J_n[K], \qquad J_n[K] = \int_{-\infty}^{\infty}\{K^*\Gamma_n K - 2K^*\gamma_n\}d\omega, \quad \Gamma_n = \Phi_n^*\Phi_n.$$

Then

$$J[K] = \int_{-\infty}^{\infty}\{K^*\Gamma K - 2K^*\gamma\}d\omega \qquad \text{for } \Gamma = \Phi^*\Phi = \sum \Phi_n^*\Phi_n, \gamma = \sum \gamma_n.$$

Let $\partial_r(\gamma_n) \ge \partial_r(\Phi_n) + 1$ and assumptions A1 and A2 hold for each n. Then

$$\partial_r\left(\sum \gamma_n\right) \ge \min\{\partial_r(\gamma_n)\},$$

as usual, but

$$\partial_r\left(\sum \Gamma_n\right) = \partial_r\left(\sum \Phi_n^* \Phi_n\right) = \min\{\partial_r(\Gamma_n)\}$$

because the numerator of the first term is a sum of para-Hermitian functions, each one strictly positive on the imaginary axis, which implies that its degree is the maximum degree of the parcels. See [9] for a complete development of this argument. Therefore,

$$\partial_r\left(\sum \gamma_n\right) \geq \min\{\partial_r(\gamma_n)\} \geq \partial_r(\Phi_n) + 1 = \left(\frac{1}{2}\right)\min\{\partial_r(\Gamma_n)\} + 1$$

$$= \left(\frac{1}{2}\right)\min\left\{\partial_r\left(\sum \Gamma_n\right)\right\} + 1 = \partial_r(\Phi_n) + 1,$$

completing the proof. □

Proof of Theorem 8. Here the notations $\|\,.\,\|_\Gamma$ and $\langle\,.\,,.\,\rangle_\Gamma$ from section 4 will be used. The strictly convex criterion in (5.1) is a continuous function because H_n is finite-dimensional. Ω_n is a closed convex set as the interception of the closed convex sets Ω, Θ, and H_n. The set Ω_n is nonempty if the dimension n is sufficiently large because it is the closure of $\bigcup_{n=1}^{\infty} H_n$ in $H_+^{2,-1}$ and in A_0 in the correspondent topologies, and assumption A3 is verified. Therefore, if n is sufficiently large, then (5.1) has one and only one solution $\hat{K}_n$.

(a) For all $V \in H_1$, $\|\hat{K}_n\|_\Gamma \leq \|V\|_\Gamma$ because $\Omega_1 \subset \Omega_2 \subset \cdots \subset \Omega_n \subset \cdots \subset \Omega \cap \Theta$. Then the sequence $\{\hat{K}_n, n \in \mathbb{N}\}$ is bounded, which implies the existence of a weakly convergent subsequence that converges weakly in $H_+^{2,-1}$ to a function, denoted here by $\hat{K}_w$ (see the Bolzano–Weierstrass theorem, [29, p. 26]). This subsequence will be denoted by $\{\hat{K}_m, m \in \mathbb{N}\}$. Note that $\hat{K}_w$ depends on the chosen subsequence.

(b) As $\Omega \cap \Theta$ is convex and strongly closed, it is also weakly closed (see the Mazur theorem, [29, p. 20]). Then $\hat{K}_w \in \Omega \cap \Theta$.

(c) $\hat{K}_n$ is a solution of (5.1) if and only if it verifies the following variational inequality:

$$\langle V_m, V_m - \hat{K}_m\rangle_\gamma \geq 0 \;\forall\; V_m \in \Omega_m$$

(see [31, pp. 9–11] or, in a more general setting [32, p. 76]). The weak convergence of $\hat{K}_m$ implies the convergence of the inequality above to the condition

$$\langle V_m, V_m - \hat{K}_w\rangle_\Gamma \geq 0 \;\forall\; V_m \in \Omega_m$$

for each m used in the subsequence. As the sequence of spaces $\{H_n\}$ increases, then $\bigcup_{m=1}^{\infty} H_m$ is a dense subspace of $H_+^{2,-1}$ and $\bigcup_{m=1}^{\infty} \Omega_m \cup \Omega_m$ is a dense subset of $\Omega \cap \Theta$. Taking the limit in the last inequality, we arrive at

$$\langle V, V - \hat{K}_w\rangle_{2,-1} \geq 0 \;\forall\; V \in \Omega \cap \Theta,$$

a necessary and sufficient condition to $\hat{K}_w$ being the solution of (2.3). Then $\hat{K}_w$ equals $\hat{K}$, the solution of (4.1) for any subsequence $\hat{K}_m$ of the sequence $\hat{K}_n$ generated by the Galerkin method, which implies the weak convergence of this sequence to the optimal solution to (4.1). □

Proof of Theorem 9. First, it will be considered the situation where $\breve{K}_n = \breve{K}$, when (5.2) and (5.3) are essentially the same. Second, note that Theorem 8 can be

generalized to the space $H_+^{2,-k}$ without changes, which proves the weak convergence (in $H_+^{2,-k}$) of the sequences generated by (5.2) and (5.3), when $n \in \mathbb{N}$, meaning that $\hat{K}_n$ converges weakly to $\hat{K}$ and $\hat{G}_n + \breve{K}$ converges weakly to $\hat{G} + \breve{K} = \hat{K}$ in $H_+^{2,-k}$. Under the assumptions of Theorem 9 these sequences will converge strongly in $H_+^{2,-k}$ for the same limit. Indeed, the density of $\bigcup_{n=1}^{\infty} H_n$ in $H_+^{2,-k}$ and the fact that $\hat{G} \in \Omega' \cap \Theta'$, a closed convex set, imply that for all positive real numbers ε, there is an integer N such that $\|\hat{G} - G_n\|_\Gamma < \varepsilon$ for all $n > N$ and $G_n \in \Omega_n'$. Thus, by the triangle inequality,

$$\|G_n\|_\Gamma = \|G_n + \hat{G} - \hat{G}\|_\Gamma \leq \|G_n - \hat{G}\|_\Gamma + \|\hat{G}\|_\Gamma < \|\hat{G}\|_\Gamma + \varepsilon.$$

Squaring this expression and recalling the minimizing property of $\hat{G}$ in $\Omega' \cap \Theta' \subset \Omega_n'$, we have

$$(\|\hat{G}\|_\Gamma + \varepsilon)^2 = \|\hat{G}\|_\Gamma^2 + \varepsilon(2\|\hat{G}\|_\Gamma + \varepsilon) > \|G_n\|_\Gamma^2 \geq \|\hat{G}_n\|_\Gamma^2 \geq \|\hat{G}\|_\Gamma^2.$$

When ε tends to zero, $\|\hat{G}_n\|_\Gamma^2$ converges to $\|\hat{G}\|_\Gamma^2$.

Now an argument due to Riesz shows the strong convergence of $\hat{G}_n$ to $\hat{G}$:

$$\|\hat{G}_n - \hat{G}\|_\Gamma^2 = \langle \hat{G}_n - \hat{G}, \hat{G}_n - \hat{G} \rangle_\Gamma = \|\hat{G}_n\|_\Gamma^2 - 2\langle \hat{G}_n, \hat{G} \rangle_\Gamma + \|\hat{G}\|_\Gamma^2,$$

which goes to $\|\hat{G}\|_\Gamma^2 - 2\langle \hat{G}, \hat{G} \rangle_\Gamma + \|\hat{G}\|_\Gamma^2 = 0$ as n goes to ∞ by the weak convergence of $\hat{G}_n$ to $\hat{G}$ and by the norm convergence (showed above). This ends this part of the proof.

The strong convergence of $\hat{K}_n = \hat{G}_n + \breve{K}$ to $\hat{K} = \hat{G} + \breve{K}$ is a consequence of the continuity of the sum in Hilbert spaces.

Now, if $\breve{K}_n$ is the projection of $\breve{K}$ in Ω_n, then $\hat{K}_n = \hat{G}_n + \breve{K}_n$, where $\{\hat{G}_n\}$ is exactly the sequence considered above. As $\hat{G}_n$ converges strongly to $\hat{G}$ and $\breve{K}_n$ converges strongly to $\breve{K}$ (by the continuity of convex projections in Hilbert spaces [32, pp. 157–158]), $\hat{K}_n$ converges strongly to $\hat{K}$ in $H_+^{2,-k}$.

The strong convergence of $\hat{G}_n$ to $\hat{G}$, in the case where $\breve{K}_n$ is the projection of $\breve{K}$ in Ω_n, is now a consequence of the equivalence between (4.2) and (4.3).

To end the proof, note that $\Omega_n \subset H_+^{2,-1}$, which implies that $\hat{K}_n$ belongs to $H_+^{2,-1}$. Then the convergence of $\hat{K}_n$ in $H_+^{2,-k}$ implies the convergence in $H_+^{2,-1}$ to the same limit by the inverse isometry of Theorem 2(a). □

Proofs of Remark 11 *and Theorem* 10. First, note that

$$H_+^\infty \subset H_+^{2,-1} \approx (H_+^{2,-1})' = H_+^{2,1} \subset (H_+^\infty)',$$

H_+^∞ being dense in $H_+^{2,-1}$ and $(H_+^{2,-1})'$ being weak-star dense in $(H_+^\infty)'$ (apply the corollary in [25, p. 298] and T2, XIX, 7; 5, [25, p. 299]). Then $\hat{K}_n$, $\hat{K}$ above can be identified with functions in $(H_+^{2,-1})' \subset (H_+^\infty)'$ by $K \approx G_K(f) = \langle \Phi_{-1}K, \Phi_{-1}f \rangle_2$ for $f \in H_+^\infty$. Second, "$F_n \in (H_+^\infty)'$ converges to $F \in (H_+^\infty)'$ in the weak-star topology" means $F_n(g) \to F(g)$ for all $g \in H_+^\infty$.

If $\hat{K}_n \to \hat{K}$ weakly in $H_+^{2,-1}$, $\langle \Phi_{-1}\hat{K}_n, \Phi_{-1}g \rangle_2 \to \langle \Phi_{-1}\hat{K}, \Phi_{-1}g \rangle_2$ for each $g \in H_+^\infty$, which proves the sequence weak-star convergence in $(H_+^\infty)'$ and Remark 11.

Now, for Theorem 10, let $\hat{K}_n$ converge to $\hat{K}$ strongly in $H_+^{2,-1}$. For each functional $G \in (H_+^\infty)'$, let F_m be a functional sequence in $(H_+^{2,-1})'$ approaching G in the $(H_+^\infty)'$ weak-star topology, i.e., $F_m(g) \to G(g)$ for each $g \in H_+^\infty$. By the Banach–Steinhaus

TABLE 4

Rational functions	Degree	Numerator coefficients	Denominator coefficients
$A(s)$	s^{14}	0	1.000000000000000e+000
	s^{13}	1.381024580093296e+000	1.211595042867158e+002
	s^{12}	1.686008580615162e+000	5.458713034337957e+003
	s^{11}	7.703019538462904e+003	1.125395291285975e+005
	s^{10}	1.634508577704982e+005	1.076974629099070e+006
	s^9	1.671129690051401e+006	4.927775329734212e+006
	s^8	8.766081844981248e+006	1.233200905234427e+007
	s^7	2.566570185137830e+007	1.843165614381086e+007
	s^6	4.288376514277657e+007	1.679171067887532e+007
	s^5	3.638011973651214e+007	8.937537473511269e+006
	s^4	1.128209738056106e+007	2.752841097866921e+006
	s^3	1.881317702823051e+006	4.943826116024184e+005
	s^2	1.129883884492416e+005	4.781899120202310e+004
	s^1	2.747279969974730e+003	1.733418128119080e+003
	s^0	2.330461985378970e+001	1.960830911271398e+001
$B(s)$	s^{10}	0	1.000000000000000e+001
	s^9	−1.000012371305996e+000	1.863782215050616e+001
	s^8	−1.562399952832152e+001	1.181106597540888e+002
	s^7	−1.161078545444695e+002	3.834756797815088e+002
	s^6	−4.837125168666809e+002	7.437879887932461e+002
	s^5	−1.205831279245822e+003	9.129112097836384e+002
	s^4	−1.706884574538163e+003	7.095696057600543e+002
	s^3	−1.220645462582847e+003	3.376552213167196e+002
	s^2	−3.711128214145705e+002	9.277136443643682e+001
	s^1	−3.777832594378928e+001	1.316517110219557e+001
	s^0	−7.343714254348321e−001	7.250551615217723e−001
$F_0(s)$	s^2	0	1.000000000000000e+000
	s^1	−2.488088793672762e+000	−2.263724821234260e+000
	s^0	8.620956412513727e−001	8.843128062448000e−001

theorem [25], the set $\{F_m\}$ is equicontinuous in $H_+^{2,-1}$. Thus $F_n(\hat{K}_n)$ converges to $G(\hat{K})$. Indeed,

$$|G(\hat{K}) - F_n(\hat{K}_n)| \le |F_n(\hat{K}_n) - F_n(\hat{K})| + |F_n(\hat{K}) - G(\hat{K})|,$$

the first term in the right going to zero because $\{F_m\}$ is equicontinuous and $\hat{K}_n \to \hat{K}$ strongly in $H_+^{2,-1}$, the second term in the right going to zero because $F_m(g) \to G(g)$ for each $g \in H_+^\infty$. Therefore, $G(\hat{K}_n)$ converges to $G(\hat{K})$ for each functional $G \in (H_+^\infty)'$, proving the weak convergence in H_+^∞. □

Appendix B. In this appendix we provide the numerical data for the example in section 5. Rational functions described in Table 4 have been calculated from the data $P(s)$, $\phi_w(s)$, $\phi_d(s)$, and $\phi_v(s)$ given in section 5 by solving some diophantine equations (arising from the parameterization of stabilizing controllers), a variable change to reduce the robustness condition to Nehari form was applied, and some cancellations of coincident poles and zeros were made. The first two calculations were performed by state variable methods, as exposed in [6], the cancellation being performed by model order reduction using the Hankel singular value technique. The use of double precision calculations was imperative.

Acknowledgments. The authors are indebted to an unknown referee for his (her) sound criticism and to the encouraging friendship of Professor Carlos Kubrusly.

REFERENCES

[1] G. O. Corrêa, D. M. Sales, and T. M. Soares, *Approximate solutions to H_2-Cost/H_∞-constraint optimization problems*, Internat. J. Control, 61 (1997), pp. 475–491.

[2] D. C. Youla, H. A. Jabr, and J. J. Bongiorno, Jr., *Modern Wiener-Hopf design of optimal controllers, Part* II: *The multivariable case*, IEEE Trans. Automat. Control, 21 (1976), pp. 319–338.

[3] V. Kučera, *Discrete Linear Control: The Polynomial Approach*, Wiley, New York, 1979.

[4] K. Park and J. J. Bongiorno, Jr., *A general theory for the Wiener-Hopf design of multivariable control systems*, IEEE Trans. Automat. Control, 34 (1989), pp. 619–626.

[5] P. P. Khargonekar and M. A. Rotea, *Mixed H_2/H_∞ control: A convex optimization approach*, IEEE Trans. Automat. Control, 36 (1991), pp. 824–837.

[6] K. Zhou, K. Glover, B. Bodenheimer, and J. Doyle, *Mixed H_2 and H_∞ performance objectives* I: *Robust performance analysis*, IEEE Trans. Automat. Control, 39 (1994), pp. 1564–1574.

[7] J. C. Doyle, K. Zhou, K. Glover, and B. Bodenheimer, *Mixed H_2 and H_∞ performance objectives* II: *Optimal control*, IEEE Trans. Automat. Control, 39 (1994), pp. 1575–1587.

[8] G. O. Corrêa and M. A. da Silveira, *On H_2-optimal control of linear systems with tracking/disturbance rejection constraints*, Internat. J. Control, 55 (1992), pp. 1115–1139.

[9] G. O. Corrêa and M. A. da Silveira, *On the design of servomechanisms via H_2-optimization*, Internat. J. Control, 61 (1995), pp. 475–491.

[10] D. S. Bernstein and W. M. Haddad, *LQG control with an H_∞ performance bound: A Riccati equation approach*, IEEE Trans. Automat. Control, 34 (1989), pp. 293–305.

[11] K. Glover and D. Mustafa, *Derivation of the maximum entropy H_∞-controller and a state space formula for its entropy*, Internat. J. Control, 50 (1989), pp. 899–916.

[12] G. Zames and B. A. Francis, *Feedback, minimax sensitivity, and optimal robustness*, IEEE Trans. Automat. Control, 28 (1983), pp. 585–601.

[13] J. C. Doyle, B. A. Francis, and A. R. Tannenbaum, *Feedback Control Theory*, Macmillan, New York, 1992.

[14] S. P. Boyd, V. Balakrishnan, C. G. Barratt, N. M. Kraishi, X. Li, D. G. Meyer, and S. A. Norma, *A new CAD method and associated architectures for linear controllers*, IEEE Trans. Automat. Control, 33 (1988), pp. 268–283.

[15] C. W. Scherer, *Multiobjective H_2/H_∞ control*, in Selected Topics in Identification Modelling and Control, Delft University Press, Holland, 1993, pp. 85–94.

[16] M. Sznaier, *An exact solution to general SISO mixed H_2/H_∞ problems via convex optimization*, IEEE Trans. Automat. Control, 39 (1994), pp. 2511–2517.

[17] H. Rotstein and M. Sznaier, *An exact solution to general 4-blocks discrete-time mixed H_2/H_∞ control problem via convex optimization*, IEEE Trans. Automat. Control, 43 (1988), pp. 1475–1480.

[18] M. Sznaier, H. Rotstein, J. Bu, and A. Sideris, *An exact solution to continuous-time mixed H_2/H_∞ problems*, IEEE Trans. Automat. Control, 45 (2000), pp. 2095–2101.

[19] A. Megretski, *On the order of optimal controllers in mixed H_2/H_∞ control*, in Proceedings of 1998 American Control Conference, Lake Buena Vista, FL, IEEE, 1998, pp. 3173–3174.

[20] C. Scherer, P. Gahinet, and M. Chilali, *Multiobjective output-feedback control via LMI optimization*, IEEE Trans. Automat. Control, 42 (1997), pp. 896–911.

[21] M. A. da Silveira and R. Ades, *Robust Optimal Controllers (Mixed H_2/H_∞ Problem) via Galerkin's Method, Preprints of World IFAC Congress Proceedings*, Beijing, Pergamon Press, Cambridge, UK, 1998.

[22] R. Ades, *Problema H_2/H_∞-soluções Approximadas por Meio de Expansão de Bases*, Doctoral thesis, Electrical Engineering Department, PUC-Rio, Rio de Janeiro, 1999.

[23] P. L. Duren, *The Theory of H^p-Spaces*, Academic Press, New York, 1970.

[24] M. Vidyasagar, *Control System Synthesis*, MIT Press, Cambridge, MA, 1985.

[25] L. Schwartz, *Analyse*, Hermann, Paris, 1970.

[26] V.-K. Khoan, *Distributions, Analyse de Fourier, Opérateurs aux Derivées Partielles*, Tome 2, Vuibert, Paris, 1976.

[27] L. Hormander, *Linear Partial Differential Operators*, Springer-Verlag, Berlin, 1976.

[28] U. Shaked, *A general-transfer function approach to the steady-state LQG stochastic control problem*, Internat. J. Control, 24 (1976), pp. 771–800.

[29] V. Balakrishnan, *Applied Functional Analysis*, Springer-Verlag, Berlin, 1976.

[30] R. F. Curtain and H. J. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, Berlin, 1995.

[31] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-

Verlag, Berlin, 1971.
[32] R. B. Holmes, *A Course on Optimization and Best Approximation*, Lecture Notes in Math. 257, Springer-Verlag, Berlin, 1972.
[33] R. G. Bartle, *The Elements of Integration*, Wiley, New York, 1966.
[34] D. H. Luecking and L. A. Rubel, *Complex Analysis*, Springer-Verlag, Berlin, 1984.
[35] M. G. Safonov and R. Y. Chiang, *CACSD using the state-space L_∞ theory—A design example*, IEEE Trans. Automat. Control, 33 (1988), pp. 477–479.
[36] D. M. Sales, *Controle H_2/H_∞ Soluções Aproximadas Baseadas em Sequências de Problemas Quadráticos*, Dissertação de Mestrado. IME-RJ, Rio de Janeiro, 1994.
[37] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice–Hall, New York, 1996.
[38] J. P. Aubin, *Applied Abstract Analysis*, Wiley, New York, 1977.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 4, pp. 1118–1140

EXISTENCE OF MINIMIZERS FOR NONCONVEX, NONCOERCIVE SIMPLE INTEGRALS*

P. CELADA† AND S. PERROTTA‡

Abstract. We consider the problem of minimizing autonomous, simple integrals such as

$$(\mathcal{P}) \qquad \min\left\{\int_0^T f\left(x(t)\,,x'(t)\right)\,dt\colon\ x\in AC([0\,,T]),\,x(0)=x_0,\,x(T)=x_T\right\},$$

where $f:\mathbb{R}\times\mathbb{R}\to[0,\infty]$ is a possibly nonconvex function with either superlinear or slow growth at infinity. Assuming that the relaxed problem $(\mathcal{P}^{**})$—obtained from $(\mathcal{P})$ by replacing f with its convex envelope f^{**} with respect to the derivative variable x'—admits a solution, we prove attainment for $(\mathcal{P})$ under mild regularity and growth assumptions on f and f^{**}. We discuss various instances of growth conditions on f that yield solutions to the corresponding relaxed problem $(\mathcal{P}^{**})$, and we present examples that show that the hypotheses on f and f^{**} considered here for attainment are essentially sharp.

Key words. nonconvex and noncoercive minimum problem, simple integrals, existence of solutions

AMS subject classifications. 49J05, 49K05

PII. S0363012901387999

1. Introduction.

This paper deals with the existence of solutions to variational problems for autonomous, simple integrals such as

$$(\mathcal{P}) \qquad \min\left\{\int_0^T f(x(t)\,,x'(t))\,dt\colon\ x\in AC([0\,,T]),\,x(0)=x_0,\,x(T)=x_T\right\},$$

where the Lagrangian function $f\colon\mathbb{R}\times\mathbb{R}\to[0\,,\infty]$ is a possibly nonconvex function of its second argument x'. Though the emphasis here is on the lack of convexity of f, we remark that we wish to consider either problems with slow growth, i.e., $f(\eta\,,\xi)$ has no superlinear growth as $|\xi|\to\infty$, or problems with an extended-valued Lagrangian f as it happens in the case of one-sided constraints on derivatives like $x'\geq 0$ or $x'>0$ almost everywhere on $[0\,,T]$.

As is well known, the lack of convexity of $f(\eta\,,\xi)$ with respect to ξ affects the sequential lower semicontinuity of the integral with respect to weak convergence in $AC([0\,,T])$, thus ruling out the possibility of establishing the existence of optimal configurations by means of the direct method of the calculus of variations. However, attainment is a quite typical behavior for variational, simple integrals, and the basic question for nonconvex minimum problems like $(\mathcal{P})$ becomes that of finding which conditions other than convexity of $f(\eta\,,\xi)$ with respect to ξ yield solutions to $(\mathcal{P})$.

This question has been widely investigated in recent years, mainly when f has a special form like $f(\eta\,,\xi)=h(\xi)+g(\eta)$ or $f(\eta\,,\xi)=g(\eta)h(\xi)$ with nonnegative g and h. In either case, a fairly complete understanding of attainment versus nonattainment

*Received by the editors April 16, 2001; accepted for publication (in revised form) January 15, 2002; published electronically October 8, 2002.
http://www.siam.org/journals/sicon/41-4/38799.html

†Dipartimento di Matematica, Università degli Studi di Parma, Via M. D'Azeglio 85, I-43100 Parma, Italy (pietro.celada@unipr.it).

‡Dipartimento di Matematica Pura ed Applicata "G. Vitali," Università degli Studi di Modena e Reggio Emilia, Via Campi 213/B, I-41100 Modena, Italy (perrotta@mail.unimo.it).

phenomena is now available: roughly speaking, attainment occurs provided $g \in \mathcal{C}(\mathbb{R})$ is such that (i) every point $t \in \mathbb{R}$ lies between two intervals where g is monotone, i.e., g does not oscillate too fast, and (ii) g has no strict, local minima. Moreover, well-known Bolza-type examples such as

$$\min\left\{\int_0^T \left[\left(|x'(t)|^2-1\right)^2+|x(t)|^2\right]\,dt\colon\ x\in AC([0\,,T]),\,x(0)=x(T)=0\right\}$$

and

$$\min\left\{\int_0^T \left(1+|x(t)|^2\right)\left[1+\left(|x'(t)|^2-1\right)^2\right]\,dt\colon\ x\in AC([0\,,T]),\,x(0)=x(T)=0\right\}$$

show that attainment is not to be expected to hold in general if the latter assumption on g is dropped, unless h is supposed to be convex at zero, i.e., the values at zero of h and its convex envelope h^{**} coincide (see [12] and [17] for a complete discussion of this issue). Among the many related papers, we refer to [1], [18], [4], [5], [6], [16], [8], and [9] for sum-like integrals and to [14], [15], and [2] for product-like integrals. We mention also the above-mentioned [12] and [17] for a somewhat different point of view on the subject.

As regards the case of nonconvex Lagrangian functions f of general form, we mention [19], [14], [13], [15], and [20]. Roughly speaking, in these papers, assuming that either f and its convex envelope f^{**} with respect to ξ are smooth, attainment for $(\mathcal{P})$ is proved when the continuous function

$$f^{**}(\eta\,,\xi)-\xi\frac{\partial f^{**}}{\partial\xi}(\eta\,,\xi) \tag{1.1}$$

is either monotone or concave as a function of η for every ξ or possibly on the sections with fixed ξ of the set $\{f^{**}<f\}$ only (see [20]). Note that, letting f^* be the polar function of $f(\eta\,,\xi)$ with respect to ξ, the function above coincides with

$$-f^*\left(\eta\,,\frac{\partial f^{**}}{\partial\xi}(\eta\,,\xi)\right), \tag{1.2}$$

i.e., the value at the origin of the supporting affine function to the graph of $\xi\to f^{**}(\eta\,,\xi)$ through the point $(\eta\,,\xi)$. Hence, according to these papers, attainment for $(\mathcal{P})$ seems to require a very special, global (or possibly local as in [20]) behavior of either (1.1) or (1.2) as a function of η for every ξ like monotonicity or concavity. However, in the special cases of variational problems $(\mathcal{P})$ featuring smooth sum-like or product-like Lagrangian functions f, (1.1) and (1.2) turn in

$$\begin{cases}\left[h^{**}(\xi)-\xi\,(h^{**})'\,(\xi)\right]+g(\eta)=-h^*\left((h^{**})'\,(\xi)\right)+g(\eta),\\ g(\eta)\left[h^{**}(\xi)-\xi\,(h^{**})'\,(\xi)\right]=g(\eta)\left[-h^*\left((h^{**})'\,(\xi)\right)\right],\end{cases}\qquad(\eta\,,\xi)\in\mathbb{R}\times\mathbb{R},$$

respectively. Hence, the monotonicity or concavity assumptions on (1.1) and (1.2) as functions of η reduce to the requirement that g share the same property on the whole real line in cases of sum-like integrals and that g be monotone or possibly convex, provided $h^*\left((h^{**})'\,(\xi)\right)\geq 0$ for every ξ in the product-like case. By contrast, the existence results mentioned before for these special problems call only for weaker properties of g, namely, no oscillations on increasingly smaller scales and no strict local minima.

Thus, there is a gap between the available attainment results for sum- or product-like, nonconvex, variational problems on one hand and the same problems with a Lagrangian f of general form on the other hand, and the aim of this paper is precisely to fill this gap. Indeed, we are going to show that the hypotheses on f^{**} that yield attainment for $(\mathcal{P})$ in the general case actually look even weaker than they appear in the case of sum-like or product-like integrals.

To this aim, provided f enjoys mild regularity and growth assumptions (see Theorem 2.2), we associate with the convex envelope f^{**} of f with respect to ξ a function $Ef^{**}\colon \mathbb{R}\times\mathbb{R} \to [-\infty, \infty]$ whose value at a point (η, ξ) is, roughly speaking, the value at the origin of the supporting affine function to the graph of $\xi \to f^{**}(\eta, \xi)$ through the point (η, ξ) and which reduces to (1.1) and (1.2) for smooth convex envelopes f^{**}. Then, assuming also that the relaxed problem $(\mathcal{P}^{**})$ obtained from $(\mathcal{P})$ by replacing f with its convex envelope f^{**}, i.e.,

$$(\mathcal{P}^{**}) \qquad \min\left\{\int_0^T f^{**}(x(t), x'(t))\,dt\colon\ x \in AC([0,T]),\ x(0) = x_0,\ x(T) = x_T\right\},$$

admits a solution, we prove attainment for $(\mathcal{P})$ provided Ef^{**} and f^{**} have the following qualitative, local behavior on the set $\{f^{**} < f\}$:

(i) If $f^{**}(\eta_0, \xi_0) < f(\eta_0, \xi_0)$, there is $\delta = \delta(\eta_0, \xi_0) > 0$ such that the restriction $\eta \to Ef^{**}(\eta, \xi_0)$ is monotone on both intervals $[\eta_0 - \delta, \eta_0]$ and $[\eta_0, \eta_0 + \delta]$;

and, whenever the section of $\{f^{**} < f\}$ with $\xi = 0$ is not empty,

(ii) the function $\eta \to f^{**}(\eta, 0)$ has no strict, local minima on such sections.

We wish to point out that, in the Bolza-type examples mentioned above, the set $\{f^{**} < f\}$ is given in either case by $\mathbb{R}\times(-1, 1)$, that $f^{**}(\eta, 0)$ is given by η^2 and $1 + \eta^2$, respectively, and that all the other assumptions of our result are satisfied. Thus, nonattainment for those problems is a direct consequence of the failure of (ii).

We refer to section 2 for the exact statement of our result, a more detailed discussion of its hypotheses, and some examples.

Finally, we wish to remark that the existence result for the nonconvex problem $(\mathcal{P})$ we are going to prove is based on the assumption of attainment for the corresponding relaxed problem $(\mathcal{P}^{**})$ and thereby can be applied to nonconvex problems featuring either superlinear or slow growth at infinity, provided the associated relaxed problem admits a solution. Indeed, besides the standard case of functions f having superlinear growth at infinity (see Corollary 2.3) for which the existence of solutions for the corresponding relaxed problem $(\mathcal{P}^{**})$ follows immediately from the direct method of the calculus of variations, we consider also the case of functions f with slow growth at infinity (see Corollary 2.4) for which attainment for the relaxed problem $(\mathcal{P}^{**})$ can be obtained by applying the existence result of [7].

The remaining part of the paper is organized as follows. In the next section, we introduce some notation, we recall some well-known preliminary results, and we state the main result (Theorem 2.2) and prove its consequences (Corollaries 2.3 and 2.4). Then, in section 3, we prove some technical results that will be needed in the proof of Theorem 2.2, presented in section 4.

2. Notation and statement of the main results. We begin by recalling some elementary definitions, notation, and results, mostly from convex analysis and measure theory.

If $A \subset \mathbb{R}^n$, we let $\operatorname{int}(A)$, $\overline{A}$, and ∂A be the interior, the closure, and the boundary of A, respectively.

The *effective domain* of a function $g\colon A \to (-\infty\,,\infty]$ is the subset of A defined by

$$\operatorname{dom}(g) = \{\xi \in A\colon\ g(\xi) < \infty\}\,,$$

and g itself is said to be *proper* whenever its effective domain is not empty. Now, let $g\colon \mathbb{R} \to [0\,,\infty]$ be a proper, lower semicontinuous function. We recall that g is said to be *subdifferentiable* at a point $\xi \in \operatorname{dom}(g)$ if there exists $d \in \mathbb{R}$ such that

$$g(\zeta) \geq g(\xi) + d(\zeta - \xi), \qquad \zeta \in \mathbb{R}. \tag{2.1}$$

Every such d is a *subgradient* of g at ξ, and the set of all such numbers d is the *subdifferential* $\partial g(\xi)$ of g at ξ. When g is also convex, $\partial g(\xi)$ is a nonempty, compact interval for every $\xi \in \operatorname{int}(\operatorname{dom}(g))$, and g turns out to be locally Lipschitz continuous on $\operatorname{int}(\operatorname{dom}(g))$ so that $\partial g(\xi) = \{g'(\xi)\}$ for a.e. $\xi \in \operatorname{int}(\operatorname{dom}(g))$.

We recall also that if $g\colon \mathbb{R} \to [0\,,\infty]$ is proper and lower semicontinuous, the *polar function* of g is the proper, lower semicontinuous convex function $g^*\colon \mathbb{R} \to (-\infty\,,\infty]$ defined by

$$g^*(\zeta) = \sup\{\xi\zeta - g(\xi)\colon\ \xi \in \mathbb{R}\}\,, \qquad \zeta \in \mathbb{R}$$

(see [11]), and that the *bipolar function* or *convex envelope* of g is the polar $g^{**}\colon \mathbb{R} \to [0\,,\infty]$ of g^*. Thus, g^{**} is a proper, lower semicontinuous convex function such that

(2.2) $g^{**}(\xi) \leq g(\xi)$ for every $\xi \in \mathbb{R}$;

(2.3) $g^{**}(\xi) = g(\xi)$ for every $\xi \in \mathbb{R} \setminus \operatorname{int}(\operatorname{dom}(g^{**}))$;

(2.4) the set $\{g^{**} < g\}$ is open;

(2.5) g^{**} is affine on the connected components of $\{g^{**} < g\}$;

(2.6) the closure of each connected component of the set $\{g^{**} < g\}$ is contained in $\operatorname{dom}(g^{**})$.

Moreover, we recall that, whenever $d \in \mathbb{R}$ is a subgradient of g^{**} at some point $\xi \in \operatorname{dom}(g^{**})$, the values of $g^{**}(\xi)$ and $g^*(d)$ are related by

$$g^{**}(\xi) + g^*(d) = d\xi \tag{2.7}$$

(see [11]) because of the equality $g^{***} = g^*$. Hence, writing (2.1) with g^{**} instead of g, it follows that the value at the origin of the supporting affine function to the graph of g^{**} through the point $(\xi\,, g^{**}(\xi))$ with slope d is given by $-g^*(d)$.

As to measure theoretic notations and results, we denote the Lebesgue measure of a measurable subset E of $\mathbb{R}$ by $|E|$, and we recall that a family of nondegenerate, compact intervals $\mathcal{K}$ is said to *shrink* at some point $x \in \mathbb{R}$ if $x \in K$ for every $K \in \mathcal{K}$ and $\inf\{|K|\colon\ K \in \mathcal{K}\} = 0$. We recall also that a *Vitali covering* of a measurable set E is a family of nondegenerate, compact intervals $\mathcal{K}$ such that, for a.e. $x \in E$, the subfamily of those intervals $K \in \mathcal{K}$ containing x shrinks at x itself. We emphasize that in the definitions above, the intervals K associated with x need not be centered at x or nested. Then Vitali's covering theorem states that every such covering contains an (at most) countable subfamily of sets $\{K_n\}_n$ consisting of pairwise disjoint intervals that cover E up to a negligible set, i.e., $|E \setminus (\cup_n K_n)| = 0$. We also recall that a point $x \in \mathbb{R}$ is said to be a *density point* for a measurable set E if

$$\frac{1}{2\varepsilon}\,|(x-\varepsilon\,, x+\varepsilon) \cap E| \to 1 \qquad \text{or, equivalently,} \qquad \frac{1}{2\varepsilon}\,|(x-\varepsilon\,, x+\varepsilon) \setminus E| \to 0$$

as $\varepsilon \to 0_+$ and that Lebesgue's differentiation theorem states that almost every point of E is a density point. It is plain that for every such point x,

$$\frac{|K \setminus E|}{|K|} \to 0 \qquad \text{as } K \in \mathcal{K} \text{ and } |K| \to 0 \tag{2.8}$$

whenever $\mathcal{K}$ shrinks at x.

As regards functional theoretic notations, we let T be a positive number, use standard notations for the Lebesgue space of integrable functions on $[0\,,T]$ and its norm, and write $AC([0\,,T])$ for the space of absolutely continuous functions on $[0\,,T]$, which turns out to be a Banach space with respect to the Sobolev norm

$$\|x\|_{1,1} = \int_0^T \left[|x(t)| + |x'(t)|\right] dt, \qquad x \in AC([0\,,T]).$$

We also denote the space of all smooth, compactly supported functions on the real line by $\mathcal{D}(\mathbb{R})$.

Now, we introduce the class of functionals we are going to consider in what follows. Given a proper and lower semicontinuous function $f\colon \mathbb{R}\times\mathbb{R} \to [0\,,\infty]$ we consider the integral functional

$$I(x) = \int_0^T f(x(t)\,,x'(t))\, dt, \qquad x \in AC([0\,,T]),$$

and the associated minimum problem

$$(\mathcal{P}) \qquad \min\left\{I(x)\colon\ x \in AC([0\,,T]) \text{ with } x(0) = x_0 \text{ and } x(T) = x_T\right\}$$

with x_0, $x_T \in \mathbb{R}$. We denote the polar and the bipolar functions of f with respect to the second variable ξ by $f^*\colon \mathbb{R}\times\mathbb{R} \to (-\infty\,,\infty]$ and $f^{**}\colon \mathbb{R}\times\mathbb{R} \to [0\,,\infty]$, respectively, and, for every $\eta \in \mathbb{R}$, we denote also the subdifferential of $\xi \to f^{**}(\eta\,,\xi)$ at the point $\xi \in \mathbb{R}$ by $\partial f^{**}(\eta\,,\xi)$. Then we consider the functional

$$I^{**}(x) = \int_0^T f^{**}(x(t)\,,x'(t))\, dt, \qquad x \in AC([0\,,T]),$$

and the associated minimum problem

$$(\mathcal{P}^{**}) \qquad \min\left\{I^{**}(x)\colon\ x \in AC([0\,,T]) \text{ with } x(0) = x_0 \text{ and } x(T) = x_T\right\},$$

which we loosely refer to as the relaxed functional and the relaxed minimum problem, respectively. It is plain that $I^{**} \le I$ on $AC([0\,,T])$ so that any solution x to $(\mathcal{P}^{**})$ satisfying $f^{**}(x\,,x') = f(x\,,x')$ almost everywhere on $[0\,,T]$ is a solution to $(\mathcal{P})$ as well. Moreover, I^{**} is sequentially weakly lower semicontinuous on the set of competing functions $\{x \in AC([0\,,T])\colon\ x(0) = x_0 \text{ and } x(T) = x_T\}$.

Throughout this paper, we shall consider the following assumptions on the functions f and f^{**}:

(H1) $\operatorname{dom}(f) = \operatorname{dom}(f^{**}) = \mathbb{R}\times C$, where C is a nondegenerate interval;

(H2) f and f^{**} are continuous on $\mathbb{R}\times\operatorname{int}(C)$.

If the function f satisfies (H1), it follows that $f^{**}(\eta,\xi) = f(\eta,\xi)$ for every $\eta \in \mathbb{R}$ and $\xi \in \mathbb{R} \setminus \operatorname{int}(C)$ because of (2.3); moreover, if f also satisfies (H2), the detachment set $\mathcal{D}$ defined by

$$\mathcal{D} = \{(\eta,\xi) \in \mathbb{R}\times\mathbb{R}\colon\ f^{**}(\eta,\xi) < f(\eta,\xi)\} \tag{2.9}$$

is an open subset of $\mathbb{R} \times \operatorname{int}(C)$. In what follows, we shall denote the sections of the detachment set $\mathcal{D}$ with η and ξ fixed by $\mathcal{D}_\eta = \{\xi \in \mathbb{R}\colon\ (\eta,\xi) \in \mathcal{D}\}$ and $\mathcal{D}^\xi = \{\eta \in \mathbb{R}\colon\ (\eta,\xi) \in \mathcal{D}\}$, respectively.

Now, consider the mapping $Ef^{**}\colon \mathbb{R}\times\mathbb{R} \to [-\infty,\infty]$ defined by

$$Ef^{**}(\eta,\xi) = \sup\{-f^*(\eta,d)\colon\ d \in \partial f^{**}(\eta,\xi)\}, \qquad (\eta,\xi) \in \mathbb{R}\times\mathbb{R},$$

where, as usual, the supremum of the empty set is set equal to $-\infty$. Note that if it happens that f^{**} is smooth, say $f^{**} \in \mathcal{C}^1(\mathbb{R}\times\mathbb{R})$, then Ef^{**} reduces to the continuous function already considered in (1.1), i.e.,

$$Ef^{**}(\eta,\xi) = f^{**}(\eta,\xi) - \xi\frac{\partial f^{**}}{\partial \xi}(\eta,\xi), \qquad (\eta,\xi) \in \mathbb{R}\times\mathbb{R},$$

because of the basic equality (2.7).

On the function f, we shall consider also the following growth assumption:

$$\lim_{|\xi|\to\infty} \sup\{Ef^{**}(\eta,\xi)\colon\ |\eta| \le R\} = -\infty \qquad \text{for every } R \ge 0. \tag{H3}$$

Note that all functions f satisfying (H2) and (H3) have the following property: For every positive number R, there exist two numbers $\alpha > 0$ and $\beta \ge 0$ depending on R such that

$$f^{**}(\eta,\xi) \ge \alpha|\xi| - \beta \qquad \text{for every } |\eta| \le R \text{ and every } \xi \in \mathbb{R}. \tag{2.10}$$

The growth condition (H3) is strictly weaker than superlinearity at infinity. Indeed, it is easy to see that if a proper and lower semicontinuous function $f\colon \mathbb{R}\times\mathbb{R} \to [0,\infty]$ satisfies (H1) and (H2) and has the further property that, for every given $R \ge 0$, $f(\eta,\xi) \ge \theta(|\xi|)$ for every $|\eta| \le R$ and $\xi \in \mathbb{R}$ for some suitable function $\theta\colon [0,\infty) \to \mathbb{R}$ depending on R such that $\theta(|\xi|)/|\xi| \to \infty$ as $|\xi| \to \infty$, then (H3) is also satisfied (see [7], for instance). By contrast, the function

$$f(\eta,\xi) = f(\xi) = |\xi| - \log(1+|\xi|), \qquad \xi \in \mathbb{R},$$

provides a simple example of a convex function satisfying (H3) and having linear growth at infinity. We refer to [3] for interesting results on the relationship between (H3) and the regularity of solutions to $(\mathcal{P}^{**})$.

The properties of the restriction of the function Ef^{**} to the detachment set $\mathcal{D}$ are gathered in the following proposition.

PROPOSITION 2.1. *Let $f\colon \mathbb{R}\times\mathbb{R} \to [0,\infty]$ be a proper and lower semicontinuous function satisfying* (H1) *and* (H2) *and let $\mathcal{D}$ be the detachment set defined by* (2.9). *Then*

(i) *there exists $d\colon \mathcal{D} \to \mathbb{R}$ such that $\partial f^{**}(\eta,\xi) = \{d(\eta,\xi)\}$ for every $(\eta,\xi) \in \mathcal{D}$;*
(ii) *$Ef^{**}(\eta,\xi) = -f^*(\eta, d(\eta,\xi))$ for every $(\eta,\xi) \in \mathcal{D}$;*
(iii) *Ef^{**} is finite-valued on $\mathcal{D}$ and both Ef^{**} and d are continuous on $\mathcal{D}$;*

(iv) *the restrictions* $\xi \in \mathcal{D}_\eta \to Ef^{**}(\eta,\xi)$ *and* $\xi \in \mathcal{D}_\eta \to d(\eta,\xi)$ *are constant on the connected components of* $\mathcal{D}_\eta$*;*

(v) *if* $\mathcal{D}^0 \neq \emptyset$*, then* $Ef^{**}(\eta,0) = f^{**}(\eta,0)$ *for every* $\eta \in \mathcal{D}^0$.

Proof. For every nonempty section $\mathcal{D}_\eta$ the function $\xi \in \mathbb{R} \to f^{**}(\eta,\xi)$ is affine on the connected components of $\mathcal{D}_\eta$ because of (2.5). Hence, it is differentiable at every point ξ of $\mathcal{D}_\eta$ so that (i) holds with $d(\eta,\xi)$ given by the partial derivative of f^{**} with respect to ξ at the point (η,ξ) and (ii) follows from (i) and the definition of Ef^{**}. Then recall that

$$f^*(\eta, d(\eta,\xi)) = \xi d(\eta,\xi) - f^{**}(\eta,\xi), \qquad (\eta,\xi) \in \mathcal{D}, \tag{2.11}$$

because of (2.7). The right-hand side of this equality is finite since $\mathcal{D} \subset \mathbb{R}\times\mathrm{int}\,(C) \subset \mathrm{dom}\,(f^{**})$ and, moreover, f^{**} is continuous on $\mathcal{D}$ because of (H2). As to d, its restriction $\xi \in \mathcal{D}_\eta \to d(\eta,\xi)$ is constant on the connected components of $\mathcal{D}_\eta$ so that for every rectangle $Q = [\eta_1,\eta_2]\times[\xi_1,\xi_2]$ contained in the detachment set $\mathcal{D}$, we have

$$d(\eta,\xi) = \frac{f^{**}(\eta,\xi_1) - f^{**}(\eta,\xi_2)}{\xi_2 - \xi_1}, \qquad (\eta,\xi) \in Q. \tag{2.12}$$

Thus, d too is continuous on $\mathcal{D}$ and (iii) and (iv) follow from (ii), (2.11), and (2.12). Finally, (v) follows immediately from (i) and (2.11). ∎

After these preliminaries, we can state the main result of the paper.

THEOREM 2.2. *Let* $f\colon \mathbb{R}\times\mathbb{R} \to [0,\infty]$ *be a proper and lower semicontinuous function satisfying* (H1)*,* (H2)*, and* (H3)*. Assume also that the following properties hold:*

(2.13) *For every* $(\eta_0,\xi_0) \in \mathcal{D}$*, there is* $\delta = \delta(\eta_0,\xi_0) > 0$ *such that* $[\eta_0-\delta,\eta_0+\delta] \subset \mathcal{D}^{\xi_0}$ *and such that the restriction* $\eta \in [\eta_0-\delta,\eta_0+\delta] \to Ef^{**}(\eta,\xi_0)$ *is monotone on each interval* $[\eta_0-\delta,\eta_0]$ *and* $[\eta_0,\eta_0+\delta]$*;*

and, if $\mathcal{D}^0 \neq \emptyset$*,*

(2.14) *the restriction* $\eta \in \mathcal{D}^0 \to f^{**}(\eta,0)$ *has no strict, local minima on* $\mathcal{D}^0$.

Then, if the relaxed problem $(\mathcal{P}^{**})$ *has a solution, the nonconvex problem* $(\mathcal{P})$ *has a solution too.*

As already pointed out in section 1, the hypothesis (2.14) cannot be dropped without affecting attainment for $(\mathcal{P})$.

Then we complete the previous result by presenting two instances of growth hypotheses on f ensuring the existence of solutions to the relaxed problem $(\mathcal{P}^{**})$ and hence to the nonconvex problem $(\mathcal{P})$ by Theorem 2.2. The first one is the familiar case of functions f having superlinear growth at infinity, whereas the second, a simple application of the existence result of [7], applies to problems featuring Lagrangian functions f with slow growth at infinity. We wish to remark that both results apply to nonconvex problems featuring one-sided constraints on a derivative like $x' \geq 0$ or $x' > 0$ almost everywhere on $[0,T]$.

COROLLARY 2.3. *Let* $f\colon \mathbb{R}\times\mathbb{R} \to [0,\infty]$ *be a proper and lower semicontinuous function such that all the hypotheses of Theorem* 2.2 *hold with* (H3) *replaced by the following:*

(2.15) $f(\eta,\xi) \geq \alpha|\xi| - \beta$*,* $(\eta,\xi) \in \mathbb{R}\times\mathbb{R}$*, for some* $\alpha > 0$ *and* $\beta \geq 0$*;*

(2.16) *for every* $R \geq 0$*, there exists* $\theta\colon [0,\infty) \to \mathbb{R}$ *such that* $f(\eta,\xi) \geq \theta(|\xi|)$ *for every* $|\eta| \leq R$ *and* $\xi \in \mathbb{R}$ *and* $\theta(|\xi|)/|\xi| \to \infty$ *as* $|\xi| \to \infty$*.*

Then the nonconvex problem $(\mathcal{P})$ *admits (at least) a solution for every boundary data* $x_0,\ x_T \in \mathbb{R}$.

Proof. First, recall that (2.16) implies (H3) so that all the hypotheses of Theorem 2.2 hold. Then assume there is some feasible function $\overline{x} \in AC([0\,,T])$ such that $I^{**}(\overline{x}) = c < \infty$; otherwise there is nothing to prove, and set

$$\mathcal{A} = \{x \in AC([0\,,T])\colon\ x(0) = x_0,\ x(T) = x_T \text{ and } I^{**}(x) \le c\}\,.$$

All functions $x \in \mathcal{A}$ are uniformly bounded because of (2.15). Hence, I^{**} is coercive on $\mathcal{A}$ by (2.16) and lower semicontinuous on the same set with respect to weak convergence in $AC([0\,,T])$ (see Theorem 2.1, Chapter 8 in [11], for instance). Thus, $(\mathcal{P}^{**})$ admits a solution and the conclusion follows from Theorem 2.2. □

COROLLARY 2.4. *Let* $f\colon \mathbb{R}\times\mathbb{R} \to [0\,,\infty]$ *be a proper and lower semicontinuous function such that all the hypotheses of Theorem* 2.2 *hold. Assume also that*

(2.17) C *is a cone;*

(2.18) $\partial f^{**}(\eta\,,\xi) \neq \emptyset$ *for every* $(\eta\,,\xi) \in \mathbb{R}\times C$;

(2.19) $f(\eta\,,\xi) \ge \alpha\,|\xi| - \beta$, $(\eta\,,\xi) \in \mathbb{R}\times\mathbb{R}$, *for some* $\alpha > 0$ *and* $\beta \ge 0$.

Then the nonconvex problem $(\mathcal{P})$ *admits (at least) a solution for every boundary data* $x_0,\ x_T \in \mathbb{R}$.

Recall that a cone in $\mathbb{R}$ is either $\mathbb{R}$ itself or any open or closed half line starting at zero. Note that (2.18) is automatically fulfilled if C is open, and recall also that the detachment set $\mathcal{D}$ is contained in $\mathbb{R}\times\operatorname{int}(C)$. Hence, unless C is the whole real line, the section $\mathcal{D}^0$ of $\mathcal{D}$ with $\xi = 0$ is empty so that (2.14) is automatically fulfilled too.

Proof of Corollary 2.4. The very same computations of [2, Corollary 1.4] show that the hypotheses of the existence result of [7] hold for the relaxed problem $(\mathcal{P}^{**})$. Thus, $(\mathcal{P}^{**})$ admits a solution and the conclusion follows from Theorem 2.2. □

Finally, we end this section by presenting two examples of nonconvex problems which the previous results apply to. They are not meant to be meaningful from the point of view of applications. We just want to illustrate the scope of application of the previous results by showing examples of problems to which the previously known attainment results do not apply.

Example 2.5. Let $f\colon \mathbb{R}\times\mathbb{R} \to [0\,,\infty)$ be defined by

$$f(\eta\,,\xi) = [\xi - a(\eta)]^2\,[\xi - b(\eta)]^2 + c(\eta), \qquad (\eta\,,\xi) \in \mathbb{R}\times\mathbb{R},$$

where the coefficients $a\,,b\,,c \in \mathcal{C}(\mathbb{R})$ are such that

$$c(\eta) \ge \alpha \max\{|a(\eta)|\,,|b(\eta)|\}\,, \qquad \eta \in \mathbb{R}, \tag{2.20}$$

for some $\alpha > 0$. Note that if a and b are bounded, the condition above can always be satisfied by any lower bounded function c by possibly adding a positive constant to c itself. Here, we obviously assume that $a(\eta_0) \neq b(\eta_0)$ for some η_0; otherwise f would be convex with respect to ξ and we can also assume without loss of generality that $a(\eta) \le b(\eta)$ for every $\eta \in \mathbb{R}$.

Then the convex envelope f^{**} of f with respect to ξ is given by

$$f^{**}(\eta\,,\xi) = \begin{cases} [\xi - a(\eta)]^2\,[\xi - b(\eta)]^2 + c(\eta), & \xi \le a(\eta) \text{ or } \xi \ge b(\eta), \\ c(\eta), & a(\eta) \le \xi \le b(\eta), \end{cases} \qquad \eta \in \mathbb{R},$$

and the detachment set $\mathcal{D}$ by $\mathcal{D} = \{(\eta\,,\xi)\colon\ a(\eta) < \xi < b(\eta)\}$.

It is clear that (H1) and (H2) hold. Moreover, it is easy to check that (2.20) yields (2.15) and that, for every $R \geq 0$, $f(\eta, \xi) \geq \xi^4/2 - M$ for every $|\eta| \leq R$ and every $\xi \in \mathbb{R}$ for some suitable $M \geq 0$ depending on R.

Now, note that $Ef^{**}(\eta, \xi) = c(\eta)$ for every $(\eta, \xi) \in \mathcal{D}$ so that the main hypotheses (2.13) and (2.14) of Theorem 2.2 are satisfied, for instance, by every smooth function c whose derivative has only isolated zeros and which has no strict, local minima on $\mathcal{D}^0 = \{a < 0\} \cap \{b > 0\}$. In such cases, all the hypotheses of Corollary 2.3 are satisfied and the corresponding nonconvex minimum problem $(\mathcal{P})$ has at least one solution.

Example 2.6. Let $f\colon \mathbb{R} \times \mathbb{R} \to [0, \infty]$ be defined by

$$f(\eta, \xi) = \begin{cases} \infty & \text{for } \eta \in \mathbb{R} \text{ and } \xi \leq 0, \\ \xi - \log \xi + a(\eta) e^{-b(\eta)[\xi - c(\eta)]^2} & \text{for } \eta \in \mathbb{R} \text{ and } \xi > 0, \end{cases}$$

where the coefficients $a, b, c \in \mathcal{C}(\mathbb{R})$ are positive functions. For suitable choices of a, b, and c, f fails to be convex with respect to ξ.

It is plain that (H1) and (H2) hold with $C = (0, \infty)$ so that (2.17) and (2.18) obviously follow. Moreover, the growth assumption (2.19) holds too; choose $\alpha = 1/2$ and $\beta = 0$, for instance. As to (H3), note that Ef^{**} off the set $\mathcal{D}$ is given by

$$Ef^{**}(\eta, \xi) = 1 - \log \xi + a(\eta) \{1 + 2b(\eta)\xi [\xi - c(\eta)]\} e^{-b(\eta)[\xi - c(\eta)]^2}, \qquad (\eta, \xi) \notin \mathcal{D}.$$

Since it is easy to check that, for every $R \geq 0$, there is $M = M(R) > 0$ such that $f(\eta, \xi) = f^{**}(\eta, \xi)$ for every $|\eta| \leq R$ and every $\xi \geq M$, we conclude that (H3) holds.

Now, for every $(\eta, \xi) \in \mathcal{D}$, there is $\xi' > \xi$ such that $(\eta, \xi') \in \partial\mathcal{D}$ and $Ef^{**}(\eta, \xi) = Ef^{**}(\eta, \xi')$. Therefore, (2.13) holds true provided a, b, and c have the appropriate behavior; for instance, they are smooth with only isolated zeros of the derivatives. As $C = (0, \infty)$, the section $\mathcal{D}^0$ of the detachment set $\mathcal{D}$ is empty and the existence of solutions to the corresponding nonconvex problem $(\mathcal{P})$ follows from Corollary 2.4.

3. Some technical results. The proof of our attainment result (Theorem 2.2) is based on the following idea: Let z be a solution to the relaxed problem $(\mathcal{P}^{**})$ and let t be a differentiability point of z such that $(z(t), z'(t))$ lies in the detachment set $\mathcal{D}$. Then we locally modify z around this point t, thus finding a family of new solutions to $(\mathcal{P}^{**})$ which have the further property that they lie, together with their derivatives, on the "boundary" of the detachment set $\mathcal{D}$, i.e., where f and f^{**} coincide, almost everywhere on shrinking neighborhoods of the point t. Then a covering argument allows us to select and glue some of these new solutions so as to find a further new solution x to $(\mathcal{P}^{**})$ satisfying $f^{**}(x, x') = f(x, x')$ almost everywhere on $[0, T]$, thus proving attainment for $(\mathcal{P})$.

The main steps towards the proof of Theorem 2.2 are gathered in this section. Indeed, the program outlined above calls first for a local description of the "boundary" of the detachment set $\mathcal{D}$ which is given in Proposition 3.1 below and then calls for defining new solutions to $(\mathcal{P}^{**})$ which stay locally on the "boundary" of $\mathcal{D}$. These latter functions will be defined as extremal solutions to suitable, convex-valued differential inclusions related to the detachment set $\mathcal{D}$.

PROPOSITION 3.1. *Let $f\colon \mathbb{R} \times \mathbb{R} \to [0, \infty]$ be a proper and lower semicontinuous function satisfying* (H1), (H2), *and* (H3). *Then, for every $(\eta_0, \xi_0) \in \mathcal{D}$, there exists $\delta = \delta(\eta_0, \xi_0) > 0$ and two functions $a, b\colon [\eta_0 - \delta, \eta_0 + \delta] \to \mathbb{R}$ such that*

(i) *$a(\eta) < \xi_0 < b(\eta)$ for every $\eta \in [\eta_0 - \delta, \eta_0 + \delta]$;*

(ii) *a and b are bounded, upper and lower semicontinuous functions, respectively;*

(iii) $\{(\eta\,,\xi)\colon\ |\eta-\eta_0|\le\delta$ *and* $a(\eta)<\xi<b(\eta)\}\subset\mathcal{D}$*;*

(iv) $f^{**}(\eta\,,a(\eta))=f(\eta\,,a(\eta))<\infty$ *and* $f^{**}(\eta\,,b(\eta))=f(\eta\,,b(\eta))<\infty$ *hold for every* $\eta\in[\eta_0-\delta\,,\eta_0+\delta]$.

In other words, (iii) and (iv) of Proposition 3.1 say that every connected component of every sufficiently narrow, vertical strip of $\mathcal{D}$ is the plane set contained between the graphs of two functions a and b satisfying (i) and (ii).

Proof of Proposition 3.1. Let (η_0,ξ_0) be a point of $\mathcal{D}$, choose $\delta=\delta(\eta_0,\xi_0)>0$ such that $(\eta\,,\xi_0)\in\mathcal{D}$ when $|\eta-\eta_0|\le\delta$, and consider the corresponding nonempty, open sections $\mathcal{D}_\eta$.

Since $\xi\in\mathcal{D}_\eta\ \to\ Ef^{**}(\eta\,,\xi)$ is constant on the connected components of $\mathcal{D}_\eta$ by (iv) of Proposition 2.1, the growth assumption (H3) implies that $\mathcal{D}_\eta$ has bounded connected components. Hence, the functions a and b defined by

$$\begin{cases} a(\eta)=\inf\,\{\xi\le\xi_0\colon\ f^{**}(\eta\,,\xi')<f(\eta\,,\xi')\text{ for every }\xi'\in[\xi\,,\xi_0]\}\,,\\ b(\eta)=\sup\,\{\xi\ge\xi_0\colon\ f^{**}(\eta\,,\xi')<f(\eta\,,\xi')\text{ for every }\xi'\in[\xi_0,\xi]\}\,, \end{cases}\qquad |\eta-\eta_0|\le\delta,$$

are finite and properties (i) and (iii) hold by construction. Moreover, the open interval $\big(a(\eta)\,,b(\eta)\big)$ is the connected component of $\mathcal{D}_\eta$ containing ξ_0 whence we obtain the equalities in (iv). Also, the closure of every connected component of $\mathcal{D}_\eta$ is contained in C by (2.6) so that f and f^{**} are finite at the points $(\eta\,,a(\eta))$ and $(\eta\,,b(\eta))$ for every η within δ from η_0 because of (H1).

Thus, we are left to prove (ii). As regards semicontinuity, suppose, for instance, that b fails to be lower semicontinuous at some point η' in $[\eta_0-\delta\,,\eta_0+\delta]$ so that

$$\liminf_{\eta\to\eta'}b(\eta)<M<b(\eta')$$

for some real number $M>\xi_0$. It follows that $f^{**}(\eta',\xi)<f(\eta',\xi)$ for every $\xi\in[\xi_0,M]$. Hence, the compact segment $\{\eta'\}\times[\xi_0,M]$ is contained in the open set $\mathcal{D}$ whence $[\eta'-\sigma\,,\eta'+\sigma]\times[\xi_0,M]$ is in $\mathcal{D}$ too for some positive σ. This yields that $b(\eta)\ge M$ for all η such that $|\eta-\eta'|\le\sigma$ and $|\eta-\eta_0|\le\delta$, and this gives a contradiction. Finally, to prove that a and b are bounded, note that (iii) and (iv) of Proposition 2.1 imply that $Ef^{**}(\eta\,,\xi)\ge-M$ for every $\xi\in\big(a(\eta)\,,b(\eta)\big)$ and every $\eta\in[\eta_0-\delta\,,\eta_0+\delta]$ for some $M\ge0$ and that the growth assumption (H3) yields that

$$\sup\,\{Ef^{**}(\eta\,,\xi)\colon\ |\eta-\eta_0|\le\delta\}<-M,\qquad\qquad |\xi|\ge R,$$

for some large enough R. Thus, $-R<a(\eta)<b(\eta)<R$ for every $|\eta-\eta_0|\le\delta$, completing the proof. □

The following lemma is proved in [2]. It is a technical result whose statement is long though its proof is elementary.

LEMMA 3.2. *Let* $z\in AC([0\,,T])$ *be differentiable at some point* $s\in(0\,,T)$ *and let* α, $\beta\in\mathbb{R}$ *be such that*

$$\alpha<z'(s)<\beta.$$

Then for every $\delta>0$*, there exist* $\varepsilon_0=\varepsilon_0(s\,,\delta)>0$*, two families of compact subintervals* $\mathcal{H}_s^{\pm}=\{H_{s,\varepsilon}^{\pm}\colon\ 0<\varepsilon\le\varepsilon_0\}$ *of* $(0\,,T)$*, and two families of functions* $\mathcal{Z}_s^{\pm}=\{z_{s,\varepsilon}^{\pm}\colon\ 0<\varepsilon\le\varepsilon_0\}$ *in* $AC([0\,,T])$ *such that, setting*

$$\begin{cases} J_{s,\varepsilon}^{+}=\left(s-\dfrac{\varepsilon}{\beta-z'(s)}\,,s+\dfrac{\varepsilon}{z'(s)-\alpha}\right),\\ J_{s,\varepsilon}^{-}=\left(s-\dfrac{\varepsilon}{z'(s)-\alpha}\,,s+\dfrac{\varepsilon}{\beta-z'(s)}\right), \end{cases}\qquad\qquad \varepsilon>0, \tag{3.1}$$

the following properties hold for every $0 < \varepsilon \leq \varepsilon_0$*:*

$$J^{\pm}_{s,\varepsilon/2} \subset H^{\pm}_{s,\varepsilon} \subset J^{\pm}_{s,2\varepsilon} \subset (0\,,T); \tag{3.2}$$

$$z^{\pm}_{s,\varepsilon} = z \text{ on } [0\,,T] \setminus \operatorname{int}\left(H^{\pm}_{s,\varepsilon}\right); \tag{3.3}$$

$$z(t) < z^{+}_{s,\varepsilon}(t) \leq z(t) + \delta \text{ for every } t \in \operatorname{int}\left(H^{+}_{s,\varepsilon}\right); \tag{3.4+}$$

$$z(t) - \delta \leq z^{-}_{s,\varepsilon}(t) < z(t) \text{ for every } t \in \operatorname{int}\left(H^{-}_{s,\varepsilon}\right); \tag{3.4−}$$

$$\varepsilon \geq z^{+}_{s,\varepsilon}(t) - [z(s) + z'(s)(t-s)] \geq \varepsilon/2 \text{ for every } t \in J^{+}_{s,\varepsilon/2}; \tag{3.5+}$$

$$-\varepsilon/2 \geq z^{-}_{s,\varepsilon}(t) - [z(s) + z'(s)(t-s)] \geq -\varepsilon \text{ for every } t \in J^{-}_{s,\varepsilon/2}; \tag{3.5−}$$

$$\left(z^{\pm}_{s,\varepsilon}\right)'(t) \in \{\alpha\,,\beta\} \text{ for a.e. } t \in H^{\pm}_{s,\varepsilon}. \tag{3.6}$$

Setting

$$E = \{s \in (0\,,T)\colon \ z \text{ is differentiable at } s \text{ with } \alpha < z'(s) < \beta\}\,,$$

note in particular that either family of compact sets $\mathcal{H}^{\pm} = \{H^{\pm}_{s,\varepsilon}\colon \ 0 < \varepsilon \leq \varepsilon_0(s\,,\delta), s \in E\}$ constitute a Vitali covering of the measurable set E itself.

Then we construct the local solutions that will be used in the proof of Theorem 2.2. As mentioned above, they will be defined as extremal solutions to suitable, convex-valued differential inclusions.

PROPOSITION 3.3. *Let* $a, b\colon [\eta_0 - \delta\,, \eta_0 + \delta] \to \mathbb{R}$ *be two bounded, upper and lower semicontinuous functions, respectively, such that*

(i) $a(\eta) < \xi_0 < b(\eta)$ *for every* $\eta \in [\eta_0 - \delta\,, \eta_0 + \delta]$

for some $\xi_0 \in \mathbb{R}$ *and assume that there exist* $y \in AC([0\,,T])$ *and* $t_0 \in (0\,,T)$ *such that*

(ii) y *is differentiable at* t_0 *with* $y(t_0) = \eta_0$ *and* $y'(t_0) = \xi_0$.

Then there exist $\varepsilon_0 = \varepsilon_0(t_0, \delta) > 0$*, two families of compact subintervals* $\mathcal{K}^{\pm}_{t_0} = \{K^{\pm}_{t_0,\varepsilon}\colon \ 0 < \varepsilon \leq \varepsilon_0\}$ *of* $(0\,,T)$ *such that*

$$\text{each set } K^{\pm}_{t_0,\varepsilon} \text{ is a neighborhood of } t_0 \text{ and each family } \mathcal{K}^{\pm}_{t_0} \text{ shrinks at } t_0, \tag{3.7}$$

and two families of functions $\mathcal{Y}^{\pm}_{t_0} = \{y^{\pm}_{t_0,\varepsilon}\colon \ 0 < \varepsilon \leq \varepsilon_0\}$ *in* $AC([0\,,T])$ *such that the following properties hold for every* $0 < \varepsilon \leq \varepsilon_0$*:*

$$y^{\pm}_{t_0,\varepsilon} = y \text{ on } [0\,,T] \setminus \operatorname{int}\left(K^{\pm}_{t_0,\varepsilon}\right); \tag{3.8}$$

$$y(t) < y^{+}_{t_0,\varepsilon}(t) \leq y(t) + \varepsilon \text{ for every } t \in \operatorname{int}\left(K^{+}_{t_0,\varepsilon}\right); \tag{3.9+}$$

$$y(t) - \varepsilon \leq y^{-}_{t_0,\varepsilon}(t) < y(t) \text{ for every } t \in \operatorname{int}\left(K^{-}_{t_0,\varepsilon}\right); \tag{3.9−}$$

$$\left|y^{\pm}_{t_0,\varepsilon}(t) - \eta_0\right| \leq \delta \text{ for every } t \in K^{\pm}_{t_0,\varepsilon}; \tag{3.10}$$

$$\left(y^{\pm}_{t_0,\varepsilon}\right)'(t) \in \left\{a\left(y^{\pm}_{t_0,\varepsilon}(t)\right)\,, b\left(y^{\pm}_{t_0,\varepsilon}(t)\right)\right\} \text{ for a.e. } t \in K^{\pm}_{t_0,\varepsilon}. \tag{3.11}$$

Proof. We are going to treat the + case, the other one being entirely equivalent. Therefore, to simplify the notations, we shall drop the superscript + from now on. Our strategy is the following: Relying on Lemma 3.2, for every small enough ε, we are going to define the compact set $K_{t_0,\varepsilon}$ and a sequence of functions $y_{k,t_0,\varepsilon}$ in $AC([0\,,T])$ satisfying (3.8), (3.9+), and (3.10) which are "approximated" solutions to the differential inclusion (3.11). The remarkable point is that although the derivatives of the "approximated" solutions $y_{k,t_0,\varepsilon}$ oscillate faster and faster as $k \to \infty$ in order to solve (3.11), the functions $y_{k,t_0,\varepsilon}$ can be defined in such a way that they do converge strongly in $AC([0\,,T])$. The limit function will be $y_{t_0,\varepsilon}$.

To this purpose, set $I = [\eta_0 - \delta\,, \eta_0 + \delta]$ to simplify the notation and define

$$\text{(3.12)} \qquad \begin{cases} M = \max\{a(\eta)\colon\ \eta \in I\}\,, \\ m = \min\{b(\eta)\colon\ \eta \in I\} \end{cases} \qquad \text{and} \qquad \sigma = \min\{m - \xi_0, \xi_0 - M\}$$

so that $\sigma > 0$ because of (i). Then choose $\varepsilon_0' = \varepsilon_0'(t_0, \delta\,, \sigma) > 0$ small enough so that

$$\text{(3.13)} \qquad \begin{cases} \varepsilon_0' < \min\{\delta/4\,, \sigma/16\,, \sigma t_0/4\,, \sigma(T - t_0)/4\}\,, \\ |t - t_0| \le 4\varepsilon_0'/\sigma \quad \Longrightarrow \quad |y(t) - \eta_0| < \min\{\delta/4\,, \sigma/32\} \end{cases}$$

hold and define also

$$\text{(3.14)} \qquad \begin{cases} a_k(\eta) = \max\{a(\eta') - k\,|\eta' - \eta|\colon\ \eta' \in I\} + \sigma/2k, \\ b_k(\eta), = \min\{b(\eta') + k\,|\eta' - \eta|\colon\ \eta' \in I\} - \sigma/2k, \end{cases} \qquad \eta \in I.$$

These functions a_k and b_k are (up to the constants $\sigma/2k$) the Moreau–Yosida approximations of a and b, respectively. They enjoy the following properties (see [10], for instance):

(3.15) a_k and b_k are Lipschitz continuous on I with Lipschitz constant k;

(3.16) $M + \sigma/2 \ge a_1(\eta)$ and $a_k(\eta) - a_{k+1}(\eta) \ge \Delta_k$ for every $\eta \in I$ and $k \ge 1$;

(3.17) $m - \sigma/2 \le b_1(\eta)$ and $b_{k+1}(\eta) - b_k(\eta) \ge \Delta_k$ for every $\eta \in I$ and $k \ge 1$;

where $\Delta_k = \frac{\sigma}{2k(k+1)}$, $k \ge 1$, and

(3.18) $a_k \to a$ and $b_k \to b$ pointwise on I.

Now, the preparatory work is over and the proof will be completed by proving the following three claims.

Claim 1. There exist $\varepsilon_0 = \varepsilon_0(t_0, \delta\,, \sigma) \in (0\,, \varepsilon_0']$ and a family of compact subintervals $\mathcal{K}_{t_0} = \{K_{t_0,\varepsilon}\colon\ 0 < \varepsilon \le \varepsilon_0\}$ of $(0\,, T)$ satisfying (3.7) with the further property that for every sequence of positive numbers $\{\omega_k\}_{k\ge1}$, the following holds for every $\varepsilon \in (0\,, \varepsilon_0]$: There exists a sequence of functions $\{y_{k,t_0,\varepsilon}\}_{k\ge0}$ in $AC([0\,, T])$ such that $y_{0,t_0,\varepsilon} = y$ and

(3.19) $y_{k,t_0,\varepsilon} = y$ on $[0\,, T] \setminus \text{int}\,(K_{t_0,\varepsilon})$;

(3.20) $0 < y_{k,t_0,\varepsilon}(t) - y_{k-1,t_0,\varepsilon}(t) < \min\left\{\dfrac{\varepsilon}{2^k}\,, \omega_k, \dfrac{\Delta_k}{4(k+1)}\right\}$, $t \in \text{int}\,(K_{t_0,\varepsilon})$;

(3.21) $|y_{k,t_0,\varepsilon}(t) - \eta_0| < \displaystyle\sum_{1\le h\le k} \delta/2^h$, $t \in K_{t_0,\varepsilon}$;

(3.22) $$y'_{k,t_0,\varepsilon}(t) \in \Big(a_{k+1}\left(y_{k,t_0,\varepsilon}(t)\right)\,, a_k\left(y_{k,t_0,\varepsilon}(t)\right)\Big) \cup \Big(b_k\left(y_{k,t_0,\varepsilon}(t)\right)\,, b_{k+1}\left(y_{k,t_0,\varepsilon}(t)\right)\Big) \text{ for a.e. } t \in K_{t_0,\varepsilon}$$

for every $k \ge 1$.

Claim 2. There exists a sequence of positive numbers $\{\omega_k\}_k$ such that for every $\varepsilon \in (0\,, \varepsilon_0]$, the sequence of functions $\{y_{k,t_0,\varepsilon}\}_k$ converges strongly in $AC([0\,, T])$ to a function $y_{t_0,\varepsilon} \in AC([0\,, T])$.

Claim 3. For the same sequence of numbers $\{\omega_k\}_k$, we also have that

$$\begin{cases} a_{k+h}\left(y_{k,t_0,\varepsilon}(t)\right) \to a\left(y_{t_0,\varepsilon}(t)\right), \\ b_{k+h}\left(y_{k,t_0,\varepsilon}(t)\right) \to b\left(y_{t_0,\varepsilon}(t)\right) \end{cases} \qquad \text{as } k \to \infty \tag{3.23}$$

for every $h \geq 0$ and $t \in [0\,,T]$.

Once the previous claims have been proved, we conclude immediately that $y_{t_0,\varepsilon}$ satisfies (3.8), (3.9+), and (3.10) for every $\varepsilon \in (0\,,\varepsilon_0]$ because of the corresponding properties of the functions $y_{k,t_0,\varepsilon}$. Moreover, Claim 2 implies that some subsequence of $\{y'_{k,t_0,\varepsilon}\}_k$ converges to $y'_{t_0,\varepsilon}$ almost everywhere on $[0\,,T]$ so that (3.11) follows from (3.22) and Claim 3.

Proof of Claim 1. We apply Lemma 3.2 in the + case with $s = t_0$, $z = y$ and $\alpha = [a_1(\eta_0) + a_2(\eta_0)]/2$ and $\beta = [b_1(\eta_0) + b_2(\eta_0)]/2$ thus finding $\varepsilon_0 = \varepsilon_0(t_0, \delta\,, \sigma) \in (0\,, \varepsilon'_0]$, a family of compact subintervals $\mathcal{H}_{t_0} = \{H_{t_0,\varepsilon}\colon\ 0 < \varepsilon \leq \varepsilon_0\}$ of $(0\,,T)$ which are all neighborhoods of t_0, and a family of functions $\mathcal{Z}_{t_0} = \{z_{t_0,\varepsilon}\colon\ 0 < \varepsilon \leq \varepsilon_0\}$ in $AC([0\,,T])$ such that (3.2), (3.3), (3.4+), and (3.6) hold. Relying on these properties, it is easy to check that $(z_{t_0,\varepsilon} - y) \to 0_+$ uniformly on $[0\,,T]$ as $\varepsilon \to 0$ so that for every $0 < \varepsilon \leq \varepsilon_0$, we can choose $0 < \varepsilon' \leq \varepsilon$ such that

$$0 < z_{t_0,\varepsilon'}(t) - y(t) < \min\left\{\frac{\varepsilon}{2}\,, \omega_1, \frac{\Delta_1}{4(1+1)}\right\}, \qquad t \in \text{int}\,\left(H_{t_0,\varepsilon'}\right).$$

Hence, recalling that $y_{0,t_0,\varepsilon} = y$ for every $\varepsilon \in (0\,,\varepsilon_0]$ by definition and setting $K_{t_0,\varepsilon} = H_{t_0,\varepsilon'}$ and $y_{1,t_0,\varepsilon} = z_{t_0,\varepsilon'}$ for every $\varepsilon \in (0\,,\varepsilon_0]$, we conclude that $K_{t_0,\varepsilon}$ satisfies (3.7), that $y_{1,t_0,\varepsilon}$ satisfies (3.19) and (3.20) with $k = 1$, and, moreover, that the derivative of $y_{1,t_0,\varepsilon}$ is such that

$$y'_{1,t_0,\varepsilon}(t) \in \left\{\frac{a_1(\eta_0) + a_2(\eta_0)}{2}\,, \frac{b_1(\eta_0) + b_2(\eta_0)}{2}\right\} \qquad \text{for a.e. } t \in K_{t_0,\varepsilon}.$$

As regards (3.21), note that each compact interval $K_{t_0,\varepsilon}$ is contained in H_{t_0,ε_0} and that this latter interval is contained in $[t_0 - 4\varepsilon'_0/\sigma\,, t_0 + 4\varepsilon'_0/\sigma]$ by (3.2). This latter inclusion, together with the equality $y_{0,t_0,\varepsilon} = y$ and the choice of ε'_0 made in (3.13), implies that

$$|y_{0,t_0,\varepsilon}(t) - \eta_0| \leq \min\left\{\frac{\delta}{4}\,, \frac{\Delta_1}{4(1+1)}\right\} \tag{3.24}$$

since $\sigma/32 = \Delta_1/8$. Hence,

$$\begin{aligned} |y_{1,t_0,\varepsilon}(t) - \eta_0| &\leq |y_{1,t_0,\varepsilon}(t) - y_{0,t_0,\varepsilon}(t)| + |y_{0,t_0,\varepsilon}(t) - \eta_0| \\ &< \varepsilon/2 + \delta/4 \leq \delta/4 + \delta/4 = \delta/2 \end{aligned} \tag{3.25}$$

holds for every $t \in K_{t_0,\varepsilon}$ because of (3.20) with $k = 1$ and (3.24). At last, to complete the first step, we have to check that $y_{1,t_0,\varepsilon}$ is an "approximated" solution to the differential inclusion (3.11) on the set $K_{t_0,\varepsilon}$, i.e., that (3.22) holds for $k = 1$. Indeed, let $y_{1,t_0,\varepsilon}$ be differentiable at some point $t \in \text{int}\,(K_{t_0,\varepsilon})$ with derivative $y'_{1,t_0,\varepsilon}(t) = [a_1(\eta_0) + a_2(\eta_0)]/2$. By elementary computations based on (3.15), (3.16), (3.20) for $k = 1$ and (3.24), we get

$$\begin{aligned} a_2\left(y_{1,t_0,\varepsilon}(t)\right) &\leq a_2(\eta_0) + 2\Big(|y_{1,t_0,\varepsilon}(t) - y_{0,t_0,\varepsilon}(t)| + |y_{0,t_0,\varepsilon}(t) - \eta_0|\Big) \\ &< a_2(\eta_0) + \frac{\Delta_1}{2} = \frac{a_1(\eta_0) + a_2(\eta_0)}{2} \end{aligned}$$

and, similarly,

$$a_1\left(y_{1,t_0,\varepsilon}(t)\right) > \frac{a_1(\eta_0)+a_2(\eta_0)}{2}.$$

Thus, $a_2\left(y_{1,t_0,\varepsilon}(t)\right) < y'_{1,t_0,\varepsilon}(t) < a_1\left(y_{1,t_0,\varepsilon}(t)\right)$, and the very same kind of computations in the case that $y'_{1,t_0,\varepsilon}(t)$ is $[b_1(\eta_0)+b_2(\eta_0)]/2$ yield that $b_1\left(y_{1,t_0,\varepsilon}(t)\right) < y'_{1,t_0,\varepsilon}(t) < b_2\left(y_{1,t_0,\varepsilon}(t)\right)$, which is (3.22) for $k=1$.

Next, we go on defining the second "approximated" solution $y_{2,t_0,\varepsilon}$ on the same set $K_{t_0,\varepsilon}$. We shall do this for a fixed $\varepsilon \in (0\,,\varepsilon_0]$.

To this purpose, choose any point $t \in \text{int}\,(K_{t_0,\varepsilon})$ where $y_{1,t_0,\varepsilon}$ is differentiable and (3.22) holds for $k=1$ and set $\eta_t = y_{1,t_0,\varepsilon}(t)$. For every such point t, we apply Lemma 3.2 in the + case again with $s=t$, $z=y_{1,t_0,\varepsilon}$ and $\alpha=[a_2(\eta_t)+a_3(\eta_t)]/2$ and $\beta=[b_2(\eta_t)+b_3(\eta_t)]/2$, thus finding a positive number $\theta_0=\theta_0(t\,,\varepsilon)$, a family of nondegenerate, compact intervals $\mathcal{L}_t=\{L_{t,\theta}\colon\ 0<\theta\le\theta_0\}$ contained in $K_{t_0,\varepsilon}$ and a family of functions $\mathcal{Z}_t=\{z_{t,\theta}\colon\ 0<\theta\le\theta_0\}$ in $AC([0\,,T])$ such that all sets $L_{t,\theta}$ in $\mathcal{L}_t$ are neighborhoods of t and the following properties hold for every $\theta\in(0\,,\theta_0]$:

$$z_{t,\theta} = y_{1,t_0,\varepsilon} \text{ on } [0\,,T]\setminus \text{int}\,(L_{t,\theta}); \tag{3.26}$$

$$0 < z_{t,\theta}(s) - y_{1,t_0,\varepsilon}(s) \le \min\left\{\frac{\varepsilon}{2^2}\,,\omega_2,\frac{\Delta_2}{4(2+1)}\right\} \text{ for every } s\in\text{int}\,(L_{t,\theta}); \tag{3.27}$$

$$z'_{t,\theta}(s) \in \left\{\frac{a_2(\eta_t)+a_3(\eta_t)}{2}\,,\frac{b_2(\eta_t)+b_3(\eta_t)}{2}\right\} \text{ for a.e. } s\in L_{t,\theta}. \tag{3.28}$$

Moreover, we can assume that θ_0 is small enough to have

$$|y_{1,t_0,\varepsilon}(s)-\eta_t| \le \frac{\Delta_2}{4(2+1)} \qquad \text{for every } s\in L_{t,\theta} \text{ and } \theta\in(0\,,\theta_0], \tag{3.29}$$

and we note also that from (3.27), (3.13), and (3.25), it follows that

$$|z_{t,\theta}(s)-\eta_0| \le |z_{t,\theta}(s)-y_{1,t_0,\varepsilon}(s)| + |y_{1,t_0,\varepsilon}(s)-\eta_0| \le \delta/2+\delta/4 \tag{3.30}$$

for every $s\in L_{t,\theta}$ and $\theta\in(0\,,\theta_0]$. Next, we prove that

$$z'_{t,\theta}(s) \in \Big(a_3\left(z_{t,\theta}(s)\right),a_2\left(z_{t,\theta}(s)\right)\Big) \cup \Big(b_2\left(z_{t,\theta}(s)\right),b_3\left(z_{t,\theta}(s)\right)\Big) \tag{3.31}$$

for a.e. $s\in L_{t,\theta}$. Indeed, (3.27), (3.16), and (3.29) yield

$$\begin{aligned} a_3\left(z_{t,\theta}(s)\right) &\le a_3(\eta_t) + 3\Big(|z_{t,\theta}(s)-y_{1,t_0,\varepsilon}(s)| + |y_{1,t_0,\varepsilon}(s)-\eta_t|\Big) \\ &< a_3(\eta_t)+\frac{\Delta_2}{2} \le \frac{a_2(\eta_t)+a_3(\eta_t)}{2} \end{aligned}$$

and, similarly,

$$\frac{a_2(\eta_t)+a_3(\eta_t)}{2} < a_2\left(z_{t,\theta}(s)\right)$$

for a.e. $s\in L_{t,\theta}$. Thus, (3.31) holds if the derivative of $z_{t,\theta}$ exists and is equal to $[a_2(\eta_t)+a_3(\eta_t)]/2$, and a specular argument yields (3.31) if $z'_{t,\theta}(s)$ is $[b_2(\eta_t)+b_3(\eta_t)]/2$.

So far, we have defined a family of functions $z_{t,\theta}$ satisfying (3.19),(3.20), (3.21), and (3.22) for $k=2$ around every "good" point t in the interior of $K_{t_0,\varepsilon}$ and, to complete the definition of $y_{2,t_0,\varepsilon}$, we just have to glue these functions $z_{t,\theta}$ by a covering

argument. Indeed, by the remark following Lemma 3.2, the family of nondegenerate, compact intervals $\mathcal{L} = \{L_{t,\theta}\colon\ 0 < \theta \le \theta_0(t\,,\varepsilon) \text{ and } t \in E_1\}$ constitutes a Vitali covering of the measurable set

$$E_1 = \left\{t \in \text{int }(K_{t_0,\varepsilon})\colon\ y'_{1,t_0,\varepsilon}(t) \text{ exists and (3.22) holds with } k=1\right\},$$

which is a full measure subset of int $(K_{t_0,\varepsilon})$. Hence, Vitali's covering theorem yields (at most) countably many points $t_j \in E_1$ and positive numbers $\theta_j \in (0, \theta_0(t_j, \varepsilon)]$ such that the corresponding compact intervals $L_j = L_{t_j,\theta_j}$ are pairwise disjoint sets that cover E_1, and hence $K_{t_0,\varepsilon}$ as well, up to a null set.

Now, we define the second "approximated" solution $y_{2,t_0,\varepsilon}$ by setting

$$y_{2,t_0,\varepsilon}(t) = y_{1,t_0,\varepsilon}(t) + \sum_{j\ge 1} [z_j(t) - y_{1,t_0,\varepsilon}(t)]\,, \qquad t \in [0\,,T], \tag{3.32}$$

where $z_j = z_{t_j,\theta_j}$. As the supports L_j of the functions $z_j - y_{1,t_0,\varepsilon}$ are disjoint, the series in (3.32) is actually a finite sum for every t and, moreover, the functions $z_j - y_{1,t_0,\varepsilon}$ have (essentially) uniformly bounded derivatives on $[0\,,T]$. Thus, $y_{2,t_0,\varepsilon}$ is Lipschitz continuous on $[0\,,T]$ by the Ascoli–Arzelà theorem, and the fulfillment of (3.19), (3.20), (3.21), and (3.22) for $k = 2$ follows straightforwardly from the corresponding properties (3.26), (3.27), (3.30), and (3.31) of the functions $z_j = z_{t_j,\theta_j}$.

Finally, all the remaining functions $y_{k,t_0,\varepsilon}$ are defined recursively in the very same way we have got $y_{2,t_0,\varepsilon}$ from $y_{1,t_0,\varepsilon}$, and this completes the proof of Claim 1.

Proof of Claim 2. We are going to choose the positive numbers ω_k so as to have strong convergence in $L^1([0\,,T])$ of the derivatives $y'_{k,t_0,\varepsilon}$.

To this purpose, let $\varphi \in \mathcal{D}(\mathbb{R})$ be the standard mollifying kernel and set, as usual, $\varphi_r(t) = r^{-1}\varphi(t/r)$ for every $t \in \mathbb{R}$ and $r > 0$. Choose also a sequence of positive numbers $0 < r_k < 2^{-k}$ in such a way that, extending each function $y_{k,t_0,\varepsilon}$ to the whole real line as a constant function off the interval $[0\,,T]$, the following inequality holds:

$$\int_{\mathbb{R}} \left|\varphi_{r_k} * y'_{k,t_0,\varepsilon}(t) - y'_{k,t_0,\varepsilon}(t)\right|\,dt \le \frac{1}{k}, \qquad k \ge 1. \tag{3.33}$$

Then let $\{\omega_k\}_k$ be the sequence defined by setting $\omega_1 = 1$ and, recursively,

$$\omega_{k+1} = r_k\omega_k, \qquad k \ge 1. \tag{3.34}$$

The reader might think that this way of choosing the numbers ω_k is inconsistent as it requires that the functions $y_{k,t_0,\varepsilon}$ be already defined. This is not the case. Indeed, we set ω_1 to be 1 and then define $y_{1,t_0,\varepsilon}$ so that (3.20) with $k = 1$ holds. Then we compute r_1 according to (3.33) with $k = 1$—which requires only that $y_{1,t_0,\varepsilon}$ be defined—and then we define the number $\omega_2 = r_1\omega_1$. Only then do we choose the function $y_{2,t_0,\varepsilon}$ so that (3.20) with $k = 2$ holds, and we restart the procedure.

Now, we claim that this choice of the numbers ω_k yields the conclusion of Claim 2, and we break the remaining part of the proof into the following three claims.

Claim 2.1. For every $\varepsilon \in (0\,,\varepsilon_0]$, the sequence $\{y_{k,t_0,\varepsilon}\}_k$ converges uniformly on $[0\,,T]$ to some function $y_{t_0,\varepsilon}$.

Indeed,

$$0 \le y_{k+1,t_0,\varepsilon}(t) - y_{k,t_0,\varepsilon}(t) \le \omega_{k+1}, \qquad t \in [0\,,T], \tag{3.35}$$

by (3.19) and (3.20). Since $0 < \omega_{k+1}/\omega_k = r_k \to 0$ as $k \to \infty$, the sequence $\{y_{k,t_0,\varepsilon}\}_k$ is uniformly Cauchy on $[0\,,T]$ and the conclusion follows.

Claim 2.2. $y_{t_0,\varepsilon} \in AC([0\,,T])$.

All the functions $y_{k,t_0,\varepsilon}$ have (essentially) uniformly bounded derivatives on the interval $K_{t_0,\varepsilon}$ since a and b are bounded and

$$a\left(y_{k,t_0,\varepsilon}(t)\right) < a_{k+1}\left(y_{k,t_0,\varepsilon}(t)\right) < y'_{k,t_0,\varepsilon}(t) < b_{k+1}\left(y_{k,t_0,\varepsilon}(t)\right) < b\left(y_{k,t_0,\varepsilon}(t)\right)$$

for a.e. $t \in K_{t_0,\varepsilon}$. Thus, $y_{t_0,\varepsilon}$ is Lipschitz continuous on $K_{t_0,\varepsilon}$. As it coincides with the absolutely continuous function y on $[0\,,T]\setminus \operatorname{int}\left(K_{t_0,\varepsilon}\right)$ because of (3.19), the claim is proved.

Claim 2.3. The sequence $\{y_{k,t_0,\varepsilon}\}_k$ converges strongly in $AC([0\,,T])$.

Indeed,

$$\begin{aligned}&\left\|y'_{k,t_0,\varepsilon} - y'_{t_0,\varepsilon}\right\|_1\\ &\le \left\|y'_{k,t_0,\varepsilon} - \varphi_{r_k} * y'_{k,t_0,\varepsilon}\right\|_1 + \left\|\varphi_{r_k} * y'_{k,t_0,\varepsilon} - \varphi_{r_k} * y'_{t_0,\varepsilon}\right\|_1 + \left\|\varphi_{r_k} * y'_{t_0,\varepsilon} - y'_{t_0,\varepsilon}\right\|_1.\end{aligned}$$

The first and the third summand on the right-hand side go to zero as $k \to \infty$ because of (3.33) and the properties of convolutions with mollifying kernels, respectively. We are thus left to prove that

$$R_k = \left\|\varphi_{r_k} * \left(y'_{k,t_0,\varepsilon} - y'_{t_0,\varepsilon}\right)\right\|_1 \to 0 \qquad \text{as } k \to \infty.$$

To this aim, note that $R_k = C r_k^{-1} \left\|y_{k,t_0,\varepsilon} - y_{t_0,\varepsilon}\right\|_\infty$, where $C = T\left\|\varphi'\right\|_1$, so that (3.34) and (3.35) yield that

$$\begin{aligned}\left\|y_{k,t_0,\varepsilon} - y_{t_0,\varepsilon}\right\|_\infty &\le \sum_{j\ge 1} \left\|y_{k+j,t_0,\varepsilon} - y_{k+j-1,t_0,\varepsilon}\right\|_\infty \le \sum_{j\ge 1} \omega_{k+j}\\ &= \omega_{k+1}\left(1 + \sum_{j\ge 1} r_{k+j}\right) \le 2\omega_{k+1} = 2 r_k \omega_k\end{aligned}$$

because $0 < r_k < 2^{-k}$ by assumption. Thus, $R_k \le 2C\omega_k \to 0$ as $k \to \infty$, and the conclusion follows.

Proof of Claim 3. Let t be in the interior of $K_{t_0,\varepsilon}$; otherwise the conclusion follows immediately from (3.19) and (3.18). For such t, we have

$$\begin{aligned}&\left|a_{k+h}\left(y_{k,t_0,\varepsilon}(t)\right) - a\left(y_{t_0,\varepsilon}(t)\right)\right|\\ &\le \left|a_{k+h}\left(y_{k,t_0,\varepsilon}(t)\right) - a_{k+h}\left(y_{t_0,\varepsilon}(t)\right)\right| + \left|a_{k+h}\left(y_{t_0,\varepsilon}(t)\right) - a\left(y_{t_0,\varepsilon}(t)\right)\right|,\end{aligned}$$

and the second summand at the right-hand side goes to zero as $k \to \infty$ because of (3.18) again, no matter what h is. As to the first one, (3.15) and the very same argument of Claim 2.3 show that it is bounded by $2(k+h)\omega_{k+1}$, and this goes to zero as $k \to \infty$ because (3.34) and the basic assumption $0 < r_k < 2^{-k}$ yield that $\omega_{k+1} \le 2^{-k}$. This proves the a case and nothing changes in the b case. □

4. Proof of the main result. In this final section, we put together and exploit the tools developed in the previous section and prove our attainment result, Theorem 2.2.

Proof of Theorem 2.2. Let R be a bounded, open rectangle whose closure is contained in $\mathcal{D}$. By (iii) and (iv) of Proposition 2.1, the function

$$q(\eta) = Ef^{**}(\eta\,,\xi), \qquad (\eta\,,\xi) \in \overline{R},$$

is well defined, and we claim that it has at most finitely many strict, local extrema in $\overline{R}$. Indeed, should this be false, there would be a converging sequence $\{m_k\}_k$ of strict, local extrema of q, say $m_k \to m_0$, and we could assume also the sequence $\{m_k\}_k$ is strictly monotone. Thus, q would fail to be monotone on both sides of m_0, and this gives a contradiction to (2.13). As $\mathcal{D}$ is a countable union of such rectangles R, we conclude that there exists an (at most) countable family of subsets of $\mathcal{D}$, say $\{m_i\} \times L_i$, with the property that, for every index i, L_i is a connected component of $\mathcal{D}_{m_i}$ and, for every $\xi \in L_i$, m_i is a strict, local extremum point of the mapping $\eta \in \mathcal{D}^\xi \to Ef^{**}(\eta\,,\xi)$. Conversely, if $(\eta\,,\xi) \in \mathcal{D}$ is such that η is a strict, local extremum point for $\eta \in \mathcal{D}^\xi \to Ef^{**}(\eta\,,\xi)$, then $\eta = m_i$ for some index i and the corresponding open interval L_i is the connected component of $\mathcal{D}_{m_i}$ containing ξ. We recall also that according to (v) of Proposition 2.1 and (2.14), a point m_i may be a strict, local minimum point for $\eta \in \mathcal{D}^\xi \to Ef^{**}(\eta\,,\xi)$ for some $\xi \in L_i$ only if $0 \notin L_i$.

Now, let $y \in AC([0\,,T])$ be a solution to $(\mathcal{P}^{**})$ and assume that $I^{**}(y) < \infty$; otherwise there is nothing else to prove. We are going to prove that y can be modified so as to find a new solution x to $(\mathcal{P}^{**})$ such that

$$f^{**}(x(t)\,,x'(t)) = f(x(t)\,,x'(t)) \qquad \text{for a.e. } t \in [0\,,T], \tag{4.1}$$

thus showing that x is a solution to $(\mathcal{P})$ as well. The proof goes through the following three steps.

Step 1. Let M be the subset of $\mathcal{D}$ defined by $M = \cup_i (\{m_i\} \times L_i)$ and note that $\mathcal{D} \setminus M$ is open. First, we prove that whenever the measurable set

$$E = \big\{t \in (0\,,T)\colon\ y \text{ is differentiable at } t \text{ and } \big(y(t)\,,y'(t)\big) \in \mathcal{D} \setminus M\big\} \tag{4.2}$$

has positive measure, we can use Lemma 3.2 and Proposition 3.3 to associate with every point $s \in E$ a family of new solutions $\mathcal{Y}_s = \{y_{s,\varepsilon}\colon 0 < \varepsilon \le \varepsilon_0(s)\}$ to $(\mathcal{P}^{**})$ such that the sets $K_{s,\varepsilon}$ defined as the closure of $\{y_{s,\varepsilon} \ne y\}$ are nondegenerate, compact intervals that shrink at s and the following properties hold for every $0 < \varepsilon \le \varepsilon_0(s)$:

$$\sup\{|y_{s,\varepsilon}(t) - y(t)|\colon\ 0 \le t \le T\} \le \varepsilon; \tag{4.3a}$$

$$f^{**}(y_{s,\varepsilon}(t)\,,y'_{s,\varepsilon}(t)) = f(y_{s,\varepsilon}(t)\,,y'_{s,\varepsilon}(t)) \text{ for a.e. } t \in K_{s,\varepsilon}. \tag{4.3b}$$

Step 2. Then we use the modified solutions of the previous step and a covering argument to define a new solution x to $(\mathcal{P}^{**})$ such that, setting

$$A = \big\{t \in (0\,,T)\colon\ x \text{ is differentiable at } t \text{ and } \big(x(t)\,,x'(t)\big) \in M\big\}\,, \tag{4.4}$$

we have that

$$f^{**}(x(t)\,,x'(t)) = f(x(t)\,,x'(t)) \qquad \text{for a.e. } t \in [0\,,T] \setminus A. \tag{4.5}$$

Step 3. Finally, we show that A is negligible. Thus, (4.5) reduces to (4.1), and this shows that x is a solution to $(\mathcal{P})$.

Proof of Step 1. Assume that the set E defined by (4.2) has positive measure, fix a point $t_0 \in E$, and set $\eta_0 = y(t_0)$ and $\xi_0 = y'(t_0)$. As $(\eta_0, \xi_0) \in \mathcal{D} \setminus M$ by assumption, the basic hypotheses (2.13) and (2.14) and the very definition of the set M itself imply that the restriction $\eta \in \mathcal{D}^{\xi_0} \to Ef^{**}(\eta\,,\xi_0)$ is monotone on the interval $[\eta_0 - \delta\,,\eta_0 + \delta]$ for some $\delta > 0$. Moreover, by possibly choosing a smaller value of δ, we can describe the upper and the lower parts of the boundary of the connected component of the

vertical strip $\mathcal{D} \cap ([\eta_0 - \delta\,, \eta_0 + \delta] \times \mathbb{R})$ of $\mathcal{D}$ containing (η_0, ξ_0) as in Proposition 3.1, i.e., as the graphs of two functions $a, b\colon [\eta_0 - \delta\,, \eta_0 + \delta] \to \mathbb{R}$ satisfying (i), (ii), (iii), and (iv) of Proposition 3.1. Also, setting $\mathcal{D}' = \{(\eta\,, \xi)\colon\ |\eta - \eta_0| \le \delta \text{ and } a(\eta) \le \xi \le b(\eta)\}$ and recalling (2.5), (2.6), and Proposition 2.1, we can write the convex envelope f^{**} of f on the set $\mathcal{D}'$ as

$$f^{**}(\eta\,, \xi) = d(\eta)\xi + q(\eta), \qquad (\eta\,, \xi) \in \mathcal{D}', \tag{4.6}$$

where the continuous functions $d, q\colon [\eta_0 - \delta\,, \eta_0 + \delta] \to \mathbb{R}$ are defined by $q(\eta) = Ef^{**}(\eta\,, \xi)$ and $d(\eta) = d(\eta\,, \xi)$ for every $(\eta\,, \xi) \in \mathcal{D}'$. Note also that q is monotone on the interval $[\eta_0 - \delta\,, \eta_0 + \delta]$ because of the corresponding property of the restriction $\eta \in \mathcal{D}^{\xi_0} \to Ef^{**}(\eta\,, \xi_0)$, that

$$f^{**}(\eta\,, \xi) \ge d(\eta)\xi + q(\eta) \qquad \text{for every } \eta \in [\eta_0 - \delta\,, \eta_0 + \delta] \text{ and } \xi \in \mathbb{R} \tag{4.7}$$

holds because of (2.1) and (2.7), and that the equalities

$$\begin{cases} f^{**}(\eta\,, a(\eta)) = f(\eta\,, a(\eta)), \\ f^{**}(\eta\,, b(\eta)) = f(\eta\,, b(\eta)), \end{cases} \qquad \eta \in [\eta_0 - \delta\,, \eta_0 + \delta], \tag{4.8}$$

follow from (iv) of Proposition 3.1.

Then we apply Proposition 3.3 and let $\mathcal{K}^{\pm}_{t_0} = \{K^{\pm}_{t_0,\varepsilon}\colon 0 < \varepsilon \le \varepsilon_0(t_0)\}$ and $\mathcal{Y}^{\pm}_{t_0} = \{y^{\pm}_{t_0,\varepsilon}\colon 0 < \varepsilon \le \varepsilon_0(t_0)\}$ be the corresponding intervals and functions. We assume in addition that $\varepsilon_0(t_0)$ is small enough so as to have

$$|y(t) - \eta_0| \le \delta \qquad \text{for every } t \in K^{\pm}_{t_0,\varepsilon} \text{ and } 0 < \varepsilon \le \varepsilon_0(t_0). \tag{4.9}$$

Now, we wish to compare $I^{**}(y^{\pm}_{t_0,\varepsilon})$ with $I^{**}(y)$. To this aim, recalling (3.8), we see that it is enough to compare

$$\int_{K^{\pm}_{t_0,\varepsilon}} f^{**}\left(y^{\pm}_{t_0,\varepsilon}(t)\,, \left(y^{\pm}_{t_0,\varepsilon}\right)'(t)\right) dt \qquad \text{and} \qquad \int_{K^{\pm}_{t_0,\varepsilon}} f^{**}(y(t)\,, y'(t))\, dt.$$

As $(y^{\pm}_{t_0,\varepsilon}(t)\,, (y^{\pm}_{t_0,\varepsilon})'(t))$ can only stay on the upper and lower parts of the boundary of $\mathcal{D}'$ for a.e. $t \in K^{\pm}_{t_0,\varepsilon}$ by (3.10) and (3.11), equation (4.6) shows that the first integral turns in

$$\int_{K^{\pm}_{t_0,\varepsilon}} f^{**}\left(y^{\pm}_{t_0,\varepsilon}(t)\,, \left(y^{\pm}_{t_0,\varepsilon}\right)'(t)\right) dt = \int_{K^{\pm}_{t_0,\varepsilon}} \left[d\left(y^{\pm}_{t_0,\varepsilon}(t)\right)\left(y^{\pm}_{t_0,\varepsilon}\right)'(t) + q\left(y^{\pm}_{t_0,\varepsilon}(t)\right)\right] dt.$$

By the fundamental theorem of calculus, the integrals of $d\left(y^{\pm}_{t_0,\varepsilon}\right)\left(y^{\pm}_{t_0,\varepsilon}\right)'$ and $d(y)y'$ over the interval $K^{\pm}_{t_0,\varepsilon}$ are equal as they are both derivatives of absolutely continuous functions having the same values at the endpoints of the interval $K^{\pm}_{t_0,\varepsilon}$. Hence, the previous computation together with (4.9) and (4.7) yields that

$$\begin{aligned} &\int_{K^{\pm}_{t_0,\varepsilon}} f^{**}\left(y^{\pm}_{t_0,\varepsilon}(t)\,, \left(y^{\pm}_{t_0,\varepsilon}\right)'(t)\right) dt \\ &\le \int_{K^{\pm}_{t_0,\varepsilon}} f^{**}(y(t)\,, y'(t))\, dt + \int_{K^{\pm}_{t_0,\varepsilon}} \left[q\left(y^{\pm}_{t_0,\varepsilon}(t)\right) - q(y(t))\right] dt. \end{aligned} \tag{4.10}$$

Then recall that q is monotone on the interval $[\eta_0 - \delta\,, \eta_0 + \delta]$ and that (3.9+) and (3.9−) hold. Therefore, setting $y_{t_0,\varepsilon} = y^-_{t_0,\varepsilon}$ and $K_{t_0,\varepsilon} = K^-_{t_0,\varepsilon}$ if q is increasing, and $y_{t_0,\varepsilon} = y^+_{t_0,\varepsilon}$ and $K_{t_0,\varepsilon} = K^+_{t_0,\varepsilon}$ otherwise, we conclude that all functions $y_{t_0,\varepsilon}$ are solutions to $(\mathcal{P}^{**})$. Moreover, (4.3a) and (4.3b) follow immediately either from (3.9−) or from (3.9+) and from (3.11) and (4.8), respectively. This completes the proof of the step.

Proof of Step 2. We assume again that the set E defined by (4.2) has positive measure; otherwise the conclusion of the step trivially holds with $x = y$.

Then, for every point $s \in E$, we let $\mathcal{K}_s = \{K_{s,\varepsilon}\colon\ 0 < \varepsilon \le \varepsilon_0(s)\}$ be the family of nondegenerate, compact intervals and $\mathcal{Y}_s = \{y_{s,\varepsilon}\colon\ 0 < \varepsilon \le \varepsilon_0(s)\}$ be the family of new solutions to the relaxed problem $(\mathcal{P}^{**})$ satisfying (4.3) that were constructed in Step 1. Moreover, we can assume that $0 < \varepsilon_0(s) \le 1$ for every $s \in E$.

Now, we are left to prove that we can select and glue some of these functions $y_{s,\varepsilon}$ from $\mathcal{Y}_s$ so as to find a new solution x to $(\mathcal{P}^{**})$ satisfying (4.5).

To this purpose, recall that the intervals $\mathcal{K} = \{K_{s,\varepsilon}\colon\ 0 < \varepsilon \le \varepsilon_0(s) \text{ and } s \in E\}$ defined in the previous step constitute a Vitali covering of E because of (3.7). Hence, Vitali's covering theorem yields (at most) countably many points $s_h \in E$ and numbers $\varepsilon_h \in (0\,, \varepsilon_0(s_h)]$ such that the corresponding intervals $K_h = K_{s_h,\varepsilon_h}$ are pairwise disjoint subsets of $(0\,, T)$ which cover E up to a null set. Also let $y_h = y_{s_h,\varepsilon_h}$ be the corresponding solution to $(\mathcal{P}^{**})$ so that the equality

$$\int_{K_h} f^{**}\left(y_h(t)\,, y_h'(t)\right)\, dt = \int_{K_h} f^{**}\left(y(t)\,, y'(t)\right)\, dt \qquad \text{for every } h \tag{4.11}$$

follows from (3.8) and, moreover,

$$f^{**}\left(y_h(t)\,, y_h'(t)\right) = f\left(y_h(t)\,, y_h'(t)\right) \qquad \text{for a.e. } t \in K_h \tag{4.12}$$

by (4.3b), i.e., the vectors $(y_h(t)\,, y_h'(t))$ keep off the set $\mathcal{D}$ for a.e. $t \in K_h$.

Then we set

$$x(t) = y(t) + \sum_h \left[y_h(t) - y(t)\right], \qquad t \in [0\,, T],$$

and, as in Claim 1 of Proposition 3.3, we show that the series above converges strongly in $AC([0\,, T])$. Indeed, the functions $y_h - y$ are absolutely continuous functions on $[0\,, T]$, whose supports K_h are pairwise disjoint so that the series above actually reduces to a finite sum for every t and its partial sums are bounded by 1 by either (3.9+) or (3.9−) and the choice of $\varepsilon_0(s)$. Thus, the series converges strongly in $L^1([0\,, T])$ by Lebesgue's dominated convergence theorem. As to the derivatives, first recall that the basic assumptions (H2) and (H3) imply that (2.10) holds, i.e., that $f^{**}(\eta\,, \xi)$ has at least linear growth at infinity as $|\xi| \to \infty$, uniformly with respect to η ranging in a bounded interval. Therefore, setting $R = \|y\|_\infty + 1$, for instance, and letting $\alpha > 0$ and $\beta \ge 0$ be the corresponding numbers as in (2.10), we get from (4.11) that

$$\sum_h \int_{K_h} |y_h'(t)|\, dt \le \frac{1}{\alpha}\left[I^{**}(y) + \beta T\right].$$

Hence,

$$\sum_h \int_0^T |y_h'(t) - y'(t)|\, dt = \sum_h \int_{K_h} |y_h'(t) - y'(t)|\, dt$$
$$\leq \sum_h \int_{K_h} |y_h'(t)|\, dt + \|y'\|_1 \leq \frac{1}{\alpha}\,[I^{**}(y) + \beta T] + \|y'\|_1 < \infty;$$

i.e., the series of the derivatives converges strongly in $L^1([0\,,T])$, and this proves the claim about the series defining x.

Finally, it is plain that x is feasible for $(\mathcal{P}^{**})$ because of (3.8) so that, adding (4.11) up for every h, we conclude that x is a solution to $(\mathcal{P}^{**})$. Moreover, $x = y$ on $[0\,,T]\setminus(\cup_h K_h)$, whereas $x = y_h$ on K_h and $x' = y_h'$ almost everywhere on the same set so that the equality $f^{**}(x\,,x') = f(x\,,x')$ almost everywhere on $\cup_h K_h$ follows from (4.12). As the intervals K_h cover E up to a null set, we conclude that (4.5) holds.

Proof of Step 3. Let x be the solution to $(\mathcal{P}^{**})$ satisfying (4.5) that was defined in Step 2 and let A be the set defined by (4.4). We have to show that A is negligible, which we will accomplish by showing that, otherwise, a feasible function $\overline{x}$ such that $I^{**}(\overline{x}) < I^{**}(x)$ would exist.

Indeed, recalling the definitions of the sets A and M in Steps 2 and 1, respectively, we see that set A itself can actually be written, up to a null set, as a countable union of sets

$$B_i = \{t \in (0\,,T)\colon\ x(t) = m_i,\ x \text{ is differentiable at } t \text{ and } x'(t) = 0\}$$

since x' vanishes almost everywhere on each level set $\{x = m_i\}$. Now, assume by contradiction that some set B_i has positive measure and, to simplify notation, set $m = m_i$, $L = L_i$, and $B = B_i$. Note also that (v) of Proposition 2.1 and our assumption (2.14) imply that $0 \in L$ and that m has to be a strict, local maximum point of $\eta \in \mathcal{D}^0 \to Ef^{**}(\eta\,,0)$.

Then, for a sufficiently small $\delta > 0$, there exist two functions $a, b\colon [m - 2\delta\,,m + 2\delta] \to \mathbb{R}$ as in Proposition 3.1 such that the functions $d, q\colon [m - 2\delta\,,m + 2\delta] \to \mathbb{R}$ defined by

$$\begin{cases} d(\eta) = d(\eta\,,\xi), \\ q(\eta) = Ef^{**}(\eta\,,\xi), \end{cases} \qquad \eta \in [m - 2\delta\,,m + 2\delta],$$

are well defined because of Proposition 2.1, no matter what $\xi \in (a(\eta), b(\eta))$ is, and, moreover, the following properties hold:

(4.13a) $(m\,,0) \in \mathcal{D}$;

(4.13b) $a(\eta) < 0 < b(\eta)$ for every $\eta \in [m - 2\delta\,,m + 2\delta]$;

(4.13c) $q(\eta) = f^{**}(\eta\,,0)$ for every $\eta \in [m - 2\delta\,,m + 2\delta]$;

(4.13d) $f^{**}(\eta\,,\xi) = d(\eta)\xi + q(\eta)$ for $\xi \in (a(\eta)\,,b(\eta))$ and $\eta \in [m - 2\delta\,,m + 2\delta]$;

(4.13e) m is a strict, local maximum point of q;

(4.13f) q is increasing on $[m - 2\delta\,,m]$ and decreasing on $[m\,,m + 2\delta]$.

Now, let $s \in B$ be a density point of B and let $\mathcal{H}_s^{\pm} = \left\{H_{s,\varepsilon}^{\pm}\colon\ 0 < \varepsilon \leq \varepsilon_0\right\}$ and $\mathcal{X}_s^{\pm} = \left\{x_{s,\varepsilon}^{\pm}\colon\ 0 < \varepsilon \leq \varepsilon_0\right\}$ be the families of intervals and functions associated with $z = x$, δ, $\alpha = \max\{a(\eta)\colon\ |\eta - m| \leq \delta\} < 0$, and $\beta = \min\{b(\eta)\colon\ |\eta - m| \leq \delta\} > 0$

by Lemma 3.2. Let also $J^{\pm}_{s,\varepsilon}$ be the intervals defined by (3.1) and assume $\varepsilon_0 = \varepsilon_0(s\,,\delta)$ is small enough so as to have $\varepsilon_0 \le 2\delta$ and $|x(t) - m| \le \delta$ for every t in $J^{\pm}_{s,2\varepsilon_0}$. Hence, $\left|x^{\pm}_{s,\varepsilon}(t) - m\right| \le 2\delta$ for every t in $H^{\pm}_{s,\varepsilon}$ and every $0 < \varepsilon \le \varepsilon_0$ by either (3.4+) or (3.4−). Each function $x^{\pm}_{s,\varepsilon}$ is feasible for $(\mathcal{P}^{**})$ because of (3.2) and (3.3) and, in order to compare $I^{**}\left(x^{\pm}_{s,\varepsilon}\right)$ with $I^{**}(x)$, it is enough that we compare

$$\int_{H^{\pm}_{s,\varepsilon}} f^{**}\left(x^{\pm}_{s,\varepsilon}(t)\,,\left(x^{\pm}_{s,\varepsilon}\right)'(t)\right)\,dt \qquad \text{and} \qquad \int_{H^{\pm}_{s,\varepsilon}} f^{**}\left(x(t)\,,x'(t)\right)\,dt.$$

Now, the very same computations of Step 1 yield that

$$\begin{aligned}&\int_{H^{\pm}_{s,\varepsilon}} f^{**}\left(x^{\pm}_{s,\varepsilon}(t)\,,\left(x^{\pm}_{s,\varepsilon}\right)'(t)\right)\,dt\\ &\le \int_{H^{\pm}_{s,\varepsilon}} f^{**}\left(x(t)\,,x'(t)\right)\,dt + \int_{H^{\pm}_{s,\varepsilon}} \left[q\left(x^{\pm}_{s,\varepsilon}(t)\right) - q(x(t))\right]\,dt\end{aligned} \tag{4.14}$$

for every $0 < \varepsilon \le \varepsilon_0$, and we claim that for small enough ε, we can choose either + or − so that the last summand at the right-hand side of (4.14) is negative, thus getting a contradiction.

To see this, choose a decreasing sequence $\{\varepsilon_k\}_k$ in $(0\,,\varepsilon_0]$ such that $\varepsilon_k \to 0$ and set

$$\eta_k = \frac{1}{\varepsilon_k}\sup\left\{|x(t) - m| :\ \ |t - s| < 2p\varepsilon_k\right\} \qquad \text{for every } k,$$

where $p = \max\left\{(\beta - x'(s))^{-1}\,,(x'(s) - \alpha)^{-1}\right\}$. Obviously, $\eta_k \to 0_+$ since x is differentiable at s with $x'(s) = 0$ by assumption and, moreover, $0 < \eta_k\varepsilon_k \le \delta$ by the choice of ε_0. Then, recalling that m is a strict, local maximum point of q and possibly extracting a subsequence that we still label as $\{\varepsilon_k\}_k$, we can assume that the minimum between $q(m - \eta_k\varepsilon_k)$ and $q(m + \eta_k\varepsilon_k)$ is actually achieved for every k by terms with the same sign inside, say $q(m + \eta_k\varepsilon_k)$, so that

$$0 < q(m) - q(m + \eta_k\varepsilon_k) = \max\left\{q(m) - q(m - \eta_k\varepsilon_k), q(m) - q(m + \eta_k\varepsilon_k)\right\} \tag{4.15}$$

holds for every k.

According to this assumption, we choose the + functions and, to simplify notation, we set $x_k = x^{+}_{s,\varepsilon_k}$ and $H_k = H^{+}_{s,\varepsilon_k}$ for every k. Of course, if the minimum between $q(m - \eta_k\varepsilon_k)$ and $q(m + \eta_k\varepsilon_k)$ was achieved by $q(m - \eta_k\varepsilon_k)$ instead, we would have chosen the − functions.

Finally, also set $J_\varepsilon = J^{+}_{s,\varepsilon}$ for $\varepsilon > 0$ and note that (3.2) reduces to

$$J_{\varepsilon_k/2} \subset H_k \subset J_{2\varepsilon_k}. \tag{4.16}$$

We prove the claim by showing that the integral

$$\int_{H_k} \left[q(x_k(t)) - q(x(t))\right]\,dt$$

is eventually negative.

To see this, set

$$A^1_k = \frac{1}{|H_k|}\int_{H_k} \left[q(m) - q(x_k(t))\right]\,dt \qquad \text{and} \qquad A^2_k = \frac{1}{|H_k|}\int_{H_k} \left[q(m) - q(x(t))\right]\,dt$$

for every k so that the claim reduces to proving that eventually $\left|A_k^1 - A_k^2\right| > 0$. Indeed, recalling (4.16), that q is decreasing on the interval $[m\,, m+2\delta]$ by (4.13f) and noting that (3.5+) reduces to

$$2\delta \geq \varepsilon_k \geq x_k(t) - m \geq \varepsilon_k/2, \qquad t \in J_{\varepsilon_k/2},$$

because $s \in B$ and because of the choice of ε_0, we find that

$$\begin{aligned} A_k^1 &\geq \frac{1}{|J_{2\varepsilon_k}|} \int_{J_{\varepsilon_k/2}} [q(m) - q(x_k(t))]\, dt \\ &\geq \frac{1}{|J_{2\varepsilon_k}|} \int_{J_{\varepsilon_k/2}} [q(m) - q(m+\varepsilon_k/2)]\, dt = \frac{1}{4}\,[q(m) - q(m+\varepsilon_k/2)] \end{aligned}$$

for every k since $\left|J_{\varepsilon_k/2}\right| / |J_{2\varepsilon_k}| = 1/4$ by (3.1). As to A_k^2, note that

$$A_k^2 = \frac{1}{|H_k|} \int_{H_k \setminus B} [q(m) - q(x(t))]\, dt \qquad \text{for every } k$$

and that $m - \eta_k\varepsilon_k \leq x(t) \leq m + \eta_k\varepsilon_k$ for $t \in H_k$ by (4.16) and the very definition of η_k. Hence,

$$\begin{aligned} 0 &\leq q(m) - q(x(t)) \leq \max\left\{q(m) - q(m-\eta_k\varepsilon_k)\,, q(m) - q(m+\eta_k\varepsilon_k)\right\} \\ &= q(m) - q(m+\eta_k\varepsilon_k) \end{aligned}$$

for every $t \in H_k$ and every k by (4.13f) and (4.15) whence we obtain

$$0 \leq A_k^2 \leq \frac{|H_k \setminus B|}{H_k}\,[q(m) - q(m+\eta_k\varepsilon_k)] \qquad \text{for every } k.$$

Since $\eta_k \to 0$, it follows that eventually $q(m) - q(m+\varepsilon_k/2) \geq q(m) - q(m+\eta_k\varepsilon_k) > 0$ by (4.13f). As s is a density point of B and the intervals $\{H_k\}_k$ shrink at s, the ratio $|H_k \setminus B| / |H_k|$goes to zero because of (2.8) and the conclusion follows. ∎

REFERENCES

[1] G. Aubert and R. Tahraoui, *Théorèmes d'existence pour des problèmes du Calcul des Variations du type:* $\inf \int_0^L f(x,u'(x))dx$ *et* $\inf \int_0^L f(x,u(x),u'(x))dx$, J. Differential Equations, 33 (1979), pp. 1–15.

[2] P. Celada and S. Perrotta, *Minimizing nonconvex, simple integrals of product type*, J. Differential Equations, 171 (2001), pp. 148–172.

[3] A. Cellina, *The classical problem of the Calculus of Variations in the autonomous case: Relaxation and Lipschitzianity of solutions*, preprint, Universitá degli Studi di Milano Bicocca, Italy, 2001.

[4] A. Cellina and G. Colombo, *On a classical problem of the Calculus of Variations without convexity assumptions*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 97–106.

[5] A. Cellina and C. Mariconda, *The existence question in the Calculus of Variations: A density result*, Proc. Amer. Math. Soc., 120 (1994), pp. 1145–1150.

[6] A. Cellina, G. Treu, and S. Zagatti, *On the minimum problem for a class of non coercive functionals*, J. Differential Equations, 127 (1996), pp. 225–262.

[7] F. H. Clarke, *An indirect method in the Calculus of Variations*, Trans. Amer. Math. Soc., 336 (1993), pp. 655–673.

[8] G. Crasta, *An existence result for noncoercive, nonconvex problems in the Calculus of Variations*, Nonlinear Anal., 26 (1996), pp. 1527–1533.

[9] G. Crasta and A. Malusa, *Existence results for noncoercive variational problems*, SIAM J. Control Optim., 34 (1996), pp. 2064–2076.

[10] G. Dal Maso, *An Introduction to* Γ*-Convergence*, Progr. Nonlinear Differential Equations Appl. 8, Birkhäuser, Boston, 1993.
[11] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, Stud. Math. Appl. 1, North Holland, Amsterdam, 1976.
[12] N. Fusco, P. Marcellini, and A. Ornelas, *Existence of minimizers for some non convex one dimensional integrals*, Portugal. Math., 55 (1998), pp. 167–185.
[13] D. Giachetti and R. Schianchi, *Minima of some nonconvex, noncoercive problems*, Ann. Mat. Pura Appl. (4), 165 (1993), pp. 109–120.
[14] P. Marcellini, *Non convex integrals of the Calculus of Variations*, in Methods of Non Convex Analysis, Lecture Notes in Math. 1446, A. Cellina, ed., Springer, Berlin, 1990, pp. 16–57.
[15] C. Mariconda, *A generalization of the Cellina-Colombo theorem for a class of nonconvex variational problems*, J. Math. Anal. Appl., 175 (1993), pp. 514–522.
[16] M. D. P. Monteiro Marques and A. Ornelas, *Genericity and existence of a minimum for scalar integrals functionals*, J. Optim. Theory Appl., 86 (1995), pp. 421–431.
[17] A. Ornelas, *Existence of scalar minimizers for nonconvex simple integrals of sum type*, J. Math. Anal. Appl., 221 (1998), pp. 559–573.
[18] J. P. Raymond, *Champs hamiltoniens, relaxation et existence de solutions en Calcul des Variations*, J. Differential Equations, 70 (1987), pp. 226–274.
[19] J. P. Raymond, *Conditions nécessaires et suffisantes d'existence des solutions en Calcul des Variations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 4 (1987), pp. 169–202.
[20] J. P. Raymond, *Existence and uniqueness results for minimization problems with nonconvex functionals*, J. Optim. Theory Appl., 82 (1994), pp. 571–591.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 4, pp. 1141–1165

FEEDBACKS FOR NONAUTONOMOUS REGULAR LINEAR SYSTEMS*

ROLAND SCHNAUBELT†

Abstract. We introduce nonautonomous well-posed and (absolutely) regular linear systems as quadruples consisting of an evolution family and output, input, and input–output maps subject to natural hypotheses. In the spirit of Weiss' work, these maps are represented in terms of admissible observation and control operators (the latter in an approximate sense) in the time domain. In this setting, the closed-loop system exists for a canonical class of "admissible" feedbacks, and it inherits the absolute regularity and other properties of the given system. In particular, we can iterate feedbacks.

Key words. input–output map, evolution family, Lebesgue extension, representation, closed-loop system, controllable, observable, robustness of exponential dichotomy, input–output stability

AMS subject classifications. Primary, 93C25; Secondary, 47D06

PII. S036301290139169X

1. Introduction.

As a motivation, we first look at the finite dimensional nonautonomous linear system

$$\begin{aligned} x'(t) &= A(t)x(t) + B(t)u(t), \quad t \geq s \geq 0, \\ y(t) &= C(t)x(t), \quad t \geq s \geq 0, \quad x(s) = x_0, \end{aligned} \tag{1.1}$$

on the state space X with control operators $B(t) : U \to X$, observation operators $C(t) : X \to Y$, the control space U, and the observation space Y. Let $T(t,s)$, $t \geq s \geq 0$, be the evolution family (propagator) on X generated by $A(\cdot)$. Then the output of (1.1) with $u = 0$, the state of (1.1) with $x_0 = 0$, and the input–output operator of (1.1) are given by

$$\begin{aligned} (\Psi_s x_0)(t) &= C(t)T(t,s)x_0, \qquad \Phi_{t,s}u = \int_s^t T(t,\tau)B(\tau)u(\tau)d\tau, \\ (\mathbb{F}_s u)(t) &= C(t)\int_s^t T(t,\tau)B(\tau)u(\tau)d\tau, \qquad t \geq s. \end{aligned} \tag{1.2}$$

If one feeds back the output via $u(t) = \Delta(t)y(t)$, the resulting closed-loop system is described by the perturbed evolution equation

$$x'(t) = [A(t) + B(t)\Delta(t)C(t)]\, x(t), \quad t \geq s \geq 0, \qquad x(s) = x_0. \tag{1.3}$$

Of course, $x(t) = T_\Delta(t,s)x_0$ solves (1.3) if T_Δ is generated by $A(t) + B(t)\Delta(t)C(t)$. This evolution family also satisfies the "variation of constants formulas"

$$T_\Delta(t,s)x = T(t,s)x + \int_s^t T(t,\tau)B(\tau)\Delta(\tau)C(\tau)T_\Delta(\tau,s)x\, d\tau, \tag{1.4}$$

$$T_\Delta(t,s)x = T(t,s)x + \int_s^t T_\Delta(t,\tau)B(\tau)\Delta(\tau)C(\tau)T(\tau,s)x\, d\tau \tag{1.5}$$

*Received by the editors July 2, 2001; accepted for publication (in revised form) April 11, 2002; published electronically October 29, 2002.
http://www.siam.org/journals/sicon/41-4/39169.html
†FB Mathematik and Informatik, Universität Halle, 06099 Halle (Saale), Germany (schnaubelt@mathematik.uni-halle.de).

for $t \geq s$ and $x \in X$. Identity (1.4) is the integrated version of (1.3). To derive (1.5), we perturb T_Δ by $-B(t)\Delta(t)C(t)$. There are formulas analogous to (1.4) and (1.5) relating the maps from (1.2) with the corresponding ones of the closed-loop system. These formulas are needed to show further properties of the closed-loop system. For instance, the closed-loop system is observable (controllable) if and only if the open-loop system is observable (controllable). In this framework, one can also show the equivalence of internal stability with input/output stability, detectability, and stabilizability. We establish infinite dimensional versions of these results in section 5.

If we pass to an infinite dimensional state space X, it is no longer clear that (1.3) possesses differentiable solutions for "many" initial values even if the Cauchy problem for $A(\cdot)$ is well-posed; cf. [7], [9, section VI.9], [10]. Nevertheless, the formulas (1.2) still work, and there is an evolution family T_Δ fulfilling (1.4) and (1.5). Thus $x(t) = T_\Delta(t,s)x_0$ is the "mild" solution of (1.3) [7]. However, point or boundary control and observation lead to input and output operators $B(t) : U \to \overline{X}_t$ and $C(t) : \underline{X}_t \to Y$ for spaces $\underline{X}_t \subsetneqq X \subsetneqq \overline{X}_t$, where $C(t)$ usually is not closable; see, e.g., [3], [16]. In order to solve (1.4) in this more general setting, we may restrict ourselves to "admissible" observation and control operators—roughly speaking, those for which the expressions (1.2) make sense. Then we are also faced with the question of whether the operators $B(t)$ and $C(t)$ are again admissible for the perturbed evolution family T_Δ, which is necessary to verify (1.5) or to iterate feedbacks.

The resulting perturbation problem (1.3) generalizes the settings of both the Desch–Schappacher theorem (where $\Delta(t) = C(t) = I$) and the Miyadera theorem (where $\Delta(t) = B(t) = I$) from semigroup theory [9, section III.3], [19]. In the control literature, there is a rich perturbation theory for the autonomous case (i.e., $A(t) = A$, $B(t) = B$, $C(t) = C$, $\Delta(t) = \Delta$). Linear systems belonging to the *Pritchard–Salamon class* [18] were exhaustively treated in [6]. Salamon and Weiss introduced the larger class of *well-posed linear systems* in [21] and [28], [29], [30], [31]. Here the semigroup T is given, and the operators Φ, Ψ, and $\mathbb{F}$ are defined in an abstract way by certain algebraic relations. One can then construct admissible control and observation operators B and C and obtain formulas such as (1.2) if the system satisfies a quite natural *regularity* hypothesis. Weiss established a powerful feedback theory for regular systems in the Hilbert space situation [32]. We refer to section 4, [3, section 3.3], [17], [33], and, in particular, to Staffans' monograph [25] for further information and literature.

For nonautonomous systems in variational form, there is the well-known approach due to Lions [16]; see also [1] and [3, Chap. 2]. In a general setting, Hinrichsen, Jacob, and Pritchard [10], [12], [14] constructed an evolution family solving (1.4) for initial values x contained in a dense subspace $\underline{X}$ of X under rather weak assumptions covering autonomous regular systems. However, (1.5) and the admissibility of the perturbed system was investigated only in [12] requiring stronger hypotheses of Pritchard–Salamon type.

In the present work, we combine the direct approach of Hinrichsen, Jacob, and Pritchard with some of Weiss' ideas: In Definition 2.6, we introduce "Lebesgue extensions" of given observation operators $C(t)$ (cf. [28]) which allow the study of (1.4) and (1.5) for all $x \in X$ and simplify several technical details of the proofs considerably. For similar reasons, we mostly work with *nonautonomous (absolutely) regular systems*, which are defined in the spirit of Weiss' work (see Definitions 3.6 and 3.10) as opposed to *admissible systems*, which have been used in [10], [12], [14] and are given directly by operators $B(t)$ and $C(t)$ (see Definition 3.8). In Theorem 2.7, Proposition 3.5, and Theorem 3.11, we represent a given regular system similar as in (1.2). It is known

[27, Ex. 6] that (1.3) can only be solved if the feedback is not "too large." We thus introduce *admissible* feedbacks in Definition 4.1; cf. [25, section 7.1], [32, section 3]. In our main theorem, Theorem 4.4, we then establish the existence of an absolutely regular closed-loop system for a given absolutely regular nonautonomous system with admissible time varying feedback.

However, the extension of Weiss' theory to the nonautonomous case is limited by two serious obstacles: One cannot apply transform methods, and, in contrast to semigroups (see, e.g., [2, Chap. V], [9, section II.5]), we do not have a general extrapolation theory for evolution families. The first point excludes the use of transfer functions (being crucial in [32]) but leads us to arguments which work in a Banach space setting (as in [25, Chap. 7]). The second point forces us to employ approximation formulas for the representation of control systems in Proposition 3.5. A similar problem occurs in the computation of the feedback system and in the context of (1.5); cf. Remark 4.7.

In section 5 we derive analogues of (1.4) and (1.5) for the operators given in (1.2). It is also seen that the closed-loop system is controllable (or observable) if and only if the given system is controllable (or observable). Moreover, iterated feedbacks behave as one would expect. We further prove that the feedback system inherits the exponential dichotomy (or stability) of T. Results of this type are important tools in investigating the long-term behavior of evolution equations but have not yet been obtained for perturbations mapping from a subspace of X to a larger space. Finally, the equivalence of internal stability with input–output stability, detectability, and stabilizability is established, extending theorems from [5], [6], [17], [20], [33] to the present setting. As a sample of possible applications, we treat in section 6 a parabolic problem with point observation and control in space dimension $n \leq 3$ which can be generalized in various directions.

Notation. We denote the space of bounded linear operators from X to Y by $\mathcal{L}(X,Y)$ and put $\mathcal{L}(X) := \mathcal{L}(X,X)$, where X, Y, U, Z always designate Banach spaces. $C_b(\mathbb{R}_+, \mathcal{L}_s(X,Y))$ and $L^\infty(\mathbb{R}_+, \mathcal{L}_s(X,Y))$ are the spaces of (essentially) bounded strongly continuous and strongly measurable operator-valued functions, respectively. We set $a \vee b = \max\{a,b\}$, $a \wedge b = \min\{a,b\}$, $a^+ = a \vee 0$, and $a^- = (-a)^+$ for $a, b \in \mathbb{R}$ and write $\mathbb{1}_N$ for the characteristic function of $N \subset M$. Unless otherwise stated, p is a number contained in $[1,\infty)$. The spaces $L^p_{loc}([s,\infty), Z)$ and $C([s,\infty), Z)$ are endowed with their standard Fréchet topologies. We mostly use the same symbol for a function on $J \subset \mathbb{R}$ and its restrictions to subintervals.

2. Nonautonomous observation systems.

DEFINITION 2.1. *A set* $T = (T(t,s))_{t \geq s \geq 0} \subseteq \mathcal{L}(X)$ *is an* evolution family *if*

(E1) $T(t,s) = T(t,r)T(r,s)$, $T(s,s) = I$,

(E2) $(t,s) \mapsto T(t,s)$ *is strongly continuous, and*

(E3) $\|T(t,s)\| \leq Me^{w(t-s)}$

for $t \geq r \geq s \geq 0$ *and constants* $M \geq 1$ *and* $w \in \mathbb{R}$. *We also define* $(\mathbb{K}_s f)(t) = \int_s^t T(t,\tau) f(\tau)\, d\tau$ *for* $t \geq s \geq 0$ *and* $f \in L^1_{loc}([s,\infty), X)$ *and put* $\mathbb{K} = \mathbb{K}_0$.

Evolution families arise as solution operators of nonautonomous evolution equations, although not every evolution family solves such a problem. We refer to [4], [9, section VI.9], and the references therein for further information. Condition (E3) is needed only in the study of asymptotic properties in section 5; see Remark 4.5.

DEFINITION 2.2. *Let* T *be an evolution family on* X *and* $\Psi_s : X \to L^p_{loc}([s,\infty), Y)$, $s \geq 0$, *be linear operators satisfying*

$$\Psi_s x = \Psi_t T(t,s)x \quad \text{on } [t,\infty) \quad \text{and} \quad \int_s^{s+t_0} \|(\Psi_s x)(t)\|_Y^p \, dt \leq \gamma^p \|x\|_X^p \tag{2.1}$$

for $t \geq s \geq 0$, $x \in X$, and some $t_0 > 0$, $\gamma = \gamma(t_0) > 0$. Then $(T, \Psi) = (T, \{\Psi_s : s \geq 0\})$ is a nonautonomous observation system. *We extend the map $\Psi_s x$ by 0 to $\mathbb{R}$.*

LEMMA 2.3. *Let (T, Ψ) be a nonautonomous observation system. Then one can replace the constant t_0 in (2.1) by every $t_1 > 0$ and $\gamma = \gamma(t_0)$ by $\gamma(t_1) = c_0 M \gamma(t_0) c(t_1)$, where $c(t) = e^{w^+ t}$ for $w \neq 0$, $c(t) = (1 + \frac{t}{t_0})^{\frac{1}{p}}$ for $w = 0$, and c_0 depends on t_0, w, p.*

Proof. The case $t_1 \leq t_0$ is obvious. So let $t_1 = nt_0 + \tau$ for some $n \in \mathbb{N}$ and $\tau \in [0, t_0)$. Setting $I_k = [s + kt_0, s + (k+1)t_0]$, we deduce from Definition 2.2 that

$$\|\Psi_s x\|^p_{L^p([s,s+t_1],Y)} \leq \sum_{k=0}^{n} \|\Psi_{s+kt_0} T(s + kt_0, s)x\|^p_{L^p(I_k,Y)} \leq M^p \, \gamma(t_0)^p \sum_{k=0}^{n} e^{wpt_0 k}$$

for $x \in X$ and $s \geq 0$. The assertion then follows easily. □

DEFINITION 2.4. *Let T be an evolution family on X and $C(s) : D(C(s)) \subseteq X \to Y$, $s \geq 0$, be densely defined linear operators such that $T(\cdot, s)x \in D(C(\cdot), s) := \{f \in L^p_{loc}([s, \infty), X) : f(t) \in D(C(t))$ for a.e. $t \geq s$, $C(\cdot)f(\cdot) \in L^p_{loc}([s, \infty), Y)\}$ and*

$$\int_s^{s+t_0} \|C(t)T(t, s)x\|^p_Y \, dt \leq \gamma^p \, \|x\|^p_X \tag{2.2}$$

for $s \geq 0$, $x \in D(C(s))$, and some constants $\gamma, t_0 > 0$. Then we say that $C(s)$, $s \geq 0$, are (T-)admissible observation operators.

LEMMA 2.5. *Let $C(s)$, $s \geq 0$, be T-admissible observation operators. Then (2.2) holds for all $t_0 > 0$ with a possibly different γ. Let $\Psi_s : X \to L^p_{loc}([s, \infty), Y)$, $s \geq 0$, be the continuous extension of the map $D(C(s)) \ni x \mapsto C(\cdot)T(\cdot, s)x$. Then (T, Ψ) is a nonautonomous observation system.*

Proof. The first claim can be established as Lemma 2.3; one has only to replace $s + kt_0$ by points $s_k \approx s + kt_0$ such that $T(s_k, s)x \in D(C(s_k))$; see [23, Lem. 4.13]. So we can define Ψ_s as in the claim. Given $t \geq s \geq 0$ and $x \in D(C(s))$, we take $z_n \in D(C(t))$ converging in X to $T(t, s)x$ and $t_n \searrow t$ such that $T(t_n, s)x, T(t_n, t)z_n \in D(C(t_n))$. Since $\Psi_t T(t, s)x = \lim_{n\to\infty} \mathbb{1}_{[t_n, t+t_0]} \, C(\cdot)T(\cdot, t)z_n$ in $L^p([t, t + t_0], Y)$, we obtain

$$\begin{aligned}
&\|\Psi_t T(t, s)x - C(\cdot)T(\cdot, s)x\|^p_{L^p([t,t+t_0],Y)} \\
&= \lim_{n\to\infty} \left[\int_{t_n}^{t+t_0} \|C(\tau)T(\tau, t_n) \, [T(t_n, t)z_n - T(t_n, s)x]\|^p d\tau + \int_t^{t_n} \|C(\tau)T(\tau, s)x\|^p d\tau \right] \\
&\leq \gamma^p \lim_{n\to\infty} \|T(t_n, t)z_n - T(t_n, s)x\|^p = 0.
\end{aligned}$$

Therefore, (2.1) holds for $x \in D(C(s))$ and thus for $x \in X$ by approximation. □

We note that different admissible observation operators $C_1(s)$ and $C_2(s)$ may yield the same observation system as shown in [28, Ex. 1.2]. However, if the observation operators $C(s)$ are closable, then one easily verifies that $\Psi_s x = \overline{C(\cdot)}T(\cdot, s)x$ for the induced observation system. We now proceed in the converse direction and represent a given observation system by admissible observation operators; cf. [28, Def. 4.1].

DEFINITION 2.6. *For a nonautonomous observation system (T, Ψ), we define*

$$C(s)x = \lim_{\tau \searrow 0} \frac{1}{\tau} \int_s^{s+\tau} (\Psi_s x)(\sigma) \, d\sigma \qquad \textit{(in } Y\textit{)} \tag{2.3}$$

for $x \in \underline{X}_s := \{x \in X :$ *the limit in* (2.3) *exists*$\}$ *and*

$$\|x\|_{\underline{X}_s} = \|x\|_X + \sup_{0<\tau\le 1} \left\| \frac{1}{\tau} \int_s^{s+\tau} (\Psi_s x)(\sigma)\, d\sigma \right\|_Y$$

for $x \in \underline{X}_s$ *and* $s \ge 0$. *The space* $D(C(\cdot), s)$ *is defined as in Definition* 2.4 *by replacing* $D(C(t))$ *with* $\underline{X}_t$.

Clearly, $\|\cdot\|_{\underline{X}_s}$ is a norm on the subspace $\underline{X}_s$ and $C(s) : \underline{X}_s \to Y$ is linear and continuous. As in [28, Prop. 4.3], one verifies that $(\underline{X}_s, \|\cdot\|_{\underline{X}_s})$ is complete.

We say that $t \in \mathbb{R}$ is a *p-Lebesgue point* of $f \in L^p_{loc}(\mathbb{R}, Z)$, $1 \le p < \infty$, if

$$\lim_{|J|\to 0} \frac{1}{|J|} \int_J \|f(s) - f(t)\|^p\, ds = 0,$$

where the limit is taken over compact intervals J containing t (of length $|J|$). If $p = 1$, then t is called the *Lebesgue point*. Recall that a.e. t is a p-Lebesgue point of $f \in L^p_{loc}(\mathbb{R}, Z)$; see, e.g., [31, Lem. 6.1] or [26, section I.1.8]. The next representation theorem extends [28, Thm. 4.5] to nonautonomous observation systems. A different representation of output functions was given in [11] applying Weiss' theory to the "evolution semigroup" on $L^p([0, t_0], X)$ associated with T; cf. [4].

THEOREM 2.7. *Let* (T, Ψ) *be a nonautonomous observation system, and let* $C(s) \in \mathcal{L}(\underline{X}_s, Y)$ *be given as in Definition* 2.6. *Let* $x \in X$ *and* $t \ge s \ge 0$. *Then* $T(t, s)x \in \underline{X}_t$ *if and only if* $1/\tau \int_0^\tau (\Psi_s x)(t + \sigma)\, d\sigma$ *converges as* $\tau \searrow 0$. *If this is the case, then the limit equals* $C(t)T(t, s)x$. *Thus* $(\Psi_s x)(t) = C(t)T(t, s)x$ *for all Lebesgue points* t *of* $\Psi_s x$.

Proof. The theorem follows from the identity

$$\frac{1}{\tau} \int_t^{t+\tau} [\Psi_s x](\sigma)\, d\sigma = \frac{1}{\tau} \int_t^{t+\tau} [\Psi_t T(t, s)x](\sigma)\, d\sigma. \qquad \square$$

This theorem shows that the operators $C(t)$, $t \ge 0$, introduced in Definition 2.6 are admissible observation operators. According to Lemma 2.5, they generate an observation system $(\tilde{\Psi}, T)$. It is easy to see that, in fact, $(\Psi_s x)(t) = (\tilde{\Psi}_s x)(t)$ for each $x \in X$ and a.e. $t \ge s$. We say that the operators $C(t)$ from Definition 2.6 *represent the observation system* (T, Ψ).

In the remainder of this section, we establish several properties of Ψ_s which will be important for our main perturbation result.

LEMMA 2.8. *Let* (T, Ψ) *be a nonautonomous observation system,* $f \in L^p_{loc}(\mathbb{R}_+, X)$, *and* $t_0 > 0$. *Then the map* $[0, t_0] \ni s \mapsto \Psi_s f(s) \in L^p([0, t_0], Y)$ *is measurable, and*

$$\int_0^{t_0} \|\Psi_s f(s)\|^p_{L^p([0,t_0],Y)}\, ds \le \gamma(t_0)^p \, \|f\|^p_{L^p([0,t_0],X)}\,. \tag{2.4}$$

Proof. For $f \in C(\mathbb{R}_+, X)$, the map $s \mapsto \Psi_s f(s)$ is continuous from the right since

$$\|\Psi_s f(s) - \Psi_r f(r)\|^p_{L^p([0,t_0],Y)} = \|\Psi_s(f(s) - T(s, r)f(r))\|^p_{L^p([s,t_0],Y)} + \|\Psi_r f(r)\|^p_{L^p([r,s],Y)}$$

for $0 \le r \le s \le t_0$. Functions $f \in L^p_{loc}(\mathbb{R}_+, X)$ can be treated by approximation. The estimate (2.4) follows from (2.1). $\square$

The nonclosedness of $C(t)$ is a major obstacle for the analysis of observation systems and input–output operators; for instance, it is a priori not clear whether $C(t)$

can be taken out of an integral. As in the autonomous case (see, e.g., [31, section 4]), such problems can be overcome by employing the operators

$$C_\tau(s)x = \frac{1}{\tau}\int_s^{s+\tau} (\Psi_s x)(\sigma)\,d\sigma, \tag{2.5}$$

$x \in X$, $s \geq 0$, and $\tau \in (0,1]$. Due to this definition, $C_\tau(s)$ belongs to $\mathcal{L}(X,Y)$ with norm less than or equal to $\gamma(1)\tau^{-\frac{1}{p}}$, $C_\tau(s)x$ converges as $\tau \to 0$ if and only if $x \in \underline{X}_s$, and then the limit equals $C(s)x$. Let $C_c(\mathbb{R}_+)$ be the space of continuous functions with compact support in $[0,\infty)$. We also define

$$\mathcal{D}_s = \operatorname{span}\{\varphi(\cdot)T(\cdot,r)x : x \in X,\, r \geq s,\, \varphi \in C_c(\mathbb{R}_+),\, \varphi(t) = 0 \text{ for } s \leq t < r\} \tag{2.6}$$

for $s \geq 0$ (setting $T(t,s) := 0$ for $t < s$), and we put $\mathcal{D} = \mathcal{D}_0$. This space is dense in $L^p([s,\infty),X)$ and in $C_0([s,\infty),X)$, the space of continuous functions vanishing at infinity. This fact can be seen by an obvious modification of the proof of [4, Thm. 3.12].

LEMMA 2.9. *Let (T,Ψ) be a nonautonomous observation system represented by $C(s)$, and let $C_\tau(s)$ be given by (2.5). Then $(s,\tau) \mapsto C_\tau(s)x$ is continuous on $\mathbb{R}_+ \times (0,1]$,*

$$\|C_\tau(\cdot)T(\cdot,s)x\|_{L^p([s,s+t_0],Y)} \leq \gamma(t_0+1)\,\|x\|, \quad \textit{and} \tag{2.7}$$

$$\Psi_s x = \lim_{\tau\to 0} C_\tau(\cdot)T(\cdot,s)x \quad \textit{in} \quad L^p_{loc}([s,\infty),Y) \tag{2.8}$$

for $x \in X$, $s \geq 0$, $\tau \in (0,1]$, and $t_0 > 0$.

Proof. If $f \in \mathcal{D}$, then $(t,\tau) \mapsto C_\tau(t)f(t)$ is continuous since

$$C_\tau(t)f(t) = \sum_{k=1}^n \varphi_k(t)\,\frac{1}{\tau}\int_t^{t+\tau} (\Psi_{r_k}x_k)(\sigma)\,d\sigma$$

for $\tau > 0$, $t \geq 0$, and suitable $n \in \mathbb{N}$, $r_k \geq 0$, $x_k \in X$, $\varphi_k \in C_c(\mathbb{R}_+)$. The first assertion follows by approximation. We further estimate

$$\begin{aligned}&\|C_\tau(\cdot)T(\cdot,s)x\|^p_{L^p([s,s+t_0],Y)}\\ &\leq \int_s^{s+t_0+\tau} \frac{1}{\tau}\int_{\sigma-\tau}^{\sigma} \|(\Psi_s x)(\sigma)\|^p\,dt\,d\sigma \leq \gamma(t_0+1)^p\,\|x\|^p\end{aligned}$$

using Hölder's inequality and Fubini's theorem. Similarly, (2.8) follows from

$$\begin{aligned}&\|C_\tau(\cdot)T(\cdot,s)x - \Psi_s x\|^p_{L^p([s,s+t_0],Y)}\\ &\leq \frac{1}{\tau}\int_0^\tau \int_s^{s+t_0} \|(\Psi_s x)(t+\sigma) - (\Psi_s x)(t)\|^p\,dt\,d\sigma. \qquad \square\end{aligned}$$

We want to show that $C(\cdot)\mathbb{K}_s : L^p([s,s+t_0],X) \to L^p([s,s+t_0],Y)$ is well defined and bounded. This fact is crucial for Theorem 4.4, and its proof is somewhat technical. We set

$$\varphi(t;\,\tau,\sigma,f) = (C_\tau(t) - C_\sigma(t))\int_0^t T(t,s)f(s)\,ds,$$

$$o_f(t) = \overline{\lim_{\tau,\sigma\to 0}}\,\|\varphi(t;\,\tau,\sigma,f)\| = \overline{\lim_{n\to\infty}} \sup_{m\geq n} \sup_{\tau,\sigma\in[1/m,1/n]} \|\varphi(t;\,\tau,\sigma,f)\|$$

for $f \in L^1_{loc}(\mathbb{R}_+, X)$, $t \geq 0$, $\tau, \sigma \in (0,1]$. Observe that o_f is measurable. We further need the *maximal operator* given by

$$M\psi(t) = \sup_{\tau>0} \frac{1}{\tau} \int_t^{t+\tau} |\psi(s)|\, ds \ \in [0,\infty]$$

for all $t \in \mathbb{R}$ and $\psi \in L^1_{loc}(\mathbb{R})$. Recall that

$$\|M\psi\|_{L^p(\mathbb{R})} \leq c_p \, \|\psi\|_{L^p(\mathbb{R})} \tag{2.9}$$

for $\psi \in L^p(\mathbb{R})$, $1 < p \leq \infty$, and a constant c_p; see [26, Thm. I.1].

LEMMA 2.10. *Let (T, Ψ) be a nonautonomous observation system, $p \in (1,\infty)$, and $f \in L^p_{loc}(\mathbb{R}_+, X)$. Then $C_\tau(t)(\mathbb{K}f)(t) \to C(t)(\mathbb{K}f)(t)$ as $\tau \searrow 0$ for a.e. $t \geq 0$.*

Proof. Take $g \in \mathcal{D}$ and $f \in L^p_{loc}(\mathbb{R}_+, X)$. Observe that $o_g = 0$ a.e. because of

$$C_\tau(t)(\mathbb{K}g)(t) = \sum_{k=1}^n \int_0^t \varphi_k(s)\, ds \, \frac{1}{\tau} \int_t^{t+\tau} (\Psi_{r_k} x_k)(\sigma)\, d\sigma$$

for $\tau > 0$, $t \geq 0$, and suitable $n \in \mathbb{N}$, $r_k \geq 0$, $x_k \in X$, and $\varphi_k \in C_c(\mathbb{R}_+)$. Due to Lemma 2.8, there is a measurable function $\psi_{f-g} : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\text{for a.e. } \ s \geq 0, \qquad \|[\Psi_s(f(s) - g(s))](t)\|_Y = \psi_{f-g}(t,s) \quad \text{for a.e. } \ t \geq s.$$

(Here we set $[\Psi_s(f(s) - g(s))](t) = \psi_{f-g}(t,s) = 0$ for $t < s$, $t > t_0$, or $s > t_0$, where $t_0 > 0$ is fixed but arbitrary.) Employing these facts, we estimate

$$\begin{aligned} o_f(t) \leq o_{f-g}(t) &\leq \sup_{\tau,\sigma\in(0,1]} \int_0^t \|(C_\tau(t) - C_\sigma(t))\, T(t,s)(f(s) - g(s))\|\, ds \\ &\leq 2 \sup_{\tau\in(0,1]} \int_0^t \frac{1}{\tau} \int_t^{t+\tau} \|[\Psi_s(f(s) - g(s))](\rho)\|\, d\rho\, ds \\ &\leq 2 \int_0^t [M\psi_{f-g}(\cdot, s)](t)\, ds \end{aligned} \tag{2.10}$$

for t not contained in a set of measure 0 depending on g. Approximating $0 \leq \phi \in L^1_{loc}(\mathbb{R}^2)$ by continuous functions, one sees that $(t,s) \mapsto [M\phi(\cdot,s)](t)$ is measurable. We can now use (2.10), Fubini's theorem, and the maximal inequality (2.9) to derive

$$\begin{aligned} |\{t \in [0,t_0] : o_f(t) > \varepsilon\}| &\leq \frac{2}{\varepsilon} \int_0^{t_0} \int_0^t \left[M\psi_{f-g}(\cdot,s)\right](t)\, ds\, dt \\ &\leq \frac{c}{\varepsilon} \int_0^{t_0} \|M\psi_{f-g}(\cdot,s)\|_{L^p(\mathbb{R})}\, ds \\ &\leq \frac{c'}{\varepsilon} \int_0^{t_0} \|\psi_{f-g}(\cdot,s)\|_{L^p([s,t_0])}\, ds \\ &= \frac{c'}{\varepsilon} \int_0^{t_0} \|\Psi_s(f(s) - g(s))\|_{L^p([s,t_0],Y)}\, ds \\ &\leq \frac{\gamma c'}{\varepsilon} \|f - g\|_{L^1([0,t_0],X)} \end{aligned}$$

for each $\varepsilon > 0$ and constants c, c' not depending on f, g, ε. Since g is arbitrary, the set $\{o_f > \varepsilon\}$ has Lebesgue measure 0. This fact implies the assertion. ∎

PROPOSITION 2.11. *Let (T,Ψ) be a nonautonomous observation system represented by $C(t)$, $p\in(1,\infty)$, and let $C_\tau(t)$ be given by (2.5). Then $\mathbb{K}_s f\in D(C(\cdot),s)$,*

$$\|C(\cdot)\mathbb{K}_s f\|_{L^p([s,s+t_0],Y)}\le c(t_0)\,\|f\|_{L^p([s,s+t_0],X)}\,,$$

and $C_\tau(\cdot)\mathbb{K}_s f\to C(\cdot)\mathbb{K}_s f$ in $L^p_{loc}([s,\infty),Y)$ as $\tau\to 0$ for $s\ge 0$, $t_0>0$, $f\in L^p_{loc}(\mathbb{R}_+,X)$, and a constant $c(t_0)$ independent of f and s.

Proof. By Lemma 2.10, $C(\cdot)\mathbb{K}_s f$ is a well-defined measurable function. Further,

$$\begin{aligned}\int_s^{s+t_0}\|C_\tau(t)(\mathbb{K}_s f)(t)\|^p\,dt &\le c\int_s^{s+t_0}\int_s^t\|C_\tau(t)T(t,r)f(r)\|^p\,dr\,dt\\ &= c\int_s^{s+t_0}\int_r^{s+t_0}\|C_\tau(t)T(t,r)f(r)\|^p\,dt\,dr\\ &\le c\tilde\gamma^p\,\|f\|^p_{L^p([s,s+t_0],X)}\end{aligned}\tag{2.11}$$

for a constant c by Hölder's inequality, Fubini's theorem, and Lemma 2.9. Similarly,

$$\|(C_\tau(\cdot)-C_\sigma(\cdot))\mathbb{K}_s f\|_p^p\le c\int_s^{s+t_0}\int_r^{s+t_0}\|(C_\tau(t)-C_\sigma(t))T(t,r)f(r)\|^p\,dt\,dr,$$

and the right side tends to 0 as $\tau,\sigma\to 0$ by Lemma 2.9 and the dominated convergence theorem. Hence $C_\tau(\cdot)\mathbb{K}_s f$ also converges in $L^p([s,s+t_0],Y)$ to $C(\cdot)\mathbb{K}_s f$. The asserted estimate then follows from (2.11). ∎

3. Well-posed and regular nonautonomous systems.

DEFINITION 3.1. *Let T be an evolution family on X, and let $\Phi_{t,s}=\Phi(t,s):L^p_{loc}([s,\infty),U)\to X$, $t\ge s\ge 0$, be linear operators satisfying*

$$\Phi_{t,s}u=\Phi_{t,r}(u|[r,\infty))+T(t,r)\Phi_{r,s}u,\quad t\ge r\ge s\ge 0,\quad \textit{and}\tag{3.1}$$

$$\|\Phi_{t,s}u\|_X\le\beta\,\|u\|_{L^p([s,t],U)},\quad 0\le t-s\le t_0,\tag{3.2}$$

for $u\in L^p(\mathbb{R}_+,U)$ and constants $t_0>0$, $\beta=\beta(t_0)>0$. Then $(T,\Phi)=(T,\{\Phi_{t,s}:t\ge s\ge 0\})$ is called a nonautonomous control system.

Observe that the above definition implies that $\Phi_{t,t}=0$ and $\Phi_{t,s}u=\Phi_{t,r}u$ if $u=0$ on $[s,r]\subseteq[s,t]$. Thus the control system is *causal*.

LEMMA 3.2. *Let (T,Φ) be a nonautonomous control system. Then*

$$\|\Phi_{t,s}u\|_X\le c_0'M\beta(t_0)\,c(t-s)\,\|u\|_{L^p([s,t],U)},\tag{3.3}$$

$$\|\Phi(\cdot,s)u\|_{L^p([s,t],X)}\le c_0'M\beta(t_0)\,c(t-s)\,\|u\|_{L^p([s,t],U)}\tag{3.4}$$

for $t\ge s\ge 0$, $u\in L^p([s,t],U)$, and $c_0'=c_0'(t_0,w,p)$ ($c(t)$ was defined in Lemma 2.3).

Proof. In Lemma 3.4, we show the measurability of $\Phi(\cdot,s)u$ (of course without referring to (3.4)). The assertion is clear for $s\le t\le s+t_0$. Let $s_k=s+kt_0$ for $k\in\mathbb{N}_0$, let $t\in[s_n,s_{n+1}]$ for some $n\in\mathbb{N}$, and let u_k be the restriction of u to $[s_k,s_{k+1}]\cap[s,t]$ for $k=0,\dots,n$. Then

$$\Phi(t,s)u=\Phi(t,s_n)u_n+\sum_{k=1}^n T(t,s_k)\Phi(s_k,s_{k-1})u_{k-1}\,,\tag{3.5}$$

$$\|\Phi(t,s)u\|\le\beta\,\|u_n\|_p+M\beta\sum_{k=1}^n e^{w(t-s_k)}\,\|u_{k-1}\|_p\le M\beta e^{w^-t_0}\,(a*b)_n\,,$$

where $a_k = e^{wt_0 k}$ if $k = 0, \dots, n$ and $a_k = 0$ otherwise, $b_k = \|u_k\|_p$, $a = (a_k)_k$, and $b = (b_k)_k$. Young's inequality now implies the lemma. ∎

DEFINITION 3.3. *Let T be an evolution family on X, and let $\overline{X}_t$, $t \geq 0$, be Banach spaces in which X is densely and continuously embedded. Assume that $T(t,s)$ has a locally uniformly bounded extension $\overline{T}(t,s) : \overline{X}_s \to \overline{X}_t$ (which then satisfies* (E1) *and is strongly continuous with respect to s). We call $B(t) \in \mathcal{L}(U, \overline{X}_t)$, $t \geq 0$,* (T-)admissible control operators *if the function $\overline{T}(t,\cdot)B(\cdot)u(\cdot)$ is integrable in $\overline{X}_t$,*

$$(\overline{\mathbb{K}}_s B(\cdot)u)(t) := \int_s^t \overline{T}(t,\tau)B(\tau)u(\tau)\,d\tau \in X,$$

and there are constants $t_0, \beta > 0$ such that

$$\|(\overline{\mathbb{K}}_s B(\cdot)u)(t)\|_X \leq \beta \|u\|_{L^p([s,t],U)} \tag{3.6}$$

for all $0 \leq s \leq t \leq s + t_0$ and $u \in L^p([s,t],U)$. (We omit the subscript s if $s = 0$.)

Setting $\Phi_{t,s}u := (\overline{\mathbb{K}}_s B(\cdot)u)(t)$, we obtain, of course, a nonautonomous control system (T, Φ) if $B(t)$, $t \geq 0$, are admissible control operators. Every autonomous control system is given by a T-admissible control operator due to [29, Thm. 3.9], where $\overline{X}_t$, $t \geq 0$, coincide with the extrapolation space X_{-1} of X with respect to the semigroup T (see, e.g., [2, Chap. V], [9, section II.5]). In Proposition 3.5, we extend this result to the time dependent setting but only in an approximate sense because of the lack of an extrapolation theory for evolution families. We first show a preliminary fact.

LEMMA 3.4. *Let (T, Φ) be a nonautonomous control system, and let $u \in L^p_{loc}(\mathbb{R}_+, U)$. Then $t \mapsto \Phi_{t,s}u \in X$ is continuous from the right for $t \geq s$, $s \mapsto \Phi_{t,s}u \in X$ is continuous for $s \in [0,t]$ (locally uniformly in t), and $(t,s) \mapsto \Phi_{t,s}u \in X$ is measurable.*

Proof. Definition 3.1 implies the estimates

$$\begin{aligned}
\|\Phi(t',s)u - \Phi(t,s)u\| &\leq \|\Phi(t',t)u\| + \|(T(t',t) - I)\Phi(t,s)u\| \\
&\leq \beta \|u\|_{L^p([t,t'],U)} + \|(T(t',t) - I)\Phi(t,s)u\|, \\
\|\Phi(t,s)u - \Phi(t,s')u\| &= \|T(t,s')\Phi(s',s)u\| \leq M\beta e^{|w|(t-s)} \|u\|_{L^p([s,s'],U)},
\end{aligned}$$

where $t' \geq t \geq s' \geq s$. Thus the lemma is established. ∎

Let $u \in L^p_{loc}(\mathbb{R}, U)$, $t \geq 0 \geq s$, and $n \in \mathbb{N}$. We define $(B_n u)(t) = n\,\Phi(t, t - \frac{1}{n})u$, where $\Phi(t,s)u := \Phi(t,0)u$. Note that $B_n u \in L^\infty_{loc}(\mathbb{R}_+, X)$ because of the above lemma. To approximate Φ, we introduce

$$\Phi^n(t,s)u = \Phi^n_{t,s}u := \int_s^t T(t,\tau)(B_n u)(\tau)\,d\tau = (\mathbb{K}_s B_n u)(t) \tag{3.7}$$

for $t \geq s \geq 0$, $n \in \mathbb{N}$, and $u \in L^p_{loc}(\mathbb{R}, U)$. These operators can be expressed by

$$\begin{aligned}
\Phi^n(t,s)u &= n \int_s^t \left(\Phi\left(t, \tau - \frac{1}{n}\right) u - \Phi(t,\tau)u \right) d\tau \\
&= n \int_{s-\frac{1}{n}}^{s} \Phi(t,\tau)u\,ds - n \int_{t-\frac{1}{n}}^{t} \Phi(t,\tau)u\,d\tau \\
&= \Phi(t,s)u + nT(t,s) \int_{s-\frac{1}{n}}^{s} \Phi(s,\tau)u\,d\tau - n \int_{t-\frac{1}{n}}^{t} \Phi(t,\tau)u\,d\tau
\end{aligned} \tag{3.8}$$

due to (3.1). If we take a function $u \in L^p_{loc}([s,\infty),U)$ and extend it by 0 to $\mathbb{R}$, then

$$\Phi(t,s)u - \Phi^n(t,s)u = n\int_{t-\frac{1}{n}}^{t} \Phi(t,\tau)u\,d\tau =: r_n(t;\,u). \tag{3.9}$$

To represent Φ approximately, we define operators $B_n(t) \in \mathcal{L}(U,X)$ by

$$B_n(t)z := (B_n u_z)(t) = n\Phi\left(t, t-\frac{1}{n}\right)u_z\,, \qquad \text{where } \ u_z \equiv z,\ z \in U.$$

PROPOSITION 3.5. *Let (T,Φ) be a nonautonomous control system, $n \in \mathbb{N}$, $0 \le s \le t \le s+t_0$, $t_0 > 0$, $z \in U$, and $u \in L^p_{loc}(\mathbb{R},U)$. Then we have the following:*

1. *$\Phi^n(t,s)u \to \Phi(t,s)u$, and $\|\Phi^n(t,s)u\|_X \le 2\beta(t_0)\|u\|_{L^p([s,t],U)}$.*
2. *$(t,s) \mapsto \Phi(t,s)u$, and $t \mapsto B_n(t)z$ are continuous in X.*
3. *$[\mathbb{K}_s B_n(\cdot)u](t) \to \Phi(t,s)u$, and $\|[\mathbb{K}_s B_n(\cdot)u](t)\|_X \le \beta(t_0+1)\,\|u\|_{L^p([s,t],U)}$.*

Here the limits as $n \to \infty$ are taken in X and are locally uniform in (t,s).

Proof. For $u \in L^p_{loc}(\mathbb{R},U)$, we estimate

$$\left\| n\int_{t-\frac{1}{n}}^{t} \Phi(t,\tau)u\,d\tau \right\|_X \le \sup_{t-\frac{1}{n}\le\tau\le t} \|\Phi(t,\tau)u\|_X \le \beta\,\|u\|_{L^p([t-\frac{1}{n},t],U)}\,,$$

$$\left\| nT(t,s)\int_{s-\frac{1}{n}}^{s} \Phi(s,\tau)u\,d\tau \right\|_X \le M\beta e^{w(t-s)}\,\|u\|_{L^p([s-\frac{1}{n},s],U)}\,,$$

which yields the first part of (1) because of (3.8). This fact implies (2). The second part of (1) follows from (3.9) if we extend $u \in L^p([s,\infty),X)$ by 0 to $\mathbb{R}$. We set $\tilde{u}(\tau,\sigma) = u(\tau)$ and $u^{(n)}(\sigma) = n\int_\sigma^{\sigma+\frac{1}{n}} u(\tau)\,d\tau$ for $\sigma \ge \tau \ge 0$ and $n \in \mathbb{N}$. Taking first $u \in W^{1,p}_{loc}(\mathbb{R},U)$, using Hölder's inequality, and interchanging integrals, we estimate

$$\begin{aligned}
\|B_k u - B_k(\cdot)u\|_{L^1([s,s+t_0],X)} &\le k\int_s^{s+t_0} \left\|\Phi\left(\tau,\tau-\frac{1}{k}\right)[u-\tilde{u}(\tau,\cdot)]\right\|_X d\tau \\
&\le \beta k\int_s^{s+t_0} \left(\int_{\tau-\frac{1}{k}}^{\tau}\left(\int_\sigma^\tau \|u'(\rho)\|_U\,d\rho\right)^p d\sigma\right)^{\frac{1}{p}} d\tau \\
&\le \beta k^{1-\frac{1}{p}}\int_s^{s+t_0}\int_0^{\frac{1}{k}} \|u'(\tau-\rho)\|_U\,d\rho\,d\tau \\
&\le c\left(\int_0^{\frac{1}{k}}\int_s^{s+t_0} \|u'(\tau-\rho)\|_U^p\,d\tau\,d\rho\right)^{\frac{1}{p}}
\end{aligned}$$

so that $[\mathbb{K}_s B_k(\cdot)u](t) \to \Phi(t,s)u$ as $k \to \infty$ locally uniformly in this case. Fix now $t > s \ge 0$, and extend $u \in L^p([s,t],X)$ by 0 to $\mathbb{R}$. Then (3.1) and the above results imply

$$
\begin{aligned}
[\mathbb{K}_s B_n(\cdot)u](t) &= n\int_{s-\frac{1}{n}}^{t} \Phi(t,\tau)\left[\tilde{u}\left(\tau+\frac{1}{n},\cdot\right)-\tilde{u}(\tau,\cdot)\right]d\tau \\
&= \lim_{k\to\infty} n\int_{s-\frac{1}{n}}^{t}\int_{\tau}^{t} T(t,\sigma)B_k(\sigma)\left[u\left(\tau+\frac{1}{n}\right)-u(\tau)\right]d\sigma\,d\tau \\
&= \lim_{k\to\infty} n\int_{s-\frac{1}{n}}^{t} T(t,\sigma)B_k(\sigma)\int_{s-\frac{1}{n}}^{\sigma}\left[u\left(\tau+\frac{1}{n}\right)-u(\tau)\right]d\tau\,d\sigma \\
&= \Phi\left(t,s-\frac{1}{n}\right)u^{(n)}.
\end{aligned}
$$

Observing that $\|u^{(n)}\|_{L^p([s,t],U)} \le \|u\|_{L^p([s,t],U)}$, we deduce (3). □

We give, as in [13], the nonautonomous analogue of Weiss' definition of a *well-posed system*; see [30, Def. 1.1].

DEFINITION 3.6. *Let (T,Φ) and (T,Ψ) be nonautonomous control and observation systems. If there are linear operators $\mathbb{F}_s : L^p_{loc}([s,\infty),U) \to L^p_{loc}([s,\infty),Y)$, $s \ge 0$, satisfying*

$$\mathbb{F}_s u = \Psi_t \Phi_{t,s} u + \mathbb{F}_t(u|[t,\infty)) \quad \text{on } [t,\infty) \quad \text{and} \tag{3.10}$$

$$\|\mathbb{F}_s u\|_{L^p([s,s+t_0],Y)} \le \kappa \|u\|_{L^p([s,s+t_0],U)} \tag{3.11}$$

for $u \in L^p_{loc}([s,\infty),U)$, $t \ge s \ge 0$, and constants $t_0 > 0$, $\kappa = \kappa(t_0) > 0$, then $\Sigma = (T,\Phi,\Psi,\mathbb{F}) = (T,\Phi_{t,s},\Psi_s,\mathbb{F}_s)_{t\ge s\ge 0}$ is called a well-posed nonautonomous system, *and $\mathbb{F}_s$, $s \ge 0$, are called* input–output operators. *We put $\mathbb{F} = \mathbb{F}_0$.*

The above definition implies that $\mathbb{F}_s u = 0$ on $[s,t]$ and $\mathbb{F}_s u = \mathbb{F}_t(u|[t,\infty))$ on $[t,\infty)$ if u vanishes on $[s,t]$. Hence $\mathbb{F}_s$ is *causal*, and we can define its restriction as

$$\mathbb{F}_{t,s} = \mathbb{F}(t,s) : L^p([s,t],U) \to L^p([s,t],Y), \qquad t \ge s \ge 0.$$

LEMMA 3.7. *A well-posed nonautonomous linear system Σ satisfies* (3.11) *with t_0 replaced by each $t_1 > 0$ and $\kappa = \kappa(t_0)$ by $\kappa(t_1) = c_0''(\kappa(t_0) \vee M\beta(t_0)\gamma(t_0))c(t_1)$, where $c_0'' = c_0''(w,t_0)$ and $c(t)$ was defined in Lemma* 2.3.

Proof. The assertion is clear for $t_1 \le t_0$. So let $t_1 \in [s_n, s_{n+1})$ for some $n \in \mathbb{N}$, $s_k = s + kt_0$, $I_k = [s_k, s_{k+1}]$, and $u_k = u|I_k$ for $k \in \mathbb{N}_0$, $s \ge 0$, and $u \in L^p_{loc}([s,\infty),U)$. From Definition 3.6 and (3.5), we deduce that

$$\mathbb{F}_s u = \mathbb{F}_{s_k} u_k + \sum_{j=1}^{k} \Psi_{s_k} T(s_k,s_j)\Phi(s_j,s_{j-1})u_{j-1} \quad \text{on } I_k,$$

$$
\begin{aligned}
\|\mathbb{F}_s u\|_{L^p(I_k,Y)} &\le \kappa(t_0)\|u_k\|_p + M\beta(t_0)\gamma(t_0)\sum_{j=1}^{k} e^{wt_0(k-j)}\|u_{j-1}\|_p \\
&\le (\kappa(t_0) \vee M\beta(t_0)\gamma(t_0))\, e^{w^- t_0}\,(a * b)_k\,,
\end{aligned}
$$

$$\|\mathbb{F}_s u\|_{L^p([s,s+t_1],Y)} \le \left(\sum_{k=0}^{n}\|\mathbb{F}_s u\|^p_{L^p(I_k,Y)}\right)^{\frac{1}{p}} \le (\kappa(t_0) \vee M\beta(t_0)\gamma(t_0))\, e^{w^- t_0}\,\|a * b\|_{\ell^p}\,,$$

where the sequences a and b were defined in the proof of Lemma 3.2. Young's inequality now implies the asserted estimate. □

Also, Definition 3.6 is complemented by a concept involving admissible input and output operators; cf. [10], [12, section 1.3], [14].

DEFINITION 3.8. *Let $B(s)$ and $C(s)$, $s \geq 0$, be T-admissible control and observation operators. We call the triple $(T, B(\cdot), C(\cdot))$ an* admissible nonautonomous system *if $\overline{\mathbb{K}}_s B(\cdot)u \in D(C(\cdot), s)$ and $\|C(\cdot)\overline{\mathbb{K}}_s B(\cdot)u\|_{L^p([s,s+t_0],Y)} \leq \kappa \|u\|_{L^p([s,s+t_0],U)}$ for $s \geq 0$, $u \in L^p_{loc}([s,\infty),U)$, and constants $\kappa, t_0 > 0$.*

LEMMA 3.9. *Let $(T, B(\cdot), C(\cdot))$ be an admissible nonautonomous system. Define Ψ_s as in Lemma 2.5, $\Phi_{t,s}u := (\overline{\mathbb{K}}_s B(\cdot)u)(t)$, and $\mathbb{F}_s := C(\cdot)\overline{\mathbb{K}}_s B(\cdot)u$. Then $(T, \Phi, \Psi, \mathbb{F})$ is a well-posed nonautonomous system.*

Proof. In view of Lemma 2.5, we have only to verify (3.10) for $u \in L^p_{loc}([s,\infty),U)$ and $t \geq s \geq 0$. There are $t_n \searrow t$ such that $\Phi_{t_n,t}u, \Phi_{t_n,s}u \in D(C(t_n))$, and hence

$$\mathbb{F}_s u = \mathbb{F}_t u - \Psi_{t_n}\Phi_{t_n,t}u + \Psi_{t_n}\Phi_{t_n,s}u$$

a.e. on $[t_n,\infty)$. The third term on the right-hand side converges in L^p to $\Psi_t\Phi_{t,s}u$ due to Proposition 3.5 and the proof of Lemma 2.8. The assertion then follows from

$$\|\Psi_{t_n}\Phi_{t_n,t}u\|_{L^p([t_n,s+t_0],Y)} \leq \beta\gamma \|u\|_{L^p([t,t_n],U)}\,. \qquad \square$$

In order to prove a converse to the above lemma, we need the first of the following notions, which extends the corresponding concept due to Weiss [30, Def. 4.1].

DEFINITION 3.10. *A well-posed nonautomonous system $\Sigma = (T, \Phi, \Psi, \mathbb{F})$ is called* regular (with feedthrough $D = 0$) *if*

$$\lim_{\tau \searrow 0} \frac{1}{\tau} \int_t^{t+\tau} (\mathbb{F}_t u_z)(\sigma)\, d\sigma = 0 \tag{3.12}$$

(in Y) and absolutely regular *if*

$$\lim_{\tau \searrow 0} \frac{1}{\tau} \int_t^{t+\tau} \|(\mathbb{F}_t u_z)(\sigma)\|_Y^p\, d\sigma = 0 \tag{3.13}$$

for all $t \geq 0$ and $z \in U$, where $u_z(s) := z$ for $s \geq 0$.

We derive several useful properties of a well-posed system Σ. First, (3.11) yields

$$\left\| \frac{1}{\tau} \int_t^{t+\tau} (\mathbb{F}_t u_z)(\sigma)\, d\sigma \right\|_Y^p \leq \frac{1}{\tau} \int_t^{t+\tau} \|(\mathbb{F}_t u_z)(\sigma)\|_Y^p\, d\sigma \leq \kappa^p \|z\|^p \tag{3.14}$$

for $0 < \tau \leq t_0$, $t \geq 0$, and $z \in U$. Take $u \in L^p_{loc}(\mathbb{R}_+, U)$, and set $\tilde{u}(t,\sigma) = u(t)$ for $\sigma \geq t$ and $t \geq 0$. Then the functions

$$t \mapsto F_\tau(t) = \frac{1}{\tau} \int_t^{t+\tau} (\mathbb{F}_t u)(\sigma)\, d\sigma \quad \text{and} \quad t \mapsto \tilde{F}_\tau(t) = \frac{1}{\tau} \int_t^{t+\tau} (\mathbb{F}_t \tilde{u}(t,\cdot))(\sigma)\, d\sigma$$

are measurable for a fixed $\tau > 0$. Indeed, using (3.10), we can write

$$\begin{aligned} \tau\tilde{F}_\tau(t) - \tau\tilde{F}_\tau(r) = &\int_{r+\tau}^{t+\tau} [\mathbb{F}_t \tilde{u}(t,\cdot)](\sigma)\, d\sigma + \int_t^{r+\tau} [\mathbb{F}_t(\tilde{u}(t,\cdot) - \tilde{u}(r,\cdot))](\sigma)\, d\sigma \\ &- \int_t^{r+\tau} [\Psi_t \Phi_{t,r} \tilde{u}(r,\cdot)](\sigma)\, d\sigma - \int_r^t [\mathbb{F}_r \tilde{u}(r,\cdot)](\sigma)\, d\sigma \end{aligned}$$

for $t \geq r \geq t - \tau$. This identity and the straightforward estimates imply the left continuity of $\tilde{F}_\tau$ if u is continuous. Thus $\tilde{F}_\tau$ is measurable by approximation and

(3.14). The function F_τ can be handled in the same way. If Σ is regular, we deduce from Lebesgue's theorem and (3.14) that

$$\lim_{\tau \searrow 0} \int_0^{t_0} \|\tilde{F}_\tau(t)\|_Y^p \, dt = 0. \tag{3.15}$$

Similarly, $\varphi_u(\cdot,\tau)^{\frac{1}{p}}$ is measurable, and absolute regularity yields

$$\lim_{\tau \searrow 0} \int_0^{t_0} \frac{1}{\tau} \varphi_u(t,\tau)\, dt := \lim_{\tau \searrow 0} \int_0^{t_0} \frac{1}{\tau} \int_t^{t+\tau} \|[\mathbb{F}_t \tilde{u}(t,\cdot)](\sigma)\|_Y^p \, d\sigma \, dt = 0. \tag{3.16}$$

We now show a nonautonomous version of Weiss' representation theorem [30, Thm. 4.5].

THEOREM 3.11. *Let $\Sigma = (T, \Phi, \Psi, \mathbb{F})$ be a regular nonautonomous system, and let $C(s)$ and $C_\tau(s)$ be given by Definition 2.6 and (2.5). Then $\Phi(\cdot,s)u \in D(C(\cdot),s)$, and $\mathbb{F}_s u = C(\cdot)\Phi(\cdot,s)u$ for $s \geq 0$ and $u \in L^p_{loc}([s,\infty),U)$. Moreover, $C_\tau(\cdot)\Phi(\cdot,s)u \to \mathbb{F}_s u$ in $L^p_{loc}([s,\infty),Y)$ as $\tau \searrow 0$, and $\|C_\tau(\cdot)\Phi(\cdot,s)u\|_{L^p([s,s+t_0],Y)} \leq c \, \|u\|_{L^p([s,s+t_0],U)}$ for $\tau \in (0,1]$ and a constant $c = c(t_0)$ independent of u and s.*

Proof. Let $t \in [s,\infty)$ be a p-Lebesgue point of u and $\mathbb{F}_s u$ such that the regularity condition (3.12) holds at this point t. Setting $o_t(\sigma) = u(\sigma) - u(t)$ for $\sigma \geq t$, we have

$$\mathbb{F}_s u = \mathbb{F}_t \tilde{u}(t,\cdot) + \mathbb{F}_t o_t + \Psi_t \Phi_{t,s} u \quad \text{on} \quad [t,\infty) \quad \text{and} \tag{3.17}$$

$$\left\| \frac{1}{\tau} \int_t^{t+\tau} (\mathbb{F}_t o_t)(\sigma)\, d\sigma \right\|^p \leq \kappa^p \, \frac{1}{\tau} \int_t^{t+\tau} \|u(\sigma) - u(t)\|^p d\sigma. \tag{3.18}$$

Consequently, $C_\tau(t)\Phi_{t,s}u$ converges in Y to $(\mathbb{F}_s u)(t)$ as $\tau \to 0$ so that the first assertion holds. The estimate (3.18) and Fubini's theorem further yield

$$\begin{aligned} &\int_s^{s+t_0} \left\| \frac{1}{\tau} \int_t^{t+\tau} (\mathbb{F}_t o_t)(\sigma)\, d\sigma \right\|^p dt \leq \frac{\kappa^p}{\tau} \int_0^\tau \int_s^{s+t_0} \|u(t+\sigma) - u(t)\|^p dt \, d\sigma, \\ &\int_s^{s+t_0} \left\| \frac{1}{\tau} \int_t^{t+\tau} [(\mathbb{F}_s u)(\sigma) - (\mathbb{F}_s u)(t)]\, d\sigma \right\|^p dt \\ &\qquad\qquad \leq \frac{\kappa^p}{\tau} \int_0^\tau \int_s^{s+t_0} \|(\mathbb{F}_s u)(t+\sigma) - (\mathbb{F}_s u)(t)\|^p dt \, d\sigma, \end{aligned} \tag{3.19}$$

where both terms on the right-hand side converge to 0 as $\tau \to 0$. Combining these facts with (3.15) and (3.17), we establish that $\mathbb{F}_s u = \lim_\tau C_\tau(\cdot)\Phi(\cdot,s)u$ in $L^p_{loc}([s,\infty),Y)$. The asserted estimate follows in a similar way. ∎

The next approximation result complements Proposition 3.5 for absolutely regular systems. For technical reasons, we have to use the operators $B_n : L^p_{loc}(\mathbb{R}_+,U) \to L^\infty_{loc}(\mathbb{R}_+,X)$ rather than $B_n(t) : U \to X$. Observe that only regularity is used in the proof of estimate (3.20).

PROPOSITION 3.12. *Let Σ be an absolutely regular nonautonomous system, $p \in (1,\infty)$, and let $C(s)$ and $\Phi^n_{t,s}$ be given as in Definition 2.6 and (3.7). Then $\Phi^n(\cdot,s)u \in D(C(\cdot),s)$, $C(\cdot)\Phi^n(\cdot,s)u \to \mathbb{F}_s u$ in $L^p_{loc}([s,\infty),Y)$ as $n \to \infty$, and*

$$\|C(\cdot)\Phi^n(\cdot,s)u\|_{L^p([s,s+t_0],Y)} \leq 2\kappa(t_0) \, \|u\|_{L^p([s,s+t_0],U)} \tag{3.20}$$

for $u \in L^p_{loc}([s,\infty),U)$, $s \geq 0$, $n \in \mathbb{N}$, and $t_0 > 0$.

Proof. Due to Proposition 2.11, we have $\Phi^n(\cdot,s)u = \mathbb{K}_s B_n u \in D(C(\cdot),s)$. Formula (3.9), Proposition 2.11, and Theorem 3.11 further yield

$$\mathbb{F}_s u - C(\cdot)\Phi^n(\cdot,s)u = C(\cdot)r_n(\cdot;u) = \lim_{\tau\to 0} C_\tau(\cdot)r_n(\cdot;u) \quad (\text{in } L^p_{loc}([s,\infty),Y)). \tag{3.21}$$

Using Hölder's inequality, Fubini's theorem, and Theorem 3.11, we now derive

$$\begin{aligned}
\|C(\cdot)r_n(\cdot;u)\|^p_{L^p([s,s+t_0],Y)} &\le \lim_{\tau\to 0} n\int_s^{s+t_0}\int_{t-\frac{1}{n}}^{t} \|C_\tau(t)\Phi_{t,\sigma}u\|^p_Y \,d\sigma\,dt \\
&\le n\int_{s-\frac{1}{n}}^{s+t_0} \|\mathbb{F}_\sigma u\|^p_{L^p([\sigma,\sigma+1/n],Y)}\,d\sigma \\
&\le n\kappa(t_0)\int_0^{\frac{1}{n}}\int_{s-\frac{1}{n}}^{s+t_0} \|u(t+\sigma)\|^p\,d\sigma\,dt \le \kappa(t_0)\,\|u\|^p_p.
\end{aligned} \tag{3.22}$$

(Here we have considered a function $u\in L^p([s,s+t_0],U)$ and extended it by 0 to $\mathbb{R}$.) Thus (3.20) holds. The estimate (3.22) also gives

$$\begin{aligned}
&\|C(\cdot)r_n(\cdot;u)\|_{L^p([s,s+t_0],Y)} \\
&\le \left(\int_{s-\frac{1}{n}}^{s+t_0} n\,\|\mathbb{F}_\sigma o_\sigma\|^p_{L^p([\sigma,\sigma+\frac{1}{n}],Y)}\,d\sigma\right)^{\frac{1}{p}} + \left(\int_{s-\frac{1}{n}}^{s+t_0} n\,\|\mathbb{F}_\sigma \tilde{u}(\sigma,\cdot)\|^p_{L^p([\sigma,\sigma+\frac{1}{n}],Y)}\,d\sigma\right)^{\frac{1}{p}}.
\end{aligned}$$

The right-hand side tends to 0 as in (3.19) and (3.16). ∎

4. The main result and discussion. Let Σ be a regular nonautonomous system, $\Delta(\cdot)\in L^\infty(\mathbb{R}_+,\mathcal{L}_s(Y,U))$, and let $C(s)$ be given by Definition 2.6. For $x\in X$ and $s\ge 0$, we are looking for functions $x(\cdot)\in C([s,\infty),X)\cap D(C(\cdot),s)$ satisfying

$$x(t) = T(t,s)x + \Phi_{t,s}\Delta(\cdot)C(\cdot)x(\cdot), \qquad t\ge s, \tag{4.1}$$

or, if $\Phi(\cdot,s)u = \overline{\mathbb{K}}_s B(\cdot)u(\cdot)$ for admissible control operators $B(s)$,

$$x(t) = T(t,s)x + \int_s^t \to (t,\tau)B(\tau)\Delta(\tau)C(\tau)x(\tau)\,d\tau, \qquad t\ge s. \tag{4.2}$$

As shown by [27, Ex. 6], one cannot allow for every bounded feedback in (4.1) in general. (We note that this example gives rise to an absolutely regular autonomous system with $p=1$ and $\Delta = B = I$.) This fact motivates the next concept.

Definition 4.1. *Let $\Sigma = (T,\Phi,\Psi,\mathbb{F})$ be a well-posed nonautonomous system. We call $\Delta(\cdot)\in L^\infty(\mathbb{R}_+,\mathcal{L}_s(Y,U))$ an* admissible feedback *for Σ if there is $t_0>0$ such that $I-\mathbb{F}(s+t_0,s)\Delta(\cdot)$, $s\ge 0$, have uniformly bounded inverses on $L^p([s,s+t_0],Y)$.*

Of course, $\Delta(\cdot)$ is admissible if

$$\|\Delta(\cdot)\|_\infty < \left[\inf_{t_0>0}\sup_{s\ge 0}\|\mathbb{F}(s+t_0,s)\|\right]^{-1} =: q. \tag{4.3}$$

The right-hand side of this inequality equals ∞ if $B(t)$ and $C(t)$ are of "lower order"; see, e.g., [6] or [23, Ex. 4.11]. We point out that the invertibility of $I-\mathbb{F}(s+t_0,s)\Delta(\cdot)$ is in fact necessary for some properties of the feedback system as shown by Lemma 4.3 and Proposition 5.1. The next lemmas also indicate that our notion of admissibility is

quite flexible; see [21, Lem. 4.1], [25, section 7.1], and [32, section 3] for autonomous analogues.

LEMMA 4.2. *Let $\Delta(\cdot) \in L^\infty(\mathbb{R}_+, \mathcal{L}_s(Y,U))$ and Σ be a well-posed nonautonomous system. If $I - \mathbb{F}(s+t_0,s)\Delta(\cdot)$ is invertible on $L^p([s,s+t_0],Y)$ for all $s \geq 0$ and some $t_0 > 0$, then $I - \mathbb{F}(s+t_1,s)\Delta(\cdot)$ is invertible on $L^p([s,s+t_1],U)$ for all $s \geq 0$ and $t_1 > 0$. The notion of an admissible feedback is independent of $t_0 > 0$.*

Proof. We assume that $U = Y$ and $\Delta(s) = I$ for simplicity. First, let $t_1 \leq t_0$. Extend a given $v \in L^p([s,s+t_1],Y)$ by 0 to $\tilde{v} \in L^p([s,s+t_0],Y)$, and set $\tilde{u} = (I - \mathbb{F}(s+t_0,s))^{-1}\tilde{v}$. Then $(I - \mathbb{F}(s+t_1,s))u = v$ for the restriction u of $\tilde{u}$. If $u = \mathbb{F}(s+t_1,s)u$, then there is a function $u_1 \in L^p([s+t_1,s+t_0],U)$ such that $(I - \mathbb{F}(s+t_0,s+t_1))u_1 = \Psi_{s+t_1}\Phi_{s+t_1,s}u$. Set $\tilde{u} = u$ on $[s,s+t_1]$ and $\tilde{u} = u_1$ on $[s+t_1,s+t_0]$. Hence $\tilde{u} = \mathbb{F}(s+t_0,s)\tilde{u}$ so that $u = 0$.

It remains to consider $t_1 = nt_0$ for $n \in \mathbb{N}$. Proceeding by induction, we assume that the assertion is true for $t_1 = nt_0$. It is then clear that $I - \mathbb{F}(s+(n+1)t_0,s)$ is injective. For $v \in L^p([s,s+(n+1)t_0],Y)$, we set $u_1 = (I - \mathbb{F}(s+nt_0,s))^{-1}(v|[s,s+nt_0])$ and $u_2 = (I - \mathbb{F}(s+(n+1)t_0,s+nt_0))^{-1}\{v|[s+nt_0,s+(n+1)t_0] + \Psi_{s+nt_0}\Phi_{s+nt_0,s}u_1\}$. Putting u_1 and u_2 together, one sees that $I - \mathbb{F}(s+(n+1)t_0,s)$ is surjective. ∎

LEMMA 4.3. *For maps $T : E \to F$ and $V : F \to E$, the following are equivalent:*

1. *$I - VT$ is bijective on E.*
2. *$I - TV$ is bijective on F.*
3. *There is a map $S : E \to F$ such that $S - T = TVS = SVT$.*

Then we have $(I-TV)^{-1} = I + T(I-VT)^{-1}V = I + SV$ and $(I-VT)^{-1} = I + VS$, and S in (3) is uniquely given by $S = (I-TV)^{-1}T = T(I-VT)^{-1}$.

Thus a feedback $\Delta(\cdot)$ is admissible if and only if $I - \Delta(\cdot)\mathbb{F}(s+t_0,s)$, $s \geq 0$, have uniformly bounded inverses for some/all $t_0 > 0$ if and only if for some/all $t_0 > 0$ there are uniformly bounded operators $\mathbb{F}^\Delta(s+t_0,s)$, $s \geq 0$, such that $\mathbb{F}^\Delta(s+t_0,s) - \mathbb{F}(s+t_0,s) = \mathbb{F}^\Delta(s+t_0,s)\Delta(\cdot)\mathbb{F}(s+t_0,s) = \mathbb{F}(s+t_0,s)\Delta(\cdot)\mathbb{F}^\Delta(s+t_0,s)$.

We now solve (4.1) by constructing an evolution family T_Δ on X and show that the feedback system Σ^Δ is again absolutely regular if the unperturbed system is absolutely regular. Proposition 5.1 describes the relations between Σ and Σ^Δ in greater detail.

THEOREM 4.4. *Let $\Sigma = (T,\Phi,\Psi,\mathbb{F})$ be a regular nonautonomous system and $\Delta(\cdot) \in L^\infty(\mathbb{R}_+, \mathcal{L}_s(Y,U))$ be an admissible feedback. Then the following hold:*

(a) *There is an evolution family T_Δ on X such that $T_\Delta(\cdot,s)x \in D(C(\cdot),s)$,*

$$\|C(\cdot)T_\Delta(\cdot,s)x\|_{L^p([s,s+t_0],Y)} \leq \gamma' \|x\|, \tag{4.4}$$

$x(\cdot) = T_\Delta(\cdot,s)x$ is the unique solution of (4.1), and

$$T_\Delta(t,s)x = T(t,s)x + \Phi_{t,s}\Delta(\cdot)C(\cdot)T_\Delta(\cdot,s)x \tag{4.5}$$

for $t \geq s \geq 0$, $x \in X$, and a constant γ'. If, in addition, $\Phi(\cdot,s)u = \overline{\mathbb{K}}_s B(\cdot)u(\cdot)$ for T-admissible control operators $B(t)$, then

$$T_\Delta(t,s)x = T(t,s)x + \int_s^t \to (t,\tau)B(\tau)\Delta(\tau)C(\tau)T_\Delta(\tau,s)x\,d\tau. \tag{4.6}$$

(b) *If the system is absolutely regular and $p \in (1,\infty)$, then*

$$T_\Delta(t,s)x = T(t,s)x + \lim_{n\to\infty} \int_s^t T_\Delta(t,\tau)[B_n(\Delta(\cdot)\Psi_s x)](\tau)\,d\tau \tag{4.7}$$

for $t \geq s \geq 0$ and $x \in X$, where the limit is taken in X and is locally uniform in t. Moreover, $\Sigma^\Delta = (T_\Delta, \Phi^\Delta, \Psi^\Delta, \mathbb{F}^\Delta)$ is an absolutely regular system, where we set

$$\Psi_s^\Delta x = C(\cdot)T_\Delta(\cdot, s)x, \quad \Phi_{t,s}^\Delta u = \lim_{n\to\infty} [\mathbb{K}_s^\Delta B_n u](t),$$
$$\mathbb{F}_s^\Delta u = \lim_{n\to\infty} C(\cdot)\mathbb{K}_s^\Delta B_n u, \quad \mathbb{K}_s^\Delta f\,(t) = \int_s^t T_\Delta(t,\tau)f(\tau)\,d\tau$$

for $t \geq s \geq 0$, $x \in X$, $u \in L^p_{loc}([s,\infty), U)$, and $f \in L^p_{loc}([s,\infty), X)$, where the limits are taken in X and L^p_{loc}, respectively.

Proof. (a) We first prove the uniqueness of solutions to (4.1). If v solves (4.1) with $x = 0$, then $C(\cdot)v = \mathbb{F}_s\Delta(\cdot)C(\cdot)v$ by Theorem 3.11. Since $I - \mathbb{F}(s+t_1, s)\Delta(\cdot)$ is injective, $C(\cdot)v$ has to vanish a.e. on $[s, s+t_1]$, where $t_1 > 0$ can be chosen arbitrarily large by Lemma 4.2. Hence (4.1) implies that $v = 0$. To solve (4.1), we define

$$T_\Delta(t,s)x = T(t,s)x + \Phi_{t,s}\Delta(\cdot)(I - \mathbb{F}(s+t_1,s)\Delta(\cdot))^{-1}\Psi_s x \tag{4.8}$$

for $0 \leq t - s \leq t_1$ and $x \in X$. Clearly, $T_\Delta(t,s)$ is an exponentially bounded linear operator on X, and $T_\Delta(\cdot,s)x$ is continuous in X by Proposition 3.5. Theorems 2.7 and 3.11 further show that $T_\Delta(\cdot,s)x \in D(C(\cdot),s)$ and

$$C(\cdot)T_\Delta(\cdot,s)x = \Psi_s x + \mathbb{F}_s\Delta(\cdot)(I - \mathbb{F}(s+t_1,s)\Delta(\cdot))^{-1}\Psi_s x \tag{4.9}$$
$$= (I - \mathbb{F}(s+t_1,s)\Delta(\cdot))^{-1}\Psi_s x. \tag{4.10}$$

Hence (4.4) holds. Inserting (4.10) into (4.8), we obtain (4.5) and (4.6) and thus have solved (4.1). One verifies (E1) for T_Δ using the uniqueness of (4.1), formula (3.1), and a standard argument. It remains to establish the strong continuity of T_Δ. We first take $(t_n, s_n) \to (s_0, s_0)$ with $t_n \geq s_n \geq 0$. For $\varepsilon > 0$, $x \in X$, and large n, there is $r \in [0, s_0] \cap [0, s_n]$ such that $\|T(s_0, r)x - x\| \leq \varepsilon$. Then (4.8) and (2.1) yield

$$\begin{aligned}\|T_\Delta(t_n,s_n)x - x\| &\leq \|T_\Delta(t_n,s_n)(x - T(s_n,r)x)\| + \|T_\Delta(t_n,s_n)T(s_n,r)x - x\| \\ &\leq c\,\|x - T(s_n,r)x\| + \|T(t_n,r)x - x\| + c\left[\int_{s_n}^{t_n} \|[\Psi_r x](\sigma)\|^p d\sigma\right]^{\frac{1}{p}},\end{aligned}$$

$\overline{\lim_{n\to\infty}} \|T_\Delta(t_n,s_n)x - x\| \leq (c+1)\varepsilon$

for a constant c. Therefore, T_Δ is strongly continuous at (s_0, s_0). If $(t_n, s_n) \to (t_0, s_0)$ for some $t_0 > s_0$, we may assume that $t_n > s_n$ and $t_n > s_0$. We take $t_n \geq r_n \geq s_n \vee s_0$ with $r_n \to s_0$ and derive (E2) from the above results and the expression

$$\begin{aligned}T_\Delta(t_n,s_n)x - T_\Delta(t_0,s_0)x = {} & T_\Delta(t_n,r_n)(T_\Delta(r_n,s_n)x - T_\Delta(r_n,s_0)x) \\ & + T_\Delta(t_n,s_0)x - T_\Delta(t_0,s_0)x.\end{aligned}$$

(b) Define $\mathcal{D}_{\Delta,s}$ as in (2.6) using T_Δ. Then (4.5) and Proposition 3.5 imply that

$$\mathbb{K}_s^\Delta f\,(t) = \mathbb{K}_s f\,(t) + \lim_{n\to\infty} \int_s^t \int_\tau^t T(t,\sigma)B_n(\sigma)\Delta(\sigma)C(\sigma)T_\Delta(\sigma,\tau)f(\tau)\,d\sigma\,d\tau$$

for $f \in \mathcal{D}_{\Delta,s}$ and $s \geq 0$ since the integrand is the sum of functions of the form

$$(\tau,\sigma) \mapsto \gamma(\tau)\,T(t,\sigma)B_n(\sigma)\Delta(\sigma)C(\sigma)T_\Delta(\sigma,r)x. \tag{4.11}$$

For the same reason, $\mathbb{K}_s^\Delta f$ belongs to $D(C(\cdot),s)$, and we can apply Fubini's theorem and take $T(t,\sigma)B_n(\sigma)\Delta(\sigma)C(\sigma) \in \mathcal{L}(\underline{X}_\sigma, X)$ out of the inner integral. So we obtain

$$\mathbb{K}_s^\Delta f = \mathbb{K}_s f + \lim_{n\to\infty} \mathbb{K}_s B_n(\cdot)\Delta(\cdot)C(\cdot)\mathbb{K}_s^\Delta f = \mathbb{K}_s f + \Phi(\cdot,s)\Delta(\cdot)C(\cdot)\mathbb{K}_s^\Delta f \tag{4.12}$$

for $f \in \mathcal{D}_{\Delta,s}$ due to Proposition 3.5. Theorem 3.11 now shows that $C(\cdot)\mathbb{K}_s^\Delta f = C(\cdot)\mathbb{K}_s f + \mathbb{F}_s\Delta(\cdot)C(\cdot)\mathbb{K}_s^\Delta f$. Hence

$$C(\cdot)\mathbb{K}_s^\Delta f = (I - \mathbb{F}(s+t_1,s)\Delta(\cdot))^{-1}C(\cdot)\mathbb{K}_s f \tag{4.13}$$

on $[s, s+t_1]$ for each $t_1 > 0$. Inserting (4.13) into (4.12), we conclude that

$$\mathbb{K}_s^\Delta f = \mathbb{K}_s f + \Phi(\cdot,s)\Delta(\cdot)(I - \mathbb{F}(s+t_1,s)\Delta(\cdot))^{-1}C(\cdot)\mathbb{K}_s f \tag{4.14}$$

on $[s, s+t_1]$ for $f \in \mathcal{D}_{\Delta,s}$. This identity holds for all $f \in L^p_{loc}([s,\infty),X)$ by Proposition 2.11. So we may take $f = B_n u$ for $u \in L^p_{loc}([s,\infty),U)$ and $n \in \mathbb{N}$, and thus

$$\mathbb{K}_s^\Delta B_n u = \mathbb{K}_s B_n u + \Phi(\cdot,s)\Delta(\cdot)(I - \mathbb{F}(s+t_1,s)\Delta(\cdot))^{-1}C(\cdot)\mathbb{K}_s B_n u. \tag{4.15}$$

As a consequence, $\mathbb{K}_s^\Delta B_n u \in D(C(\cdot),s)$ and

$$C(\cdot)\mathbb{K}_s^\Delta B_n u = (I - \mathbb{F}(s+t_1,s)\Delta(\cdot))^{-1}C(\cdot)\mathbb{K}_s B_n u \tag{4.16}$$

by Proposition 2.11 and Theorem 3.11. In view of Propositions 3.5 and 3.12, we can take the limit as $n \to \infty$ in the formulas (4.15) and (4.16) (in $C([s,s+t_1],X)$ and $L^p([s,s+t_1],Y)$, respectively). It is then easy to see that $\Sigma^\Delta = (T_\Delta, \Phi^\Delta, \Psi^\Delta, \mathbb{F}^\Delta)$ defined in the assertion is a well-posed nonautonomous system. Equation (4.16) and Proposition 3.12 further yield

$$\int_t^{t+\tau} \|(\mathbb{F}_t^\Delta u_z)(\sigma)\|^p \, d\sigma \le c \int_t^{t+\tau} \|(\mathbb{F}_t u_z)(\sigma)\|^p \, d\sigma$$

for $t \ge 0$, $\tau > 0$, $u_z \equiv z$, and $z \in U$ so that additionally Σ^Δ is absolutely regular.

We now choose $u = \Delta(\cdot)\Psi_s x$ for $x \in X$ and deduce from (4.15), Propositions 3.5 and 3.12, and (4.8) that

$$\begin{aligned}\lim_{n\to\infty} \mathbb{K}_s^\Delta B_n \Delta(\cdot)\Psi_s x &= \Phi(\cdot,s)\Delta(\cdot)\Psi_s x + \Phi(\cdot,s)\Delta(\cdot)(I - \mathbb{F}(s+t_1,s)\Delta(\cdot))^{-1}\mathbb{F}_s\Delta(\cdot)\Psi_s x \\ &= \Phi(\cdot,s)\Delta(\cdot)(I - \mathbb{F}(s+t_1,s)\Delta(\cdot))^{-1}\Psi_s x = T_\Delta(\cdot,s)x - T(\cdot,s)x,\end{aligned}$$

where the limit is taken in X and is locally uniform in t. Thus (4.7) holds. ∎

We state several variants of Theorem 4.4 and compare them to related results.

REMARK 4.5. *The above theorem remains valid if we do not assume* (E3), *require* β, γ, κ *and* $\|(I - \mathbb{F}(s+t_0,s)\Delta(\cdot))^{-1}\|$ *only to be uniform with respect to* $s \in [0,a]$ *for every* $a > 0$, *and assert for the perturbed problem only the analogous properties. The proof of part* (a) *also works in the case that* T *(and then* T_Δ*) is only strongly continuous in* t *and* s *separately. Part* (a) *can be verified for admissible systems, too, if one considers only* $x \in D(C(s))$ *in* (4.2), (4.4), *and* (4.6)*; see* [23, *Thm.* 4.18] *or* (c) *below.*

REMARK 4.6. *Let* Σ *be a nonautonomous regular system with* $p \in (1,\infty)$. *It can be shown that* $\langle C(\cdot)\mathbb{K}_s B_n u, v\rangle \to \langle \mathbb{F}_s u, v\rangle$ *as* $n \to \infty$ *for all* $v \in L^q([s,s+t_0],Y^*)$, $\frac{1}{p} + \frac{1}{q} = 1$. *Thus, if* Y *is reflexive, the conclusions of Theorem* 4.4(b) *hold for merely*

regular systems except that the limits exist only weakly and that it is not clear whether Σ^Δ *is regular again. In the autonomous case, the regularity of* Σ^Δ *for regular* Σ *was established in* [32, *Thm.* 4.5, 4.7]*, but the proof given there relies on Laplace transforms and a Tauberian theorem* [31, *Thm.* 5.2] *not available here; see also* [25, *section* 7.5].

(a) *Perturbation theory of evolution equations.* Theorem 4.4 is a joint nonautonomous extension of the Desch–Schappacher and Miyadera perturbation theorem from semigroup theory (see, e.g., [9, section III.3]): First, let $B(t)$ be T-admissible control operators, and define $Y = X$, $\Psi_s = T(\cdot, s)$, and $\mathbb{F}_s = \overline{\mathbb{K}}_s B(\cdot)$; i.e., $C(t) \equiv I$. This gives an absolutely regular nonautonomous system with $\kappa(t_0) = \beta t_0^{1/p}$ so that $q = \infty$ in (4.3). Second, let (T, Ψ) be a nonautonomous observation system for $p \in (1, \infty)$ represented by $C(t)$. Setting $U = X$, $\Phi_{t,s} u = (\mathbb{K}_s u)(t)$, and $\mathbb{F}_s = C(\cdot)\mathbb{K}_s$, i.e., $B(t) \equiv I$, we obtain a well-posed nonautonomous system thanks to Proposition 2.11. Approximating u_z by $T(\cdot, t)z$, one verifies that the system is absolutely regular. A nonautonomous Miyadera theorem for closable perturbations $C(t)$ and $p \geq 1$ was proved in [19] by other methods.

(b) *Autonomous controlled systems.* Let $T(t - s) = T(t, s)$ be a C_0-semigroup generated by A and $\Delta(t) \equiv \Delta$. We say that (T, B, C) belongs to the *Pritchard–Salamon class* [18] if (2.2) holds with $\|x\|_X$ replaced by $\|x\|_{\overline{X}}$ and (3.6) holds with $\|\cdot\|_X$ replaced by $\|\cdot\|_{\underline{X}}$. The perturbation theory for this class was developed in detail in [6]. In this case, one can extend $T_\Delta(t)$ to $\overline{X}$, and the number q in (4.3) is equal to ∞.

Weiss introduced autonomous regular systems in [28], [29], [30], [31] similarly as in the above definitions by considering only the initial time $s = 0$. He solved the feedback problem in [32, Thm. 6.1] for a well-posed system with $p = 2$ on Hilbert spaces X, Y, U, allowing for nontrivial feedthrough D and assuming that (roughly speaking) $I - CR(\lambda, A_{-1})B\Delta$ is invertible on a right halfplane; see also [21]. If the system is regular, the feedback system is again regular and can be represented almost in the natural way; see [32, section 7]. The feedback theory for several classes of (non)regular systems is exhaustively studied in Chapter 7 of Staffans' monograph [25] in a Banach space setting and also for $p = 1, \infty$.

The remaining difficulties come from the fact that, in general, $T_\Delta(t)$ cannot be continuously extended to the extrapolation space X_{-1}^A corresponding to T (see [23, Ex. 4.20]); in particular, the extrapolation space X_{-1}^Δ of T_Δ may differ from X_{-1}^A. Weiss constructed subspaces W and W_Δ of X_{-1}^A and X_{-1}^Δ, respectively, such that $Jx := \lim_{\lambda\to\infty} \lambda R(\lambda, A_{-1})x$ (in X_{-1}^Δ) defines an isomorphism $J : W \to W_\Delta$; see [32, Thm. 7.7]. Note that $Jx = x$ for $x \in X$. Then

$$T_\Delta(t)x = T(t)x + \int_0^t T_{\Delta,-1}(t - \tau)JB\Delta CT(\tau)x \, d\tau$$

by (6.11), (6.1), and [32, p. 55]. In other words, Weiss managed to put the limit in (4.7) inside the integral using a different regularization. Identifying B and JB, he represented the feedback system in terms of B and C and computed the generator of T_Δ [32, section 7]; see [25, section 7.4] for a somewhat different approach.

(c) *Nonautonomous controlled systems.* Part (a) of Theorem 4.4 was proved by Hinrichsen, Jacob, and Pritchard for nonautonomous admissible systems in a slightly differing setting; see [10, Thm. 3.2] and [12], [14] also for nonlinear feedback. They work with separately strongly continuous evolution families and have some additional technical assumptions (see, e.g., Hypotheses 4 and 7 of [10]). Moreover, they obtain (4.6) with a pointwise representation of $C(\cdot)T(\cdot, s)x$ only for $x \in D(C(s))$. The

issues investigated in Theorem 4.4(b) were not considered in [10] and [14] and were considered in [12, Thm. 3.4.7] only for systems of Pritchard–Salamon type.

REMARK 4.7. *In addition to the assumptions of Theorem* 4.4(b), *we suppose that* $\Phi_{t,s}$ *is given by admissible observation operators* $B(\cdot) \in L^\infty(\mathbb{R}_+, \mathcal{L}_s(U, \overline{X}))$ *for* $\overline{X}_t \equiv \overline{X}$ *and that* $T_\Delta(t,s)$ *has a locally uniformly bounded extension* $\overline{T}_\Delta(t,s) : \overline{X} \to \overline{X}$. *Thus* $\overline{T}_\Delta$ *satisfies* (E1) *and* (E2). *We set* $(\overline{\mathbb{K}}_s^\Delta f)(t) = \int_s^t \overline{T}_\Delta(t,\tau) f(\tau)\, d\tau$ *for* $t \ge s \ge 0$ *and* $f \in L^1_{loc}([s,\infty), \overline{X})$. *Then* $\Phi^\Delta(\cdot,s)u = \overline{\mathbb{K}}_s^\Delta B(\cdot)u$, $\mathbb{F}_s^\Delta u = C(\cdot)\overline{\mathbb{K}}_s^\Delta B(\cdot)u$, *and*

$$T_\Delta(t,s)x = T(t,s)x + \int_s^t \to_\Delta (t,\tau)B(\tau)\Delta(\tau)C(\tau)T(\tau,s)x\, d\tau \tag{4.17}$$

for $u \in L^p_{loc}([s,\infty), U)$, $x \in X$, *and* $t \ge s \ge 0$.

Proof. Observe that $B_n u \to B(\cdot)u$ as $n \to \infty$ in $L^p_{loc}(\mathbb{R}_+, \overline{X})$ for $u \in L^p_{loc}(\mathbb{R}_+, \overline{X})$ because of the inequality

$$\|B_n u - B(\cdot)u\|^p_{L^p([0,t_0],\overline{X})} \le n \int_0^{\frac{1}{n}} \int_0^{t_0} \|\overline{T}(\tau+\sigma,\tau)B(\tau)u(\tau) - B(\tau+\sigma)u(\tau+\sigma)\|^p_{\overline{X}}\, d\tau\, d\sigma,$$

which is a consequence of Hölder's inequality and Fubini's theorem. Thus $\Phi^\Delta(\cdot,s)u = \overline{\mathbb{K}}_s^\Delta B(\cdot)u$, and (4.17) holds. The identities (4.15) and (4.16) then imply that

$$C(\cdot)\overline{\mathbb{K}}_s^\Delta B(\cdot)u = (I - \mathbb{F}(s+t_1,s)\Delta(\cdot))^{-1}\mathbb{F}_s u = \mathbb{F}_s^\Delta u. \qquad \square$$

The above remark and paragraph (b) indicate that an (absolutely) regular nonautonomous system and the corresponding feedback system can be represented by operators $B(t)$ (and not just approximately by $B_n(t)$) whenever we have a decent extrapolation theory for the given problem. It seems to be reasonable to study first the case that T is generated by operators $A(t)$ and consider spaces $\overline{X}_t$ related to $A(t)$. For various results on parabolic evolution equations and extrapolation spaces, we refer to [1], [2, Chap.V], [23, Prop. 2.1].

5. Further properties of the feedback system. In the setting of Theorem 4.4, we study the relationship between the open- and the closed-loop systems in more detail; see [25, Chap. 7] and [32, section 6] for similar results in the autonomous case. To put the formulas in a concise form, we define $\Psi(t,s)x = \mathbb{1}_{[s,t]}\Psi_s x$ and

$$\Sigma(t,s) = \begin{pmatrix} T(t,s) & \Phi(t,s) \\ \Psi(t,s) & \mathbb{F}(t,s) \end{pmatrix} : X \times L^p([s,t],U) \to X \times L^p([s,t],Y), \quad t \ge s \ge 0.$$

PROPOSITION 5.1. *Let* Σ *be an absolutely regular nonautonomous system, let* $p \in (1,\infty)$, *let* $\Delta(\cdot)$ *be an admissible feedback for* Σ, *and let* Σ^Δ *be the feedback system from Theorem* 4.4. *Then*

$$\mathbb{F}_s^\Delta = (I - \mathbb{F}_s\Delta(\cdot))^{-1}\mathbb{F}_s = \mathbb{F}_s(I - \Delta(\cdot)\mathbb{F}_s)^{-1} = C(\cdot)\Phi^\Delta(\cdot,s), \tag{5.1}$$

$$\Sigma^\Delta(t,s) - \Sigma(t,s) = \Sigma(t,s)\begin{pmatrix} 0 & 0 \\ 0 & \Delta(\cdot) \end{pmatrix}\Sigma^\Delta(t,s) = \Sigma^\Delta(t,s)\begin{pmatrix} 0 & 0 \\ 0 & \Delta(\cdot) \end{pmatrix}\Sigma(t,s). \tag{5.2}$$

Proof. The first equality in (5.1) is an immediate consequence of (4.16) and Proposition 3.12. Lemma 4.3 then yields the second equality in (5.1) and the expressions for $\mathbb{F}^\Delta - \mathbb{F}$ in the lower right corner of (5.2). Taking the limit in (4.15) and using

the formulas for $\mathbb{F}^\Delta$, we deduce the last equality in (5.1). The identities for $T^\Delta - T$ in the upper left corner of (5.2) were established in Theorem 4.4, and they imply the formulas for $\Psi^\Delta - \Psi$ in the lower left corner in (5.2). The first equality in the upper right corner follows from (4.15). Employing the previous results, we finally obtain

$$\begin{aligned}\Phi^\Delta(\cdot,s)\Delta(\cdot)\mathbb{F}_s &= \Phi(\cdot,s)\Delta(\cdot)\mathbb{F}_s + \Phi(\cdot,s)\Delta(\cdot)\mathbb{F}_s^\Delta\Delta(\cdot)\mathbb{F}_s \\ &= \Phi(\cdot,s)\Delta(\cdot)\mathbb{F}_s^\Delta = \Phi^\Delta(\cdot,s) - \Phi(\cdot,s). \qquad \square\end{aligned}$$

The above result allows us to prove that the following control theoretic properties (cf. [7]) remain unchanged under feedback.

DEFINITION 5.2. (a) *A nonautonomous control system* (T,Φ) *is called* exactly (approximately) controllable on $[s,t]$ *if* $\Phi(t,s)$ *is surjective (has dense range) and it is called* exactly (approximately) null controllable on $[s,t]$ *if* $T(t,s)X$ *is contained in the (closure of)* $\Phi(t,s)L^p([s,t],U)$.

(b) *A nonautonomous observation system* (T,Ψ) *is called* (continuously) initially observable on $[s,t]$ *if* $\Psi(t,s)$ *is injective (bounded from below) and* (continuously) finally observable on $[s,t]$ *if* $\ker\Psi(t,s) \subset \ker T(t,s)$ *(if* $\|T(t,s)x\| \le c\,\|\Psi(t,s)x\|_p$ *for a constant* $c>0$ *and* $x \in X$*).*

PROPOSITION 5.3. *Let* Σ *be an absolutely regular nonautonomous system, let* $p \in (1,\infty)$*, let* $\Delta(\cdot)$ *be an admissible feedback for* Σ*, and let* Σ^Δ *be the corresponding feedback system. Then* Σ *possesses one of the properties in Definition* 5.2 *if and only if* Σ^Δ *has the same property.*

Proof. (1) The assertions concerning exact (approximate) controllability and (continuous) initial observability follow from the formulas

$$\begin{aligned}\Phi^\Delta(t,s) &= \Phi(t,s)(I+\Delta(\cdot)\mathbb{F}_s^\Delta), & \Phi(t,s) &= \Phi^\Delta(t,s)(I-\Delta(\cdot)\mathbb{F}_s),\\ \Psi^\Delta(t,s) &= (I+\mathbb{F}_s^\Delta\Delta(\cdot))\Psi(t,s), & \Psi(t,s) &= (I-\mathbb{F}_s\Delta(\cdot))\Psi^\Delta(t,s),\end{aligned}$$

which are immediate consequences of (5.2).

(2) Assume that Σ is null controllable. For $x \in X$, there is $u \in L^p([s,t],U)$ such that $T(t,s)x = \Phi_{t,s}u$. Thus (5.2) yields $T_\Delta(t,s)x = \Phi^\Delta(t,s)[u - \Delta(\cdot)\mathbb{F}_s u + \Delta(\cdot)\Psi_s x]$, and Σ^Δ is null controllable. The converse implication and the equivalence for approximate null controllability are shown in the same way.

(3) Assume that Σ is continuously finally observable. Using (5.2), we estimate

$$\begin{aligned}\|T^\Delta(t,s)x\| &\le \|T(t,s)x\| + \|\Phi(t,s)\Delta(\cdot)\Psi^\Delta(t,s)x\| \\ &\le c\,\|\Psi(t,s)x\|_p + c_1\,\|\Psi^\Delta(t,s)x\|_p \\ &\le (c+c_1)\,\|\Psi^\Delta(t,s)x\|_p + \|\mathbb{F}_s\Delta(\cdot)\Psi^\Delta(t,s)x\|_p \le c_2\,\|\Psi^\Delta(t,s)x\|_p\end{aligned} \tag{5.3}$$

so that Σ^Δ is continuously finally observable. If Σ is finally observable and $\Psi^\Delta(t,s)x = 0$, then $\Psi(t,s)x = -\mathbb{F}_s\Delta(\cdot)\Psi^\Delta(t,s)x = 0$. Hence $T(t,s)x = 0$, and (5.3) yields $T^\Delta(t,s)x = 0$. The converse implications are proved similarly. $\square$

Theorem 4.4 also guarantees that repeated feedbacks behave nicely.

PROPOSITION 5.4. *Let* Σ *be an absolutely regular nonautonomous system with* $p \in (1,\infty)$*, let* $\Delta(\cdot)$ *be an admissible feedback for* Σ*, let* Σ^Δ *be the corresponding feedback system, and let* $\tilde\Delta(\cdot) \in L^\infty(\mathbb{R}_+,\mathcal{L}_s(Y,U))$*. Then* $\tilde\Delta(\cdot)$ *is admissible for* Σ^Δ *if and only if* $\Delta(\cdot)+\tilde\Delta(\cdot)$ *is admissible for* Σ*. If this is the case, then* $\Sigma^{\Delta+\tilde\Delta} = (\Sigma^\Delta)^{\tilde\Delta}$.

Proof. Proposition 5.1 implies that

$$\begin{aligned}&\mathbb{F}_s\,[I-\tilde\Delta(\cdot)\mathbb{F}_s^\Delta] = [I-\mathbb{F}_s(\Delta(\cdot)+\tilde\Delta(\cdot))]\,\mathbb{F}_s^\Delta \quad \text{and}\\ &[I-\mathbb{F}_s^\Delta\tilde\Delta(\cdot)]\,\mathbb{F}_s = \mathbb{F}_s^\Delta\,[I-(\Delta(\cdot)+\tilde\Delta(\cdot))\mathbb{F}_s].\end{aligned} \tag{5.4}$$

Assume that $\Delta(\cdot) + \tilde{\Delta}(\cdot)$ is admissible for Σ. We then deduce from (5.4) and (5.1) that

$$\mathbb{F}_s^{\Delta+\tilde{\Delta}} - \mathbb{F}_s^{\Delta} = \mathbb{F}_s^{\Delta+\tilde{\Delta}}\tilde{\Delta}(\cdot)\mathbb{F}_s^{\Delta} = \mathbb{F}_s^{\Delta}\tilde{\Delta}(\cdot)\mathbb{F}_s^{\Delta+\tilde{\Delta}}$$

so that $\tilde{\Delta}(\cdot)$ is admissible for Σ^{Δ} by Lemma 4.3. The converse implication is proved in the same way. The second claim follows similarly from (5.2) and Lemma 4.3. ∎

We introduce a basic asymptotic property of evolution equations; see, e.g., [4], [9].

DEFINITION 5.5. *An evolution family T has an* exponential dichotomy *(or is called* hyperbolic*) if there are projections $P(t)$, $t \geq 0$, and constants $N, \delta > 0$ such that $P(\cdot) \in C_b(\mathbb{R}_+, \mathcal{L}_s(X))$ and, for $t \geq s \geq 0$ and $Q(t) = I - P(t)$,*

1. *$T(t,s)P(s) = P(t)T(t,s)$,*
2. *the restriction $T_Q(t,s) : Q(s)X \to Q(t)X$ of $T(t,s)$ has the inverse $T_Q(s,t)$,*
3. *$\|T(t,s)P(s)\| \leq Ne^{-\delta(t-s)}$, and $\|T_Q(s,t)Q(t)\| \leq Ne^{-\delta(t-s)}$.*

If $P(t) \equiv I$, then T is called exponentially stable.

Persistence of dichotomy under perturbations mapping from spaces $\underline{X}_t$ into X has been studied intensively; see [4, section 5.2], [22, section 5], and the references given there. The next result also holds for admissible systems; see [23, Thm. 4.23].

THEOREM 5.6. *Assume that $(T, \Phi, \Psi, \mathbb{F})$ is a regular nonautonomous system and that $\Delta(\cdot) \in L^{\infty}(\mathbb{R}_+, \mathcal{L}_s(Y,U))$ is Σ-admissible with $k(t_0) := \sup_s \|(I - \mathbb{F}(s + t_0, s)\Delta(\cdot))^{-1}\|$. Suppose that T has an exponential dichotomy with constants $N, \delta > 0$ and projections $P(t)$. Then there is a number $\varepsilon_0 = \varepsilon_0(N, \delta, t_0) > 0$ such that*

$$k(t_0)\beta(t_0)\gamma(t_0)\,\|\Delta(\cdot)\|_{\infty} \leq \varepsilon_0$$

implies that T_{Δ} is hyperbolic with projections having the same rank as $P(t)$ and $Q(t)$.

Proof. We extend the evolution families T and T_{Δ} to the time interval $\mathbb{R}$ by setting $T(t,s) = T_{\Delta}(t,s) = \exp[(t-s)\delta(Q(0) - P(0))]$ for $0 \geq t \geq s$ and $T_{(\Delta)}(t,s) = T_{(\Delta)}(t,0)\exp[-s\delta(Q(0) - P(0))]$. Observe that we can take $k(t_1) = k(t_0)$ for $0 < t_1 \leq t_0$ by the proof of Lemma 4.2 and that $\exp[t\delta(Q(0) - P(0))] = e^{-\delta t}P(0) + e^{\delta t}Q(0)$. Therefore, (4.8) yields

$$\|T_{\Delta}(s + t_0, s) - T(s + t_0, s)\| \leq (1 + e^{\delta t_0})Nk(t_0)\beta(t_0)\gamma(t_0)\,\|\Delta(\cdot)\|_{\infty}.$$

The assertion then follows from [24, Prop. 2.3] (see also [4, Thm. 5.23] and the references therein), where $\varepsilon_0 := (1 - e^{\delta t_0})^2\,((1 + e^{\delta t_0})8N^3)^{-1}$. ∎

We finally characterize the exponential stability of T from the perspective of control theory using the following notions; cf. [5], [6], [17], [20], [25, section 8.2].

DEFINITION 5.7. *A nonautonomous control system (T, Φ) is called* stabilizable *if there exists an observation system (T_F, Ψ^F) with an exponentially stable evolution family T_F on X such that $T_F(t,s)x = T(t,s)x + \Phi_{t,s}\Psi_s^F x$ for all $x \in X$ and $t \geq s \geq 0$.*

DEFINITION 5.8. *A nonautonomous observation system (T, Ψ) is called* detectable *if there is a control system (T_K, Φ^K) with an exponentially stable evolution family T_K on X such that $T_K(t,s)x = T(t,s)x + \Phi_{t,s}^K\Psi_s x$ for all $x \in X$ and $t \geq s \geq 0$.*

The following theorem relates the exponential stability of T, i.e., *internal stability*, with the boundedness of $\mathbb{F} : L^p(\mathbb{R}_+, U) \to L^p(\mathbb{R}_+, Y)$, the so-called *input–output stability*. Versions of Theorem 5.9 for the autonomous Hilbert space setting are proved in [6, Thm. 5.8] for the Pritchard–Salamon class, in [20, Cor. 1.8] for regular systems, and in [17, Thm. 5.2] and [33, Thm. 5.3] for well-posed systems. In that case, the

input–output stability can be replaced by the equivalent condition that the transfer function $H(\lambda) = CR(\lambda, A_{-1})B$ is bounded for $\operatorname{Re}\lambda > 0$; cf. [31, Thm. 3.1]. In [5, Thm. 4.3], our theorem was shown for bounded control and observation operators using the characterization of exponential stability given in [4, Thm. 3.26]. Here we employ Datko's theorem [8, Thm. 1, Rem. 3] in order to avoid some technical problems. However, we remark that Datko's theorem can be deduced from [4, Thm. 3.26]; see [23, Thm. 1.19]. A variant of the next result holds for admissible systems [23, Thm. 4.29].

THEOREM 5.9. *Let $\Sigma = (T, \Phi, \Psi, \mathbb{F})$ be a regular nonautonomous system. Then the following assertions are equivalent:*

1. *T is exponentially stable.*
2. *(T, Φ) is stabilizable, and $\Phi(\cdot, 0) \in \mathcal{L}(L^p(\mathbb{R}_+, U), L^p(\mathbb{R}_+, X))$.*
3. *(T, Ψ) is detectable, and $\|\Psi_s x\|_{L^p([s,\infty),Y)} \le c\,\|x\|$ for $s \ge 0$ and $x \in X$.*
4. *Σ is detectable and stabilizable, and $\mathbb{F} \in \mathcal{L}(L^p(\mathbb{R}_+, U), L^p(\mathbb{R}_+, Y))$.*

Proof. Let 1 hold. Then Σ is always stabilizable (take $\Psi^F = 0$) and detectable (take $\Phi^K = 0$). The other assertions in 2–4 follow from Lemmas 2.3, 3.2, and 3.7. Extending $u \in L^p([s,\infty), U)$ by 0 to $\mathbb{R}_+$ and using causality, we see that the norms of $\Phi(\cdot, s)$ and $\mathbb{F}_s$ decrease as s increases. The assumptions in 2 and Lemma 2.3 yield $\|T(\cdot, s)x\|_{L^p([s,\infty),X)} \le c\,\|x\|_X$ for $s \ge 0$, $x \in X$, and a constant c. Thus 1 is a consequence of Datko's theorem [8, Thm. 1, Rem. 3]. The implication "3⇒1" can be proved in the same way. If Σ is stabilizable, then Theorems 2.7 and 3.11 show that the operators $C(t)$ representing Ψ are also T_F-admissible, and $\Psi_s x = C(\cdot)T_F(\cdot, s)x - \mathbb{F}_s \Psi^F_s x$ for $s \ge 0$ and $x \in X$. Hence 4 implies 1 by Lemmas 2.3 and 2.5. ∎

6. A parabolic problem with point control and observation. Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with C^2-boundary $\partial\Omega$, and let $a_{kl}, a_k, a_0 : \mathbb{R}_+ \times \overline{\Omega} \to \mathbb{R}$, $k, l = 1, \dots, n$ be bounded and uniformly Hölder continuous such that $\sum_{kl} a_{kl}(t,\xi) v_k v_l \ge \alpha\,|v|^2$ for a constant $\alpha > 0$ and $v \in \mathbb{R}^n$, $t \ge 0$, $\xi \in \overline{\Omega}$. Further, let $b, c : \mathbb{R}_+ \to \Omega$ be uniformly Lipschitz such that $|b(t) - c(t)| \ge \delta > 0$ for $t \ge 0$. Let $\varphi \in C_0(\Omega)$, $s \ge 0$, and $D_k = \frac{\partial}{\partial \xi_k}$. The unique solution $w \in C([s,\infty) \times \overline{\Omega}) \cap C^{1,2}((s,\infty) \times \Omega)$ of the problem

$$
\begin{aligned}
w_t(t,\xi) &= \sum\nolimits_{kl} a_{kl}(t,\xi) D_k D_l w(t,\xi) + \sum\nolimits_k a_k(t,\xi) D_k w(t,\xi) \\
&\quad + a_0(t,\xi) w(t,\xi), \quad t > s, \\
w(t,\xi) &= 0, \quad \xi \in \partial\Omega,\ t \ge s, \qquad w(s,\xi) = \varphi(\xi), \quad \xi \in \Omega,
\end{aligned} \tag{6.1}
$$

is given by $w(t,\xi) = \int_\Omega k(t,s,\xi,\eta)\varphi(\eta)\,d\eta$ for a continuous kernel $k(t,s,\xi,\eta)$, $t > s \ge 0$, $\xi \in \overline{\Omega}$, $\eta \in \Omega$, satisfying the Gaussian estimate

$$
|k(t,s,\xi,\eta)| \le M(t-s)^{-\frac{n}{2}} \exp\left(-\frac{w|\xi-\eta|^2}{t-s} + \tilde{w}(t-s)\right)
$$

for $0 < t - s \le t_0$ and constants $M, w > 0$ and $\tilde{w} \in \mathbb{R}$; see, e.g., [15, section IV.16]. By the uniqueness of solutions, we also have

$$
k(t,s,\xi,\eta) = \int_\Omega k(t,r,\xi,\zeta) k(r,s,\zeta,\eta)\,d\zeta, \quad t > r > s \ge 0,\ \xi,\eta \in \Omega. \tag{6.2}
$$

We take $p, q \in [1, \infty)$ and set $X = L^q(\Omega)$, $U = Y = \mathbb{C}$, $T(s,s) = I$, $\Phi(s,s)u = 0$, and

$$T(t,s)\varphi = \int_\Omega k(t,s,\cdot,\eta)\varphi(\eta)\,d\eta, \qquad \Phi_{t,s}u = \int_s^t k(t,\tau,\cdot,b(\tau))u(\tau)\,d\tau,$$
$$(\Psi_s\varphi)(t) = \int_\Omega k(t,s,c(t),\eta)\varphi(\eta)\,d\eta, \qquad (\mathbb{F}_s u)(t) = \int_s^t k(t,\tau,c(t),b(\tau))u(\tau)\,d\tau$$

for $t > s \geq 0$, $\varphi \in X$, and $u \in L^p_{loc}(\mathbb{R}_+)$. These maps correspond to the PDE (6.1) complemented by the control $B(t)u(t) = \delta_{b(t)}u(t)$ and the output $y(t) = w(t, c(t))$.

The operators $T(t,s)$ yield an evolution family on X due to standard elliptic regularity and, e.g., [2, Thm. II.4.4.1]. Since b is Lipschitz, we have

$$\exp\left(-w\,\frac{|\xi - b(s)|^2}{t-s}\right) \leq c_1 \exp\left(-w\,\frac{|\xi - b(t)|^2}{t-s}\right) \tag{6.3}$$

for $t > s \geq 0$, $\xi \in \Omega$, and a constant c_1. Hence $|(\mathbb{F}_s u)(t)| \leq \varphi * |u|\,(t)$, $s \leq t \leq s + t_0$, where we have extended $u \in L^p_{loc}([s,\infty))$ by 0 and put $\varphi(t) = c_1\, t^{-\frac{n}{2}} \exp(-\frac{w\delta^2}{t} + \tilde{w}t)$ for $t > 0$ and $\varphi(t) = 0$ otherwise. For $1 + \frac{1}{p} = \frac{1}{\lambda} + \frac{1}{\mu}$, Young's inequality yields

$$\|\mathbb{F}_s u\|_{L^p[s,s+t_0]} \leq \|\varphi\|_{L^\lambda[0,t_0]}\, \|u\|_{L^\mu[s,s+t_0]} \tag{6.4}$$

so that (3.11) holds for each $p \in [1,\infty]$. Observe that $\Psi_s\varphi$ is continuous on (s,∞) for each $\varphi \in L^1(\Omega)$ and that $t \mapsto \Phi(t,s)u \in L^1(\Omega)$ is continuous on $[s,\infty)$ for $u \in L^1_{loc}([s,\infty))$. Moreover, (6.2) implies (2.1), (3.1), and (3.10). Using (6.3), Hölder's inequality, and that the norm of the Gaussian kernel in $L^{q'}(\mathbb{R}^n)$ equals $ct^{-n/2q}$, we compute

$$\int_s^t |\Psi_s\varphi(\tau)|^p\,d\tau \leq c\,\|\varphi\|_q^p \int_s^t (\tau - s)^{-\frac{np}{2q}}\,d\tau, \tag{6.5}$$

$$\|\Phi_{t,s}u\|_q \leq c\int_s^t (t-\tau)^{-\frac{n}{2q'}}\,|u(\tau)|\,d\tau \tag{6.6}$$

for $0 < t - s \leq t_0$, $1/q + 1/q' = 1$, and constants c. Thus the operators

$$\Psi_s : L^q(\Omega) \to L^p_{loc}([s,\infty)),\ \ q > \frac{np}{2}, \quad \Phi_{t,s} : L^p_{loc}([s,t]) \to L^q(\Omega),\ \ q' > \frac{np'}{2}, \tag{6.7}$$

are continuous. As a result, $(T, \Phi, \Psi, \mathbb{F})$ is a well-posed nonautonomous system provided that $n = 1$, $q > p/2$, and $q' > p'/2$ (for instance, if $p = q = 2$). In view of (6.4), this system is absolutely regular, and every bounded feedback is admissible.

The restriction $n = 1$ was needed only to obtain the boundedness of Ψ_s and $\Phi_{t,s}$ for the same values of p and q. We now discuss to what extent the assertions of Theorem 4.4 remain valid for $n = 2, 3$. All results dealing with Ψ and Φ *separately* are true for the exponents indicated in (6.7). Observe that $\mathbb{F}_s$ satisfies (3.11), (3.13), and the assertions of Lemma 4.2 for all $p \geq 1$ and that every bounded feedback is admissible. Moreover, the proof and assertion of Theorem 3.11 work as before. Proposition 3.12 holds for $n = 2$, $u \in L^r_{loc}(\mathbb{R}_+)$ with $r > 1$, $X = L^q(\Omega)$ with $q' > r'$, and $p < q$ in the assertions. In fact, we have $B_n u \in L^\infty_{loc}(\mathbb{R}_+, L^q(\Omega))$ for $q' > r'$. So we can apply Proposition 2.11 for $p < q$ to obtain (3.21) and then proceed as before.

We now consider Theorem 4.4, where we restrict ourselves to the case where the state space X equals $L^2(\Omega)$ and the given system Σ is admissible with exponent 2.

We define $T_\Delta(t,s)\varphi$ for $\varphi \in L^q(\Omega)$ with $q > 2n/(4-n)$ as in (4.8). Then $\Psi_s\varphi \in L^p_{loc}([s,\infty))$ for all $p \in (4/(4-n), 2q/n)$, and $T_\Delta(t,s) : L^q(\Omega) \to L^2(\Omega)$ is bounded for $0 \le t-s \le t_0$ due to (6.7). Because of (6.4) with $p=\infty$ and $\mu < 4/n$ and (6.7), we have

$$\|\mathbb{F}_s\Delta(\cdot)(I - \mathbb{F}(s+t_1,s)\Delta(\cdot))^{-1}\Psi_s x\|_{L^\infty[s,s+t_0]} \le c\,\|\Psi_s\varphi\|_{L^\mu[s,s+t_0]} \le c\,\|\varphi\|_2\,.$$

Thus (4.9) and (6.5) yield

$$|C(t)T_\Delta(t,s)\varphi| \le c\,(t-s)^{-\frac{n}{4}}\,\|\varphi\|_2 \tag{6.8}$$

for $0 < t-s \le t_0$. The identity (4.5) (with $\varphi \in L^q(\Omega)$) follows as before. Using (4.5), (6.6), and (6.8), we further estimate $\|T_\Delta(t,s)\varphi\|_2 \le c\,(t-s)^{1-\frac{n}{2}}\,\|\varphi\|_2$. Therefore, we can extend $T_\Delta(t,s)$ and (4.5) to $L^2(\Omega)$. We can now argue as in the proof of Theorem 4.4 and deduce part (a) of the theorem if we replace (4.4) by (6.8), allow for a blow-up of $T_\Delta(t,s)$ as $t \to s$ if $n=3$, and consider solutions $x(\cdot)$ of (4.1) such that $x(\cdot) \in C([s,\infty), L^1(\Omega))$ and $C(\cdot)x(\cdot) \in L^1_{loc}([s,\infty))$.

Now let $n=2$. In part (b), we restrict ourselves in (4.11) to cut-off functions γ with compact support in (r,∞). (One checks as in [4, Thm. 3.12] that the set of resulting functions f is still dense in $L^p_{loc}([s,\infty),X)$. Here we need the strong continuity of $T_\Delta(t,s)$ at $t=s$ and must thus exclude $n=3$.) We proceed as before and deduce (4.15) and (4.16) for $u \in L^p_{loc}([s,\infty))$ with $p > 2$. We can take the limits as $n \to \infty$ in $C([s,s+t_1], L^2(\Omega))$ and $L^2[s,s+t_1]$, respectively, and obtain $\Phi^\Delta u$ and $\mathbb{F}^\Delta u$. Moreover, $\Phi^\Delta u$ satisfies an estimate like (6.6). We can thus apply $\Phi^\Delta_{t,s}$ on $u = \Delta(\cdot)\Psi_s\varphi$ for $\varphi \in L^2(\Omega)$ so that (4.7) holds. The other assertions of Theorem 4.4(b) can be verified as before except that Ψ^Δ and Φ^Δ have the mapping properties from (6.7) with $q=2$.

Acknowledgments. I started this project together with Yuri Latushkin and Timothy Randolph from the University of Missouri, and I profited greatly from many joint discussions.

REFERENCES

[1] P. Acquistapace and B. Terreni, *Classical solutions of nonautonomous Riccati equations arising in parabolic boundary control problems*, Appl. Math. Optim., 39 (1999), pp. 361–409.

[2] H. Amann, *Linear and Quasilinear Parabolic Problems. Volume* 1: *Abstract Linear Theory*, Birkhäuser Boston, Boston, 1995.

[3] A. Bensoussan, G. Da Prato, M.C. Delfour, and S.K. Mitter, *Representation and Control of Infinite-Dimensional Systems, Volume* I, Birkhäuser Boston, Boston, 1992.

[4] C. Chicone and Y. Latushkin, *Evolution Semigroups in Dynamical Systems and Differential Equations*, AMS, Providence, RI, 1999.

[5] S. Clark, Y. Latushkin, S. Montgomery-Smith, and T. Randolph, *Stability radius and internal versus external stability in Banach spaces: An evolution semigroup approach*, SIAM J. Control Optim., 38 (2000), pp. 1757–1793.

[6] R. Curtain, H. Logemann, S. Townley, and H. Zwart, *Well-posedness, stabilizability, and admissibility for Pritchard–Salamon systems*, J. Math. Systems Estim. Control, 7 (1997), pp. 439–476.

[7] R. F. Curtain and A. J. Pritchard, *Infinite Dimensional Linear System Theory*, Springer-Verlag, New York, 1978.

[8] R. Datko, *Uniform asymptotic stability of evolutionary processes in a Banach space*, SIAM J. Math. Anal., 3 (1972), pp. 428–445.

[9] K. J. Engel and R. Nagel, *One-Parameter Semigroups for Linear Evolution Equations*, Springer-Verlag, New York, 2000.

[10] D. Hinrichsen and A. J. Pritchard, *Robust stability of linear evolution operators on Banach spaces*, SIAM J. Control Optim., 32 (1994), pp. 1503–1541.
[11] A. Idrissi and A. Rhandi, *Admissibility of time-varying observations for non-autonomous systems*, J. Comput. Anal. Appl., to appear.
[12] B. Jacob, *Time-Varying Infinite Dimensional State-Space Systems*, Ph.D. thesis, Fachbereich Mathematik/Informatik, Universität Bremen, Bremen, 1995.
[13] B. Jacob, *Optimal control of time-varying well-posed linear systems on a finite time horizon*, in Mathematical Theory of Networks and Systems, Proceedings of the MTNS–98, A. Beghi, I. Finesso, and G. Picci, eds., Il Polografo, Padova, Italy, 1998, pp. 483–486.
[14] B. Jacob, V. Dragan, and A. J. Pritchard, *Robust stability of infinite dimensional time-varying systems with respect to nonlinear perturbations*, Integral Equations Operator Theory, 22 (1995), pp. 440–462.
[15] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, AMS, Providence, RI, 1968.
[16] J.-L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.
[17] K. A. Morris, *Justification of input–output methods for systems with unbounded control and observation*, IEEE Trans. Automat. Control, 44 (1999), pp. 81–85.
[18] A. J. Pritchard and D. Salamon, *The linear quadratic control problem for infinite-dimensional systems with unbounded input and output operators*, SIAM J. Control Optim., 25 (1987), pp. 121–144.
[19] F. Räbiger, A. Rhandi, R. Schnaubelt, and J. Voigt, *Non-autonomous Miyadera perturbations*, Differential Integral Equations, 13 (1999), pp. 341–368.
[20] R. Rebarber, *Conditions for the equivalence of internal and external stability for distributed parameter systems*, IEEE Trans. Automat. Control, 38 (1993), pp. 994–998.
[21] D. Salamon, *Infinite-dimensional linear systems with unbounded control and observation: A functional analytic approach*, Trans. Amer. Math. Soc., 300 (1987), pp. 383–431.
[22] R. Schnaubelt, *Sufficient conditions for exponential stability and dichotomy of evolution equations*, Forum Math., 11 (1999), pp. 543–566.
[23] R. Schnaubelt, *Exponential Dichotomy of Non-Autonomous Evolution Equations*, Habilitation thesis, Mathematische Fakultät, Universität Tübingen, Tübingen, 1999.
[24] R. Schnaubelt, *Asymptotically autonomous parabolic evolution equations*, J. Evol. Equ., 1 (2001), pp. 19–37.
[25] O. J. Staffans, *Well-Posed Linear Systems Part* I: *General Theory*, book manuscript dated July 9, 2001; available online from http://www.abo.fi/~staffans.
[26] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.
[27] J. Voigt, *On the perturbation theory for strongly continuous semigroups*, Math. Ann., 229 (1977), pp. 163–171.
[28] G. Weiss, *Admissible observation operators for linear semigroups*, Israel J. Math., 65 (1989), pp. 17–43.
[29] G. Weiss, *Admissibility of unbounded control operators*, SIAM J. Control Optim., 27 (1989), pp. 527–545.
[30] G. Weiss, *The representation of regular linear systems on Hilbert spaces*, in Control and Estimation of Distributed Parameter Systems (Vorau, 1988), F. Kappel, K. Kunisch, and W. Schappacher, eds., Birkhäuser Verlag, Basel, 1989, pp. 401–416.
[31] G. Weiss, *Transfer functions of regular linear systems* I: *Characterization of regularity*, Trans. Amer. Math. Soc., 342 (1994), pp. 827–854.
[32] G. Weiss, *Regular linear systems with feedback*, Math. Control Signals Systems, 7 (1994), pp. 23–57.
[33] G. Weiss and R. Rebarber, *Optimizability and estimatability for infinite-dimensional linear systems*, SIAM J. Control Optim., 39 (2000), pp. 1204–1232.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 4, pp. 1166–1193

STABILIZATION OF BEAMS WITH NONLINEAR FEEDBACK*

LARBI BERRAHMOUNE†

Abstract. This paper studies nonlinear feedback stabilization of Euler–Bernoulli beams subject to various pointwise and concentrated actuators. Strong, uniform, and nonuniform stabilization are obtained with explicit decay estimates in appropriate spaces. The results are obtained through the study of an abstract second order distributed system which encompasses the models under investigation.

Key words. stabilization, nonlinear feedback, pointwise actuator, concentrated actuator

AMS subject classifications. 93C20, 93D15

PII. S036301290138852X

1. Introduction and statement of the main abstract results.

1.1. Motivating examples. In this paper, we are concerned with the question of stabilization of second order distributed systems modelling connected vibrating beams provided with various types of actuators. In what follows, $y(x,t)$ represents the displacement of the structure in question at position x and time t. The notation y' denotes the derivative of y with respect to time.

The canonical classes of problems that we have in mind are as follows.

Model 1. *Beam equation with pointwise actuator.* Let $0 < a < 1$; we consider the system of coupled beams with spatial extent from $x = 0$ to $x = a$ and from $x = a$ to $x = 1$. We suppose that the end at $x = 0$ is simply supported and at $x = 1$ there is a shear hinge end. We consider the case where, at $x = a$, we have a rigid support joint with the discontinuity in the shear as control (see [8]). Then y satisfies the following Euler–Bernoulli equation:

$$y'' + \frac{\partial^4 y}{\partial x^4} = 0, \quad t > 0,\ x \in (0,a) \cup (a,1)\,. \tag{1.1}$$

The boundary and joint conditions are given by

$$\begin{cases} y(t,0) = \frac{\partial^2 y}{\partial x^2}(t,0) = 0, \\ \frac{\partial y}{\partial x}(t,1) = \frac{\partial^3 y}{\partial x^3}(t,1) = 0, \\ y(t,a^-) = y(t,a^+), \\ \frac{\partial y}{\partial x}(t,a^-) = \frac{\partial y}{\partial x}(t,a^+), \\ \frac{\partial^2 y}{\partial x^2}(t,a^-) = \frac{\partial^2 y}{\partial x^2}(t,a^+), \end{cases} \tag{1.2}$$

$$\frac{\partial^3 y}{\partial x^3}(t,a^+) - \frac{\partial^3 y}{\partial x^3}(t,a^-) = u(t). \tag{1.3}$$

*Received by the editors April 25, 2001; accepted for publication (in revised form) April 3, 2002; published electronically October 29, 2002.
http://www.siam.org/journals/sicon/41-4/38852.html
†Département de Mathématiques, Ecole Normale Supérieure de Rabat, BP 5118, Rabat, Morocco (berrahmoune@hotmail.com).

Model 2. *Beam equation with piezoelectric actuator.* We consider the Euler–Bernoulli beam that is subject to the action of an attached piezoelectric actuator [4]. We suppose that the beam is hinged at both ends and that the actuator is excited in a manner to produce pure bending moments. This leads to the following model [10]:

$$y'' + \frac{\partial^4 y}{\partial x^4} = u(t)\frac{d}{dx}(\delta(x-a_1) - \delta(x-a_2)), \quad t > 0,\ 0 < x < 1, \tag{1.4}$$

$$y(t,0) = y(t,1) = \frac{\partial^2 y}{\partial x^2}(t,0) = \frac{\partial^2 y}{\partial x^2}(t,1) = 0, \tag{1.5}$$

where $\delta(x-a)$ denotes the Dirac mass at a. Here the points $a_1,\ a_2 \in (0,1)$ represent the ends of the actuator and the control function $u(.)$ is the time variation of the voltage applied to the actuator.

Model 3. *Beam equation with concentrated actuator.* We consider the controlled Euler–Bernoulli beam equation

$$y'' + \frac{\partial^4 y}{\partial x^4} = u(t)g(x), \quad t > 0,\ 0 < x < 1, \tag{1.6}$$

with boundary conditions (1.5) and $g \in L^2(0,1)$. As in [12], we shall call the concentrated control (or actuator) the control relevant to (1.6). The introduction of such a system is motivated by the fact that, from an engineering viewpoint, point actuators as in examples (1.1)–(1.3) and (1.4) are idealizations. Indeed, the control actions resulting from such actuators affect a distributed part of the spatial domain and not just points. Thus, although it is generally accepted that point actuators are useful concepts which lead to the so-called unbounded controls, a more realistic model would be to consider concentrated actuators. Such considerations have been studied for beam equations in [12]. In this reference, it is shown that the uniform exponential stability is lost whenever the concentrated actuator concept is adopted for the same linear feedback. In this paper, we shall see that, even when such a concept is adopted, uniform exponential stability can be achieved in appropriate spaces by using unbounded feedbacks.

In the equations above, the function $u : (0,\infty) \to IR$ represents the control, and we suppose that $u \in L^2_{loc}(0,\infty)$. Then we consider the question of explicit (nonlinear) feedback operator based on the velocity

$$u(t) = F(y'(.,t))$$

such that the resulting equation produces a solution which is stable in appropriate spaces. Furthermore, under suitable assumptions, explicit decay estimates will be given.

1.2. Literature. As the foregoing examples indicate, the object of this paper is to study the problem of (nonlinear) feedback stabilization of various Euler–Bernoulli equations with interior control supported by points or zones. Let us state at the outset a part of the abundant literature concerned with the problem of linear feedback stabilization relative to our examples. Model 1 has been studied in [8] and [20]. In the first reference, the tool used is based on a combination of multipliers and the Lyapunov method. In the second one, a frequency domain approach is the main ingredient. As for Model 2, a strong stabilization result has been obtained in [26] via LaSalle's invariance principle. Model 3 can be treated via the general study dealing with the linear bounded feedback case in [21].

The main contributions of this paper are as follows:

(i) The problems above are presented in an abstract framework which encompasses these examples and others as well (wave equation, Kirchhoff models, etc.). Some aspects of this question will be considered in section 4.

(ii) The regularity results relative to the abstract model are sharp. The method used to this end gives an alternative way to obtain in a unified approach regularity properties already studied for their own interest (see [27] for Model 2).

(iii) The nonlinear feedback used to stabilize our models is general and less restrictive compared to those of the existing literature. Furthermore, it enables us to get the linear feedback stabilization results as particular cases (see [1] for Model 1).

(iv) To the author's knowledge and for the models above, the results deduced from the abstract framework are new in the nonlinear feedback stabilization setting. These results can be used to improve some partial results concerned with particular situations (see, for instance, [9], where boundary feedback stabilization is studied).

The plan of the paper is as follows. The rest of this section is devoted to the statement of regularity and stabilization results relative to an abstract formulation of our models. The second section is concerned with the proofs of the abstract results. In the third section, we justify that our problems can be written in abstract form and derive various stabilization results. The fourth section contains some comments on possible extensions and related questions. In the remaining part of this paper, we shall denote by C a generic positive constant which may be different at different occurrences.

1.3. Formulation of the problem and statement of the abstract results. In this subsection, we consider an abstract second order distributed control system subject to certain assumptions. In latter sections, we verify that these assumptions are natural for, and, in fact, automatically satisfied by, the models of our interest.

Let V, H be real Hilbert spaces such that $V \subset H$ with dense, continuous, and compact embedding. We denote by $|.|$ and $\|.\|$ the norms on H and V, respectively. Let us introduce the unique linear bounded operator from V to V' which is characterized by

$$\langle Av, w\rangle_{V',V} = \langle v, w\rangle_V \quad \text{for all } v, w \in V. \tag{1.7}$$

Motivated by the examples given in the introduction, we shall consider the abstract system

$$\begin{cases} y''(t) + Ay(t) = u(t)b, \quad 0 < t < T, \\ y(0) = y_0,\ y'(0) = y_1, \end{cases} \tag{1.8}$$

where $u \in L^2(0,T)$ represents the control and b is given in V'.

Recall that, by identifying H with its dual, we have $V \subset H \equiv H' \subset V'$ and

$$\langle v, w\rangle_H = \langle v, w\rangle \quad \text{for all } v, w \in H, \tag{1.9}$$

where $\langle ., .\rangle$ denotes the V', V duality pairing with respect to the H-topology. Hence, if we consider A as an operator on H with

$$D(A) = A^{-1}(H), \tag{1.10}$$

then A is self-adjoint with compact resolvent on H.

We shall consider the stability of the control system under (possibly nonlinear) feedback control of the form

$$u(t) = -f\left(\frac{d}{dt}\langle b, y(t)\rangle\right), \tag{1.11}$$

where $f : IR \to IR$ is continuous and monotone and satisfies

$$rf(r) \geq 0 \quad \text{for all } r \in IR. \tag{1.12}$$

Thus the resulting system, whose stability properties we shall investigate, is the following:

$$\begin{cases} y''(t) + Ay(t) + f(\frac{d}{dt}\langle b, y(t)\rangle)b = 0, \\ y(0) = y_0,\ y'(0) = y_1. \end{cases} \tag{1.13}$$

The main goal is to characterize those b for which (i) the solutions of (1.8) and (1.13) exist globally and (ii) the energy given by

$$E(t) = \frac{1}{2}(\|y(t)\|^2 + |y'(t)|^2) \tag{1.14}$$

decays to zero when $t \to \infty$ for every solution of (1.13). Furthermore, we shall specify the decay rate in terms of the function f.

Let us introduce the set $\{\lambda_n = \omega_n^2\}_n$ of eigenvalues of A and denote by $\{\psi_n\}_n$ the corresponding orthogonal basis in H. Then we have the standard identifications

$$V \equiv D(A^{\frac{1}{2}}),\ \ \|v\|^2 = \|v\|^2_{D(A^{\frac{1}{2}})} = \sum_n \omega_n^2 \left|\langle v, \psi_n\rangle\right|^2. \tag{1.15}$$

More generally, for $0 < \mu \leq 1$, we shall use the fractional power spaces

$$D(A^\mu) = \left\{ v \in H \,/\, \sum_n \omega_n^{4\mu} \left|\langle v, \psi_n\rangle\right|^2 < \infty \right\},\ \ \|v\|^2_{D(A^\mu)} = \sum_n \omega_n^{4\mu} \left|\langle v, \psi_n\rangle\right|^2. \tag{1.16}$$

The space $D(A^\mu)'$ will denote the dual of $D(A^\mu)$ with respect to the H-topology so that it can be characterized as the completion of H for the norm defined by

$$\|v\|^2_{D(A^\mu)'} = \sum_n \omega_n^{-4\mu} \left|\langle v, \psi_n\rangle\right|^2. \tag{1.17}$$

In the applications, the operator A will stand for $\frac{d^4}{dx^4}$ with appropriate homogenous boundary conditions so that the spaces above can be identified with Sobolev spaces in a standard way [13], [19]. We shall use later the fact that the embedding $H \subset V'$ is also compact.

Let us mention that, from standard theory, the system (1.8) has a unique solution with $y \in C\,(0, T; H) \cap C^1\,(0, T; V')$ [19, p. 311]. In fact, we shall prove that the solution can be more regular. As for the feedback system (1.13), similar regularity results will be obtained under less demanding assumptions.

Below, we shall state our main abstract results, while the proofs are relegated to section 2.

PROPOSITION 1.1 (well posedness of (1.8) on $V \times H$). *Suppose that*

$$\underline{\lim}_n \omega_{n+1} - \omega_n \geq \delta > 0 \tag{1.18}$$

and, for some positive constant C_b,

$$|\langle b, \psi_n \rangle| \leq C_b, \quad \text{for all } n. \tag{1.19}$$

Then, for any $\{y_0, y_1\} \in V \times H$, there exists a unique weak solution to (1.8) such that

$$y \in C\left(0, T; V\right) \cap C^1\left(0, T; H\right). \tag{1.20}$$

Before stating the following corollary, we introduce, for $-1 \leq \theta \leq 1$, the spaces V_θ given by

$$V_\theta = \begin{cases} D(A^\theta), & \text{if } \theta \geq 0, \\ D(A^{-\theta})' & \text{otherwise.} \end{cases} \tag{1.21}$$

Then we have the following corollary.

COROLLARY 1.2. *Suppose that (1.18) holds and, for some $0 < \mu \leq 1$ and some positive constant C_b,*

$$|\langle b, \psi_n \rangle| \leq C_b \lambda_n^\mu \quad \text{for all } n. \tag{1.22}$$

Then, for any $\{y_0, y_1\} \in V_{\frac{1}{2}-\mu} \times D(A^\mu)'$, there exists a unique weak solution to (1.8) such that

$$y \in C(0, T; V_{\frac{1}{2}-\mu}) \cap C^1\left(0, T; D(A^\mu)'\right). \tag{1.23}$$

PROPOSITION 1.3. *Suppose that the assumptions of Proposition 1.1 hold. Moreover, assume that the sequence $\{\sigma_n\}_n$ given by*

$$\sigma_n = \sum_{j \neq n} \frac{1}{\left|\sqrt{\omega_n} - \sqrt{\omega_j}\right|^2} \tag{1.24}$$

is well defined and bounded. Then, for y solution of

$$\begin{cases} y''(t) + Ay(t) = u(t)b, \quad 0 < t < T, \\ y(0) = y'(0) = 0, \end{cases} \tag{1.25}$$

the function $\frac{d}{dt}\langle b, y(.)\rangle$ is well defined in $L^2(0,T)$, and

$$\left\| \frac{d}{dt} \langle b, y(.) \rangle \right\|_{L^2(0,T)} \leq C \left\| u \right\|_{L^2(0,T)} \tag{1.26}$$

for some positive constant C.

PROPOSITION 1.4 (well posedness of (1.13) on $V \times H$). *Let $f : IR \to IR$ be a continuous monotone function satisfying (1.12). Then, for any $\{y_0, y_1\} \in V \times H$, there exists a unique weak solution to (1.13) with the regularity (1.20).*

To state our first stability result for system (1.13), we need the following assumption regarding f:

$$rf(r) > 0 \quad \text{for all } r \neq 0. \tag{1.27}$$

Then we have the following theorem.

THEOREM 1.5 (strong stability for the solution of (1.13)). *Let $f : IR \to IR$ be a continuous monotone function satisfying (1.27). Suppose that the eigenvalues λ_n are simple. Then, for any $\{y_0, y_1\} \in V \times H$, the following conditions are equivalent:*

(i) *The solution of* (1.13) *satisfies* $\lim_{t\to\infty} E(t) = 0$.
(ii) $\langle b, \psi_n\rangle \neq 0$ *for all* n.
(iii) *For any* $\{\varphi_0, \varphi_1\} \in D(A) \times V$, *the only function satisfying the conditions*

$$\begin{cases} \varphi''(t) + A\varphi(t) = 0, \quad 0 < t < \infty, \\ \varphi(0) = \varphi_0,\ \varphi'(0) = \varphi_1, \\ \langle b, \varphi'(t)\rangle = 0, \quad 0 < t < \infty, \end{cases} \tag{1.28}$$

is $\varphi \equiv 0$.

To state our second stabilization result, we consider the assumptions

$$c_f\,|r| \leq |f(r)| \leq C_f\,|r| \quad \text{for all } |r| \geq 1, \tag{1.29}$$

where c_f, C_f are positive constants and there is a concave, strictly increasing function $\xi : IR^+ \to IR$ with $\xi(0) = 0$ such that, for some $\rho \in (0, 1]$,

$$|r|^{2\rho} + |f(r)|^2 \leq \xi(rf(r)) \quad \text{for all } |r| < 1. \tag{1.30}$$

On the other hand, for T satisfying

$$T > \frac{2\pi}{\delta}, \tag{1.31}$$

we set

$$h(s) = s + \xi\left(\frac{s}{T}\right). \tag{1.32}$$

Let p denote the inverse of $C_0 h$ where C_0 is a positive constant to be precise in the proof. Then we have the following theorem.

THEOREM 1.6 (stability and general decay estimate for the solution of (1.13)). *Let* f *be a continuous monotone function satisfying* (1.27), (1.29), *and* (1.30). *Suppose that the eigenvalues* λ_n *are simple, satisfying* (1.18) *and such that the sequence* $\{\sigma_n\}_n$ *is well defined and bounded. Moreover, assume that, for some positive constants* C_b, c_b,

$$c_b \leq |\langle b, \psi_n\rangle| \leq C_b \quad \textit{for all } n. \tag{1.33}$$

Then, for some $T > 0$, *the solution of the system* (1.13) *satisfies*

$$E(t) \leq S\left(\frac{t}{T} - 1\right) \quad \textit{for all } t > T, \tag{1.34}$$

where $S(t) \to 0$ *as* $t \to \infty$ *and is the solution (contraction semigroup) of the differential equation*

$$S'(t) + q(S(t)) = 0,\ S(0) = E(0), \tag{1.35}$$

and q *is given by*

$$q(s) = s - (I + p)^{-1}(s). \tag{1.36}$$

THEOREM 1.7 (stability and decay estimate for the solution of (1.13)). *Let* f *be a continuous monotone function satisfying* (1.27), (1.29), *and the additional hypotheses*

$$\widetilde{c}_f\,|r|^{\alpha+1} \leq rf(r) \quad \textit{for all } |r| < 1, \tag{1.37}$$

$$|f(r)| \leq \widetilde{C}_f \, |r|^{\beta} \quad \textit{for all } |r| < 1, \tag{1.38}$$

where $\widetilde{C}_f$, $\widetilde{c}_f$, α, β *are positive constants with* $0 < \beta \leq 1$ *and* $\alpha \geq \beta$. *Suppose that the eigenvalues* λ_n *satisfy the assumptions of Theorem* 1.6. *Moreover, assume that* (1.33) *holds. Then, for some positive constants* ω, T, *and* K, *the solution of the system* (1.13) *satisfies*

$$E(t) \leq KE(0)e^{-\omega t} \quad \textit{for all } t > T \tag{1.39}$$

if $\alpha = \beta = 1$ *and*

$$E(t) = O(t^{-\frac{2\beta}{\alpha+1-2\beta}}) \; (t \to \infty) \tag{1.40}$$

if $\alpha + 1 > 2\beta$.

The following result gives a nonuniform stability property for (1.13) when the first inequality in (1.33) is not satisfied. More precisely, the decay rates are obtained for initial data lying in a space dense in the energy space.

THEOREM 1.8 (nonuniform stability and decay estimate for the solution of (1.13)). *Let* f *be a continuous monotone function satisfying the assumptions of Theorem* 1.7 *with* $\alpha = \beta = 1$. *Suppose that the eigenvalues* λ_n *satisfy the assumptions of Theorem* 1.6. *Assume that* $b \in H$ *and, for some positive constants* C_b, c_b *and some* $0 < \mu \leq 1$,

$$\frac{c_b}{\lambda_n^{\mu}} \leq |\langle b, \psi_n \rangle| \leq C_b \quad \textit{for all } n. \tag{1.41}$$

Then, for any $\{y_0, y_1\} \in D(A) \times V$, *the solution of the system* (1.13) *satisfies*

$$\begin{cases} E(t) = d(t)(\|y_0\|^2_{D(A)} + \|y_1\|^2), \\ d(t) = O(t^{-\frac{1}{2\mu}}) \;\; (t \to \infty). \end{cases} \tag{1.42}$$

2. Proofs of the main abstract results.

2.1. Well posedness and regularity. Without loss of generality, we shall suppose in subsubsections 2.1.1–2.1.3 that $T > \frac{2\pi}{\delta}$.

2.1.1. Proof of Proposition 1.1. It is standard that (1.8) admits a unique solution satisfying [19, p. 311]

$$y \in C(0, T; H) \cap C^1(0, T; V'). \tag{2.1}$$

On the other hand, if we consider the uncontrolled system

$$\begin{cases} \varphi''(t) + A\varphi(t) = 0, \\ \varphi(0) = \varphi_0, \; \varphi'(0) = \varphi_1, \end{cases} \tag{2.2}$$

and suppose that $\{\varphi_0, \varphi_1\} \in V \times H$, then we have

$$\frac{d}{dt} \langle b, \varphi(t) \rangle = \sum_n \left[-\omega_n \langle \psi_n, \varphi_0 \rangle \sin(\omega_n t) + \langle \psi_n, \varphi_1 \rangle \cos(\omega_n t) \right] \langle b, \psi_n \rangle . \tag{2.3}$$

If we consider the series above as a Fourier series in t and we apply Ingham's inequality, we obtain, for some positive constant C [3],

$$\left\| \frac{d}{dt} \langle b, \varphi(.) \rangle \right\|^2_{L^2(0,T)} \leq C(\|\varphi_0\|^2 + |\varphi_1|^2). \tag{2.4}$$

By adapting Proposition 3.1 in [5, p. 172], we can easily deduce (1.20). ∎

2.1.2. Proof of Corollary 1.2. Let us introduce

$$\widetilde{b} = A^{-\mu} b \tag{2.5}$$

and consider the change of variables

$$\begin{cases} \widetilde{y}(t) = A^{-\mu} y(t), \\ \widetilde{y}_0 = A^{-\mu} y_0, \\ \widetilde{y}_1 = A^{-\mu} y_1. \end{cases} \tag{2.6}$$

Then $\widetilde{b} \in V'$, and $\widetilde{y}$ is the solution of the problem

$$\begin{cases} \widetilde{y}''(t) + A\widetilde{y}(t) = u(t)\widetilde{b} \quad 0 < t < T, \\ \widetilde{y}(0) = \widetilde{y}_0,\ \widetilde{y}'(0) = \widetilde{y}_1. \end{cases} \tag{2.7}$$

On the other hand, for some positive constant C,

$$|\langle \widetilde{b}, \psi_n \rangle| \leq C \quad \text{for all } n. \tag{2.8}$$

Hence Proposition 1.1 yields $\widetilde{y} \in C\left(0, T; V\right) \cap C^1\left(0, T; H\right)$. This completes the proof of the corollary. □

2.1.3. Proof of Proposition 1.3. We first note that, by using the Galerkin method, the solution of (1.25) can be approximated by the sequence of solutions given by

$$\begin{cases} \langle y_m''(t), \psi_k \rangle + \langle A y_m(t), \psi_k \rangle = u(t) \langle b, \psi_k \rangle \quad 0 < t < T, \\ 1 \leq k \leq m,\ y_m(t) \in \mathit{span}\ (\psi_1, \ldots,\ \psi_m), \\ y_m(0) = y_m'(0) = 0. \end{cases} \tag{2.9}$$

Indeed, by adapting the methods performed in [19, chapter 3, section 8], it can be shown that we can extract a subsequence, still denoted by $\{y_m\}_m$, such that

$$y_m \to y \text{ weakly in } L^2(0, T; H), \tag{2.10}$$

$$y_m' \to y' \text{ weakly in } L^2(0, T; V'). \tag{2.11}$$

On the other hand, we notice that, for $u \in H_0^1(0, T)$, we can easily get from above that

$$y' \in C\left(0, T; V\right). \tag{2.12}$$

Hence the function $\frac{d}{dt} \langle b, y(.) \rangle$ is well defined in $C(0, T)$. We shall use the following result.

LEMMA 2.1. *Suppose that $u \in H_0^1(0, T)$. Then the sequence $\{y_m\}_m$ satisfies*

$$\{y_m'\}_m \ \textit{bounded in } C\left(0, T; V\right), \tag{2.13}$$

$$\{y_m''\}_m \ \textit{bounded in } C\left(0, T; H\right). \tag{2.14}$$

Proof. From

$$y'_m(t) = \sum_{k=1}^{m} \left\{ \int_0^t u'(s) \sin \omega_k (t-s) ds \right\} \frac{\langle b, \psi_k \rangle}{\omega_k} \psi_k \tag{2.15}$$

and using (1.19), we get, for some positive constant C,

$$\|y'_m(t)\|_V^2 \leq C \sum_{k=1}^{m} \left\{ \int_0^t u'(s) \sin \omega_k (t-s) ds \right\}^2. \tag{2.16}$$

By an easy adaptation of the auxiliary result in Lemma A.1 (see Appendix A), we obtain, for some positive constant C independent of m,

$$\|y'_m(t)\|_V \leq C \|u'\|_{L^2(0,T)} \quad \text{for all } 0 < t < T. \tag{2.17}$$

The assertion (2.14) can be obtained in the same way. The proof of the lemma is complete. □

In order to complete the proof of (1.26), we notice that, by standard compactness argument [23], there exists a subsequence, still denoted by $\{y'_m\}_m$, such that

$$y'_m \to y' \text{ in } C(0,T;H), \tag{2.18}$$

$$y'_m \to y' \text{ weakly in } L^2(0,T;V). \tag{2.19}$$

Furthermore, from (2.13) we get

$$\langle b, y'_m(.) \rangle \text{ bounded in } L^2(0,T) \tag{2.20}$$

so that we can extract a subsequence such that

$$\langle b, y'_m(.) \rangle \to \langle b, y'(.) \rangle \text{ weakly in } L^2(0,T). \tag{2.21}$$

By a density argument, it is sufficient to establish (1.26) for $u \in H_0^1(0,T)$. Taking into account (2.21) and using a weak compactness argument, this can be reduced to

$$\|\langle b, y'_m(.) \rangle\|_{L^2(0,T)} \leq C \|u\|_{L^2(0,T)} \tag{2.22}$$

for some positive constant C independent of m. To this end, let us denote by $\widetilde{u}$ the extension of u defined by

$$\widetilde{u}(t) = \begin{cases} u(t) & \text{if } 0 < t < T, \\ 0 & \text{otherwise,} \end{cases} \tag{2.23}$$

and consider the sequences $\{s_k\}_k$ and $\{S_m\}_m$ given by

$$s_k(t) = \begin{cases} |\langle b, \psi_k \rangle|^2 e^{i\omega_k t} & \text{if } 0 < t < T, \\ 0 & \text{otherwise,} \end{cases} \tag{2.24}$$

$$S_m(t) = \sum_{k=1}^{m} s_k(t). \tag{2.25}$$

Then, from (2.15), we get

$$\|\langle b, y'_m(.)\rangle\|_{L^2(0,T)} \leq \|(S_m * \widetilde{u})(.)\|_{L^2(IR)}\,. \tag{2.26}$$

For $w \in L^2(IR)$, let $\widehat{w}$ denote the Fourier transform of w defined by

$$\widehat{w}(\tau) = \int_{IR} e^{-it\tau} w(t)dt. \tag{2.27}$$

Then, using Parseval's property, we have

$$\|\langle b, y'_m(.)\rangle\|_{L^2(0,T)} \leq \|\widehat{S}_m \widehat{\widetilde{u}}\|_{L^2(IR)} \tag{2.28}$$

and

$$\|\langle b, y'_m(.)\rangle\|_{L^2(0,T)} \leq \|\widehat{S}_m\|_{L^\infty(IR)} \|\widehat{\widetilde{u}}\|_{L^2(IR)}. \tag{2.29}$$

Hence, it is sufficient to see that, for some positive constant C,

$$\|\widehat{S}_m\|_{L^\infty(IR)} \leq C \quad \text{for all } m. \tag{2.30}$$

By continuity, this can be reduced to see that, for any τ such that

$$\tau \neq \omega_k \quad \text{for all } k,$$

we have

$$|\widehat{S}_m(\tau)| \leq C \quad \text{for all } m. \tag{2.31}$$

In what follows, we shall implicitly consider this case so that

$$\widehat{S}_m(\tau) = \sum_{k=1}^{m} \widehat{s}_k(\tau) = \sum_{k=1}^{m} \frac{e^{i(\omega_k - \tau)T} - 1}{i(\omega_k - \tau)} \left|\langle b, \psi_k\rangle\right|^2. \tag{2.32}$$

If $\tau \leq 0$, then we have

$$|\widehat{S}_m(\tau)| \leq \sum_{k=1}^{m} \frac{2C_b^2}{\omega_k} \leq 2C_b^2 \left(\sigma_1 + \frac{1}{\omega_1}\right) \quad \text{for all } m. \tag{2.33}$$

On the other hand, an easy computation gives

$$\left|\frac{e^{i(\omega_k - \tau)T} - 1}{i(\omega_k - \tau)}\right| = \frac{\left|4\sin^2(\omega_k - \tau)T/2\right|^{\frac{1}{2}}}{|\omega_k - \tau|} = T\frac{|\sin(\omega_k - \tau)T/2|}{|\omega_k - \tau|\, T/2}.$$

Hence we have, for all real τ,

$$|\widehat{s}_k(\tau)| \leq TC_b^2. \tag{2.34}$$

Suppose that $\tau > 0$, and let

$$n(\tau) = \min\{k \,/\, \tau < \omega_k\}\,. \tag{2.35}$$

If $n(\tau) \leq 3$, then

$$|\widehat{S}_m(\tau)| \leq \left|\sum_{k=1}^{n(\tau)} \widehat{s}_k(\tau)\right| + \left|\sum_{k=n(\tau)+1}^{m} \widehat{s}_k(\tau)\right|$$
$$\leq 3TC_b^2 + 2C_b^2 \left|\sum_{k=n(\tau)+1}^{m} \frac{1}{\omega_k - \omega_{n(\tau)}}\right|$$
$$\leq 3TC_b^2 + 2C_b^2 \left[\sum_{k=n(\tau)+1}^{m} \frac{1}{|\sqrt{\omega_k} - \sqrt{\omega_{n(\tau)}}|^2}\right]^{\frac{1}{2}} \left[\sum_{k=n(\tau)+1}^{m} \frac{1}{|\sqrt{\omega_k} + \sqrt{\omega_{n(\tau)}}|^2}\right]^{\frac{1}{2}}.$$

Thus

$$|\widehat{S}_m(\tau)| \leq 3TC_b^2 + 2C_b^2 \sigma_{n(\tau)}. \tag{2.36}$$

If $n(\tau) > 3$, then

$$|\widehat{S}_m(\tau)| \leq \left|\sum_{k=1}^{n(\tau)-2} \widehat{s}_k(\tau)\right| + \left|\widehat{s}_{n(\tau)-1}(\tau)\right| + \left|\widehat{s}_{n(\tau)}(\tau)\right| + \left|\sum_{k=n(\tau)+1}^{m} \widehat{s}_k(\tau)\right|,$$

and, as above, we can easily obtain

$$\left|\sum_{k=1}^{n(\tau)-2} \widehat{s}_k(\tau)\right| \leq \sum_{k=1}^{n(\tau)-2} \frac{2C_b^2}{\omega_{n(\tau)-1} - \omega_k} \leq 2C_b^2 \sigma_{n(\tau)-1}$$

and

$$\left|\sum_{k=n(\tau)+1}^{m} \widehat{s}_k(\tau)\right| \leq 2C_b^2 \sigma_{n(\tau)}.$$

Hence the following holds:

$$|\widehat{S}_m(\tau)| \leq 2C_b^2 \sigma_{n(\tau)-1} + 2TC_b^2 + 2C_b^2 \sigma_{n(\tau)}. \tag{2.37}$$

The inequalities (2.33), (2.36), and (2.37), combined with the boundedness of the sequence (1.24), imply (2.30). This completes the proof of (1.26). Hence the proof of Proposition 1.3 is complete. □

2.1.4. Proof of Proposition 1.4. We shall be concerned with the unbounded case where $b \notin H$. The other case can be proved by similar arguments. Let $\widetilde{A}$ denote the (nonlinear) operator in $V \times H$ defined by

$$\widetilde{A} = \begin{pmatrix} 0 & I \\ -A & -f(\langle b, .\rangle)b \end{pmatrix}, \tag{2.38}$$

$$D(\widetilde{A}) = \{\{y_0, y_1\} \in V \times H \ / \ y_1 \in V,\ Ay_0 + f(\langle b, y_1\rangle)b \in H\}. \tag{2.39}$$

It is sufficient to see that $\widetilde{A}$ is densely defined and maximal dissipative [6]. Since the functional b is not bounded on H, the space $\widetilde{V}$ defined by

$$\widetilde{V} = \{v \in V \ / \ \langle b, v\rangle = 0\} \tag{2.40}$$

is dense in H. Since $D(\widetilde{A}) \supset D(A) \times \widetilde{V}$, $D(\widetilde{A})$ is dense in $V \times H$. On the other hand, the condition Range $(I - \widetilde{A}) = V \times H$ can be reduced to

$$\text{Range } (I + A + f(\langle b, . \rangle\, b) = V'.$$

The latter follows from the coercivity of A and the assumptions on f. Furthermore, the monotonicity results from that of f. This completes the proof of Proposition 1.4. □

2.2. Stabilization.

2.2.1. Proof of Theorem 1.5. (iii) ⇒ (i). Recall that, for $\{y_0, y_1\} \in V \times H$, the strong ω-limit set of $\{y_0, y_1\}$, denoted by $\omega(\{y_0, y_1\})$, is the (possibly empty) set given by those $\{\varphi_0, \varphi_1\} \in V \times H$ such that there exists a sequence $t_n \to \infty$ as $n \to \infty$ for which the solution of (1.13) satisfies

$$\{y(t_n), y'(t_n)\} \to \{\varphi_0, \varphi_1\} \text{ (strongly) in } V \times H \text{ as } n \to \infty.$$

By the invariance principle of LaSalle and following the outline of [11], we can reduce the proof to the following lemmas.

LEMMA 2.2. *Consider the operator defined by* (2.38) *and* (2.39). *Then* $D(\widetilde{A})$ *is compactly embedded in* $V \times H$, *and* $(I - \widetilde{A})^{-1}$ *is compact from* $V \times H$ *into itself.*

Proof. As in the proof of Proposition 1.4, it is easy to see that $(I - \widetilde{A})^{-1}$ maps continuously $V \times H$ onto $D(\widetilde{A})$ and $H \times V'$ onto $V \times H$. Then we can conclude by noting that $V \times H$ is compactly embedded in $H \times V'$. □

LEMMA 2.3. *Suppose that* $\{y_0, y_1\} \in D(\widetilde{A})$. *Then* $\omega(\{y_0, y_1\})$ *is nonempty and satisfies* $\omega(\{y_0, y_1\}) \subset D(A) \times \widetilde{V}$.

Proof. From Lemma 2.2, $\omega(\{y_0, y_1\})$ is nonempty. On the other hand, it is standard that, for all $t \geq 0$, $\{y(t), y'(t)\} \in D(\widetilde{A})$ and (see, for instance, [6, p. 54])

$$\| \widetilde{A}\, \{y(t), y'(t)\} \|_{V \times H} \leq \| \widetilde{A}\, \{y_0, y_1\} \|_{V \times H} \tag{2.41}$$

so that, from the compactness of the embedding $D(\widetilde{A}) \subset V \times H$, it is easy to see that $\omega(\{y_0, y_1\}) \subset D(\widetilde{A})$. Furthermore, from (2.41), we get, for all $t \geq 0$,

$$\| y'(t) \|_V \leq \| \widetilde{A}\, \{y_0, y_1\} \|_{V \times H}. \tag{2.42}$$

Let us consider $\{\varphi_0, \varphi_1\} \in \omega(\{y_0, y_1\})$. Then there exists a sequence $t_n \to \infty$ as $n \to \infty$ such that, as $n \to \infty$,

$$y'(t_n) \to \varphi_1 \text{ in } H. \tag{2.43}$$

From (2.42), we deduce that there exists a subsequence, still denoted by $\{t_n\}_n$, such that, as $n \to \infty$,

$$y'(t_n) \to \varphi_1 \text{ weakly in } V \tag{2.44}$$

so that, as $n \to \infty$,

$$\langle b, y'(t_n) \rangle\, f(\langle b, y'(t_n) \rangle) \to \langle b, \varphi_1 \rangle\, f(\langle b, \varphi_1 \rangle). \tag{2.45}$$

On the other hand, we have, for all $t > 0$,

$$E(t) - E(0) + \int_0^t \langle b, y'(s) \rangle\, f(\langle b, y'(s) \rangle) ds = 0. \tag{2.46}$$

Therefore,

$$\lim_{n\to\infty}\int_{t_n}^{t+t_n}\langle b,y'(s)\rangle\, f(\langle b,y'(s)\rangle)ds=\lim_{n\to\infty}\int_0^t\langle b,y'(s+t_n)\rangle\, f(\langle b,y'(s+t_n)\rangle)ds=0.$$

Let us consider the solution of the equation

$$\begin{cases}\varphi''(t)+A\varphi(t)+f(\frac{d}{dt}\langle b,\varphi(t)\rangle)b=0,\\ \varphi(0)=\varphi_0,\ \varphi'(0)=\varphi_1.\end{cases}\tag{2.47}$$

Then the dominated convergence theorem gives

$$\int_0^t\langle b,\varphi'(s)\rangle\, f\left(\frac{d}{ds}\langle b,\varphi(s)\rangle\right)ds=0\quad\text{for all } t>0.$$

By continuity, this implies that

$$\langle b,\varphi'(t)\rangle\, f(\langle b,\varphi'(t)\rangle)=0\quad\text{for all } t\geq 0.$$

Hence, by assumption (1.27), we get

$$\langle b,\varphi'(t)\rangle=0\quad\text{for all } t\geq 0.\tag{2.48}$$

This completes the proof of the lemma. □

(i) ⇒ (ii). Suppose that condition (ii) is violated for some eigenvector $\widetilde{\psi}$ associated to the eigenvalue $\widetilde{\lambda}$. Then $y(t)=e^{i\sqrt{\widetilde{\lambda}}t}\widetilde{\psi}$ is a constant energy solution of (1.28).

(ii) ⇒ (iii). Suppose that, for $\{\varphi_0,\varphi_1\}\in D(A)\times V$, we have a solution φ satisfying (1.28). Then, for all $t\geq 0$,

$$\sum_n-\omega_n\langle\varphi_0,\psi_n\rangle\langle b,\psi_n\rangle\sin(\omega_n t)+\sum_n\langle\varphi_1,\psi_n\rangle\langle b,\psi_n\rangle\cos(\omega_n t)=0.\tag{2.49}$$

The proof can be reduced to the fact that (2.49) implies

$$\varphi_0=\varphi_1=0.\tag{2.50}$$

It is easy to see that the series in (2.49) is uniformly convergent in $(-\infty,\infty)$. From the uniqueness of the Fourier series expansion for almost periodic functions, we deduce that (2.49) implies, for all n,

$$\langle\varphi_0,\psi_n\rangle\langle b,\psi_n\rangle=\langle\varphi_1,\psi_n\rangle\langle b,\psi_n\rangle=0.\tag{2.51}$$

Thus condition (ii) gives (2.50).

The proof of Theorem 1.5 is complete. □

2.2.2. Proof of Theorem 1.6. For y solution of (1.13) and every $T>0$, the following equality can be deduced by a density argument

$$E(T)+\int_0^T\frac{d}{dt}\langle b,y(t)\rangle\, f\left(\frac{d}{dt}\langle b,y(t)\rangle\right)dt=E(0).\tag{2.52}$$

Let us consider the decomposition

$$y=\varphi+\psi,\tag{2.53}$$

where φ is the solution of

$$\begin{cases} \varphi''(t) + A\varphi(t) = 0, \\ \varphi(0) = y_0,\ \varphi'(0) = y_1, \end{cases} \tag{2.54}$$

and ψ is determined by the problem

$$\begin{cases} \psi'' + A\psi = -f(\frac{d}{dt}\langle b, y(t)\rangle)b, \\ \psi(0) = \psi'(0) = 0. \end{cases} \tag{2.55}$$

Then we have, for $T > \frac{2\pi}{\delta}$ [3],

$$\int_0^T \left| \frac{d}{dt}\langle b, \varphi(t)\rangle \right|^2 dt \geq CE(0) \tag{2.56}$$

so that

$$E(T) = -\int_0^T \frac{d}{dt}\langle b, y(t)\rangle\, f\left(\frac{d}{dt}\langle b, y(t)\rangle\right) dt + E(0)$$

$$\leq C\left\{\int_0^T \frac{d}{dt}\langle b, y(t)\rangle\, f\left(\frac{d}{dt}\langle b, y(t)\rangle\right) dt + \int_0^T \left|\frac{d}{dt}\langle b, \varphi(t)\rangle\right|^2 dt\right\}.$$

On the other hand, by applying (1.26) to system (2.55),

$$\int_0^T \left|\frac{d}{dt}\langle b, \varphi(t)\rangle\right|^2 dt \leq C\left\{\int_0^T \left|\frac{d}{dt}\langle b, \psi(t)\rangle\right|^2 dt + \int_0^T \left|\frac{d}{dt}\langle b, y(t)\rangle\right|^2 dt\right\}$$

$$\leq C\left\{\int_0^T \left|f(\frac{d}{dt}\langle b, y(t)\rangle)\right|^2 dt + \int_0^T \left|\frac{d}{dt}\langle b, y(t)\rangle\right|^2 dt\right\}.$$

Let us consider the decomposition $[0,T] = J_1 \cup J_2$ such that

$$J_1 = \left\{t \in [0,T]\ /\ \left|\frac{d}{dt}\langle b, y(t)\rangle\right| \geq 1\right\},\ J_2 = \left\{t \in [0,T]\ /\ \left|\frac{d}{dt}\langle b, y(t)\rangle\right| < 1\right\}.$$

Then, from (1.29), we get

$$\int_{J_1} \left|f\left(\frac{d}{dt}\langle b, y(t)\rangle\right)\right|^2 dt + \int_{J_1} \left|\frac{d}{dt}\langle b, y(t)\rangle\right|^2 dt \leq C\int_0^T f\left(\frac{d}{dt}\langle b, y(t)\rangle\right)\frac{d}{dt}\langle b, y(t)\rangle\, dt.$$

The hypothesis (1.30) implies

$$\int_{J_2} \left|f\left(\frac{d}{dt}\langle b, y(t)\rangle\right)\right|^2 dt + \int_{J_2} \left|\frac{d}{dt}\langle b, y(t)\rangle\right|^2 dt$$

$$\leq \int_{J_2} \left|f\left(\frac{d}{dt}\langle b, y(t)\rangle\right)\right|^2 dt + \int_{J_2} \left|\frac{d}{dt}\langle b, y(t)\rangle\right|^{2\rho} dt$$

$$\leq \int_0^T \xi\left(f\left(\frac{d}{dt}\langle b, y(t)\rangle\right)\frac{d}{dt}\langle b, y(t)\rangle\right) dt.$$

Applying Jensen's inequality, we obtain

$$\int_0^T \xi\left(f\left(\frac{d}{dt}\langle b,y(t)\rangle\right)\frac{d}{dt}\langle b,y(t)\rangle\right)dt \le T\xi\left(\frac{1}{T}\int_0^T f\left(\frac{d}{dt}\langle b,y(t)\rangle\right)\frac{d}{dt}\langle b,y(t)\rangle\,dt\right).$$

Therefore, for some positive constant C_0, which will be the one introduced in the theorem, we have

$$E(T) \le C_0\left\{\int_0^T f\left(\frac{d}{dt}\langle b,y(t)\rangle\right)\frac{d}{dt}\langle b,y(t)\rangle\,dt + \xi\left(\frac{1}{T}\int_0^T f\left(\frac{d}{dt}\langle b,y(t)\rangle\right)\frac{d}{dt}\langle b,y(t)\rangle\,dt\right)\right\}$$

so that

$$E(T) \le C_0 h(E(0)-E(T)), \tag{2.57}$$

where h is the function given by (1.32). Then the function p (inverse of $C_0 h$) is obviously increasing on $[0,+\infty)$, and (2.57) gives

$$E(T)+p(E(T)) \le E(0). \tag{2.58}$$

As the estimate above remains valid in successive intervals $[kT,(k+1)T]$, we have

$$E((k+1)T)+p(E((k+1)T)) \le E(kT), \quad k=0,1,2,\dots. \tag{2.59}$$

We now apply the result of Lemma 3.3 in [17].

LEMMA 2.4 (see [17]). *Let p denote a positive increasing function such that $p(0)=0$, and consider the function $q(s)=s-(I+p)^{-1}(s)$ and the sequence $\{s_k\}_k$ of positive numbers such that*

$$p(s_{k+1})+s_{k+1} \le s_k, \quad k \ge 0. \tag{2.60}$$

Then $s_k \le S(k)$, where $S(t)$ is the solution of

$$S'(t)+q(S(t))=0, \quad S(0)=s_0. \tag{2.61}$$

Moreover, if $p(s)>0$ for $s>0$, then $S(t)\to 0$ as $t\to\infty$.

Thus applying the lemma to the sequence $s_k=E(kT)$ yields

$$E(kT) \le S(k), \quad k=0,1,2,\dots. \tag{2.62}$$

For any $t>0$, we may write $t=kT+\tau$ for some integer k and $0\le\tau<T$ so that

$$E(t) \le E(kT) \le S(k) \le S\left(\frac{t-\tau}{T}\right) \le S\left(\frac{t}{T}-1\right) \quad \text{for } t>T.$$

This completes the proof of Theorem 1.6. □

2.2.3. Proof of Theorem 1.7. We first construct a function which fulfills the assumptions prescribed for ξ in Theorem 1.6. For $|r| < 1$, the assumptions (1.37) and (1.38) give

$$|r|^{2\beta} + |f(r)|^2 \leq \widetilde{c}_f^{-\frac{2\beta}{\alpha+1}} |rf(r)|^{\frac{2\beta}{\alpha+1}} + \widetilde{C}_f^2 |r|^{2\beta} \leq \widetilde{c}_f^{-\frac{2\beta}{\alpha+1}} (1 + \widetilde{C}_f^2) |rf(r)|^{\frac{2\beta}{\alpha+1}} . \tag{2.63}$$

Therefore, we can choose

$$\xi(s) = \widetilde{c}_f^{-\frac{2\beta}{\alpha+1}} (1 + \widetilde{C}_f^2) s^{\frac{2\beta}{\alpha+1}}, \tag{2.64}$$

and we may define the functions h (which gives p) and q given by (1.32) and (1.36), respectively. We conclude the proof by itemizing as follows.

(i) If $\alpha = \beta = 1$, then p and q have the forms

$$p(s) = c_p s, \ q(s) = \omega s$$

for some positive constants c_p, ω. Then, for some positive constant C, we easily get (1.39).

(ii) If $\alpha + 1 > 2\beta$, by setting $\gamma = \frac{\alpha+1}{2\beta}$ and noting that, for some positive constants C_1, C_2, we have

$$p\left(C_1 s + C_2 \left(\frac{s}{T} \right)^{\frac{1}{\gamma}} \right) = s, \tag{2.65}$$

we obtain

$$p(s) \sim C_p s^\gamma \ \ (s \to 0) \tag{2.66}$$

for some positive constant C_p. Furthermore, from $q(s + p(s)) = p(s)$, we get

$$q(s) \sim C_q s^\gamma \ \ (s \to 0) \tag{2.67}$$

for some positive constant C_q. Then the estimate (1.40) can be deduced from the following lemma, whose proof is given in the Appendix B.

LEMMA 2.5. *Let q denote a positive function such that*

$$q(s) \sim C_q s^\gamma \ \ (s \to 0), \quad \gamma > 1. \tag{2.68}$$

Then the solution of the differential equation (2.61) *satisfies*

$$S(t) = O(t^{\frac{1}{1-\gamma}}) \ \ (t \to \infty). \tag{2.69}$$

This completes the proof of Theorem 1.7. ∎

REMARK 2.1. *Under the assumption* (1.31), *it is easy to see that the solution of* (2.54) *satisfies* (2.56) *if and only if the first inequality in* (1.33) *holds. In other words, this condition is a characterization of the observability of the abstract sytem* (2.2) *with the output*

$$w(t) = \langle b, \varphi'(t) \rangle .$$

REMARK 2.2. *In the case $\alpha = \beta = 1$ and by adapting the results established in* [2], *we can see that the first inequality in* (1.33) *is necessary for the exponential decrease* (1.39) *to hold.*

2.2.4. Proof of Theorem 1.8. One of the ingredients of Theorem 1.8 is the following lemma, whose proof is given in Appendix C.

LEMMA 2.6. *Under the assumptions of Theorem* 1.8*, the solution of* (1.13) *satisfies*

$$\left\| \frac{d}{dt} \langle b, y(.) \rangle \right\|^2_{L^2(0,T)} \geq C(\|y_0\|^2_{V_{\frac{1}{2}-\mu}} + \|y_1\|^2_{D(A^\mu)'}) \tag{2.70}$$

for some positive constant C.

By using (2.52) and (2.70), it is easy to see that

$$E(T) \leq E(0) - C(\|y_0\|^2_{V_{\frac{1}{2}-\mu}} + \|y_1\|^2_{D(A^\mu)'}). \tag{2.71}$$

On the other hand, from the interpolation identities

$$V = D(A^{\frac{1}{2}}) = [D(A), V_{\frac{1}{2}-\mu}]_{\frac{1}{1+2\mu}}, \quad H = [D(A^{\frac{1}{2}}), D(A^\mu)']_{\frac{1}{1+2\mu}},$$

we get the following interpolation inequalities [19, p. 23]:

$$\|y_0\| \leq C \|y_0\|^{\frac{2\mu}{1+2\mu}}_{D(A)} \|y_0\|^{\frac{1}{1+2\mu}}_{V_{\frac{1}{2}-\mu}}, \tag{2.72}$$

$$|y_1| \leq C \|y_1\|^{\frac{2\mu}{1+2\mu}} \|y_1\|^{\frac{1}{1+2\mu}}_{D(A^\mu)'}. \tag{2.73}$$

Then we obtain

$$\|y_0\|^2_{V_{\frac{1}{2}-\mu}} \geq C \frac{\|y_0\|^{2(1+2\mu)}}{\|y_0\|^{4\mu}_{D(A)}}, \tag{2.74}$$

$$\|y_1\|^2_{D(A^\mu)'} \geq C \frac{|y_1|^{2(1+2\mu)}}{\|y_1\|^{4\mu}}. \tag{2.75}$$

Furthermore, it is easy to see that

$$\begin{aligned} \|y_0\|^2_{V_{\frac{1}{2}-\mu}} + \|y_1\|^2_{D(A^\mu)'} &\geq C \left\{ \frac{\|y_0\|^{2(1+2\mu)}}{\|y_0\|^{4\mu}_{D(A)}} + \frac{|y_1|^{2(1+2\mu)}}{\|y_1\|^{4\mu}} \right\} \\ &\geq C \frac{(E(0))^{1+2\mu}}{(\|y_0\|^2_{D(A)} + \|y_1\|^2)^{2\mu}}. \end{aligned}$$

Then the inequality (2.71), combined with the fact that $E(t)$ is nonincreasing, gives

$$E(T) \leq E(0) - C \frac{(E(T))^{1+2\mu}}{(\|y_0\|^2_{D(A)} + \|y_1\|^2)^{2\mu}}. \tag{2.76}$$

The estimate above remains valid in successive intervals $[kT, (k+1)T]$ so that

$$E((k+1)T) \leq E(kT) - C \frac{(E((k+1)T))^{1+2\mu}}{(\|y(kT)\|^2_{D(A)} + \|y'(kT)\|^2)^{2\mu}}, \quad k = 1, 2, \ldots . \tag{2.77}$$

As $\{y(t), y'(t)\}$ defines a (nonlinear) semigroup of contraction in $D(A)\times V$, the relation above gives [6, p. 54]

$$E((k+1)T) \le E(kT) - C\frac{(E((k+1)T))^{1+2\mu}}{(\|y_0\|^2_{D(A)} + \|y_1\|^2)^{2\mu}}, \quad k = 1, 2, \dots . \tag{2.78}$$

Let us consider the function $F(s) = \frac{s}{\|y_0\|^2_{D(A)}+\|y_1\|^2}$. Then relation (2.78) would read $F(E(k+1)T)) \le F(E(kT)) - C(F(E(k+1)T)))^{1+2\mu}$. By using Lemmas 2.4 and 2.5, we can easily deduce (1.42). This completes the proof of Theorem 1.8. □

3. Applications.

3.1. Beam equation with internal pointwise actuator. Consider the system (1.1)–(1.3). By the approach introduced in [14], one can obtain for this system the following state-space equation (see [20] for details):

$$y'' + \frac{\partial^4 y}{\partial x^4} = u(t)\delta(x-a), \quad t > 0,\ 0 < x < 1, \tag{3.1}$$

$$y(t,0) = \frac{\partial^2 y}{\partial x^2}(t,0) = \frac{\partial y}{\partial x}(t,1) = \frac{\partial^3 y}{\partial x^3}(t,1) = 0. \tag{3.2}$$

This formulation has the form (1.8) if we set

$$H = L^2(0,1),\ V = \left\{ v \in H^2(0,1) \ / \ v(0) = \frac{dv}{dx}(1) = 0 \right\}, \tag{3.3}$$

$$A = \frac{d^4}{dx^4},\ D(A) = \left\{ v \in H^4(0,1) \cap V \ / \ \frac{d^2v}{dx^2}(0) = \frac{d^3v}{dx^3}(1) = 0 \right\}, \tag{3.4}$$

$$b = \delta(x-a). \tag{3.5}$$

The eigenvalues $\left\{\lambda_n = \omega_n^2\right\}_n$ and the corresponding eigenfunctions $\{\psi_n\}_n$ are given by

$$\lambda_n = \left(-\frac{\pi}{2} + n\pi\right)^4, \quad n = 1, 2, \dots, \tag{3.6}$$

$$\psi_n(x) = \sin\left(-\frac{\pi}{2} + n\pi\right)x, \quad n = 1, 2, \dots . \tag{3.7}$$

On the other hand, the analogue of the feedback (1.11) is given by

$$u(t) = -f\left(\frac{d}{dt}y(t,a)\right). \tag{3.8}$$

Theorem 1.7 can be adapted to obtain the following stabilization result.

THEOREM 3.1. *Let f be a monotone function satisfying the assumptions of Theorem* 1.7. *Suppose that a is a rational number with coprime factorization*

$$a = \frac{a_1}{a_2} \tag{3.9}$$

such that a_1 is odd. Then, for any initial conditions $\{y_0, y_1\} \in V \times L^2(0,1)$, the solution of the feedback system defined by (3.1), (3.2), *and* (3.8) *satisfies, for some positive constants ω, K, T,*

$$\left\| \frac{\partial^2 y(t,.)}{\partial x^2} \right\|^2_{L^2(0,1)} + \|y'(t,.)\|^2_{L^2(0,1)} \leq K \left\{ \left\| \frac{d^2 y_0}{dx^2} \right\|^2_{L^2(0,1)} + \|y_1\|^2_{L^2(0,1)} \right\} e^{-\omega t} \quad \text{for all } t > T \tag{3.10}$$

if $\alpha = \beta = 1$ and

$$\left\| \frac{\partial^2 y(t,.)}{\partial x^2} \right\|^2_{L^2(0,1)} + \|y'(t,.)\|^2_{L^2(0,1)} = O(t^{-\frac{2\beta}{\alpha+1-2\beta}}) \ (t \to \infty) \tag{3.11}$$

if $\alpha + 1 > 2\beta$.

Proof. We have only to see that, for some positive constant c,

$$\left| \sin\left(-\frac{\pi}{2} + n\pi\right) a \right| \geq c > 0 \quad \text{for all } n. \tag{3.12}$$

This result has been established in [20]. □

3.2. Beam equation with internal piezoelectric actuator. Consider the system defined by (1.4) and (1.5). This system has the form (1.8) if we set

$$H = L^2(0,1),\ V = H^2(0,1) \cap H_0^1(0,1), \tag{3.13}$$

$$A = \frac{d^4}{dx^4},\ D(A) = \left\{ v \in H^4(0,1) \ / \ v(0) = \frac{d^2 v}{dx^2}(0) = v(1) = \frac{d^2 v}{dx^2}(1) = 0 \right\}, \tag{3.14}$$

$$b = \frac{d}{dx}(\delta(x - a_1) - \delta(x - a_2)). \tag{3.15}$$

The eigenvalues $\{\lambda_n = \omega_n^2\}_n$ and the corresponding eigenfunctions $\{\psi_n\}_n$ are given by

$$\lambda_n = (n\pi)^4, \quad n = 1, 2, \ldots, \tag{3.16}$$

$$\psi_n(x) = \sqrt{2} \sin n\pi x, \quad n = 1, 2, \ldots . \tag{3.17}$$

It is easy to see that $b \in D(A^{\frac{1}{2}})'$ and

$$\left| \left\langle \frac{d}{dx}(\delta(x - a_1) - \delta(x - a_2)), \psi_n \right\rangle \right| = \sqrt{2} n\pi \, |\cos n\pi a_1 - \cos n\pi a_2|$$

$$= 2\sqrt{2} n\pi \left| \sin\left(n\pi \frac{a_1 + a_2}{2} \right) \sin\left(n\pi \frac{a_1 - a_2}{2} \right) \right| \leq C \lambda_n^{\frac{1}{4}}.$$

From Corollary 1.2, we deduce that, for $\{y_0, y_1\} \in H_0^1(0,1) \times H^{-1}(0,1)$, the solution of (1.4) satisfies

$$y \in C(0, T; D(A^{\frac{1}{4}}) = H_0^1(0,1)) \cap C^1(0, T; D(A^{\frac{1}{4}})' = H^{-1}(0,1)).$$

Furthermore, an easy application of Theorem 1.5 gives the following result.

THEOREM 3.2. *Let f be a monotone continuous function satisfying* (1.27)*, and consider the feedback*

$$u(t) = -f\left[\frac{d}{dt}\left(\frac{\partial y(t,a_1)}{\partial x} - \frac{\partial y(t,a_2)}{\partial x}\right)\right]. \tag{3.18}$$

Then, for any $\{y_0, y_1\} \in D(A^{\frac{1}{2}}) \times L^2(0,1)$, the solution of the feedback system defined by (1.4)*,* (1.5)*, and* (3.18) *satisfies*

$$y \in C(0,T; D(A^{\frac{1}{2}}) = H^2(0,1) \cap H_0^1(0,1)) \cap C^1(0,T; L^2(0,1)). \tag{3.19}$$

Moreover, $\lim_{t\to\infty} \|\frac{\partial^2 y(t,.)}{\partial x^2}\|^2_{L^2(0,1)} + \|y'(t,.)\|^2_{L^2(0,1)} = 0$ *if and only if*

$$\frac{a_1 + a_2}{2}, \frac{a_1 - a_2}{2} \in IR \backslash IQ. \tag{3.20}$$

In order to get decay estimates related to the position of the actuators, we shall introduce some results from the theory of Diophantine approximations. Such results have been used to derive observability results for parabolic systems in [22], exact controllability properties for system (1.4) in [27], and decay estimate for strings and beams in [15] and [1], respectively.

For a real number θ, we denote by $|||\theta|||$ the difference, taken positively, between θ and the nearest integer, i.e.,

$$|||\theta||| = \min_{n\in Z} |\theta - z|. \tag{3.21}$$

An irrational number $\theta \in (0,1)$ is said to be of constant type if the sequence$\{\theta_n\}_n$, defined by the expansion of θ as a continuous fraction, is bounded. From [16], we quote the following result.

PROPOSITION 3.3. *An irrational number $\theta \in (0,1)$ is of constant type if and only if there exists a positive constant C such that*

$$|||n\theta||| \geq \frac{C}{n}, \quad n = 1, 2, \dots . \tag{3.22}$$

Furthermore, we shall use the following [7].

PROPOSITION 3.4. *For any $\epsilon > 0$, there exists a set $B_\epsilon \subset (0,1)$ having the Lebesgue measure equal to* 1 *and a positive constant C such that, for any $\theta \in B_\epsilon$,*

$$|||n\theta||| \geq \frac{C}{n^{1+\epsilon}}, \quad n = 1, 2, \dots . \tag{3.23}$$

REMARK 3.1. *The property* (3.22) *is satisfied if θ is irrational and is a root of a second degree polynomial with rational coefficients. The property* (3.23) *holds true if θ is an algebraic irrational* [7, p. 104].

Then we have the following stabilization result.

THEOREM 3.5. *Let f satisfy the assumptions of Theorem* 1.8*, and consider the feedback given by*

$$u(t) = -f\left[\frac{d}{dt}\left(\frac{\partial(A^{-1}y)}{\partial x}(t,a_1) - \frac{\partial(A^{-1}y)}{\partial x}(t,a_2)\right)\right]. \tag{3.24}$$

Then, for any $\{y_0, y_1\} \in L^2(0,1) \times D(A^{\frac{1}{2}})'$, *the feedback system defined by* (1.4) *and* (3.24) *admits a unique solution such that*

$$y \in C(0,T; L^2(0,1)) \cap C^1(0,T; D(A^{\frac{1}{2}})'). \tag{3.25}$$

Moreover, for any $\{y_0, y_1\} \in D(A^{\frac{1}{2}}) \times L^2(0,1)$, *we have*

$$\|y(t,.)\|^2_{L^2(0,1)} + \|y'(t,.)\|^2_{D(A^{\frac{1}{2}})'} = d(t)\{\|y(t,.)\|^2_{D(A^{\frac{1}{2}})} + \|y'(t,.)\|^2_{L^2(0,1)}\}, \tag{3.26}$$

where

$$d(t) = O(t^{-\frac{2}{3}}) \ \ (t \to \infty) \tag{3.27}$$

if $\frac{a_1+a_2}{2}$, $\frac{a_1-a_2}{2}$ *are of constant type and*

$$d(t) = O(t^{-\frac{2}{3+2\epsilon}}) \ \ (t \to \infty) \tag{3.28}$$

if $\frac{a_1+a_2}{2}, \frac{a_1-a_2}{2} \in B_\epsilon$.

Proof. Suppose that $\{y_0, y_1\} \in L^2(0,1) \times D(A^{\frac{1}{2}})'$, and consider

$$\widetilde{b} = A^{-\frac{1}{2}} \frac{d}{dx}(\delta(x-a_1) - \delta(x-a_2)) \in L^2(0,1) \tag{3.29}$$

and the change of variables

$$\begin{cases} \widetilde{y}(t) = A^{-\frac{1}{2}} y(t), \\ \widetilde{y}_0 = A^{-\frac{1}{2}} y_0, \\ \widetilde{y}_1 = A^{-\frac{1}{2}} y_1. \end{cases} \tag{3.30}$$

Then we have $\{\widetilde{y}_0, \widetilde{y}_1\} \in D(A^{\frac{1}{2}}) \times L^2(0,1)$, and, by using Proposition 1.4, the feedback system

$$\begin{cases} \widetilde{y}''(t) + A\widetilde{y}(t) + f(\frac{d}{dt}\langle \widetilde{b}, \widetilde{y}(t)\rangle)\widetilde{b} = 0, \\ \widetilde{y}(0) = \widetilde{y}_0, \ \widetilde{y}'(0) = \widetilde{y}_1, \end{cases} \tag{3.31}$$

admits a unique solution satisfying the analogous regularity to the one given by (1.20). This implies that the feedback system defined by (1.4) and (3.24) admits a unique solution satisfying (3.25). Moreover, we have

$$|\langle \widetilde{b}, \psi_n\rangle| = \frac{2\sqrt{2}}{n\pi} \left| \sin\left(n\pi \frac{a_1+a_2}{2}\right) \sin\left(n\pi \frac{a_1-a_2}{2}\right) \right| \leq \frac{2\sqrt{2}}{\pi} \quad \text{for all } n \tag{3.32}$$

and, if $\frac{a_1+a_2}{2}$ and $\frac{a_1-a_2}{2}$ are of constant type,

$$|\langle \widetilde{b}, \psi_n\rangle| \geq C\lambda_n^{-\frac{3}{4}} \quad \text{for all } n. \tag{3.33}$$

Hence, under the assumption $\{y_0, y_1\} \in D(A^{\frac{1}{2}}) \times L^2(0,1)$, we have $\{\widetilde{y}_0, \widetilde{y}_1\} \in D(A) \times D(A^{\frac{1}{2}})$ so that, by using Theorem 1.8, the solution of (3.31) satisfies

$$\begin{cases} \|\widetilde{y}(t)\|^2_{D(A^{\frac{1}{2}})} + \|\widetilde{y}'(t)\|^2_{L^2(0,1)} = d(t)\{\|\widetilde{y}_0\|^2_{D(A)} + \|\widetilde{y}_1\|^2_{D(A^{\frac{1}{2}})}\}, \\ d(t) = O(t^{-\frac{2}{3}}) \ \ (t \to \infty). \end{cases} \tag{3.34}$$

This gives (3.26) and (3.27). The remaining part of the proof can be obtained by similar arguments. This ends the proof of Theorem 3.5. ∎

3.3. Beam equation with concentrated actuator. Consider the system defined by (1.6) with the boundary conditions (1.5). This system has the form (1.8) by setting (3.13), (3.14), and $b = g \in L^2(0,1)$. The appropriate eigenvalues and the corresponding eigenfunctions are given by (3.16) and (3.17). Furthermore, it is standard that, for any $\{y_0, y_1\} \in D(A^{\frac{1}{2}}) \times L^2(0,1)$, there exists a unique solution to (1.5)–(1.6) satisfying (3.19). On the other hand, applying Theorem 1.6, we can obtain the following theorem.

THEOREM 3.6. *Let f be a monotone continuous function satisfying* (1.27)*, and consider the feedback*

$$u(t) = -f\left(\int_0^1 g(x)y'(t,x)dx\right). \tag{3.35}$$

Then, for any $\{y_0, y_1\} \in D(A^{\frac{1}{2}}) \times L^2(0,1)$, the solution of the feedback system defined by (1.5), (1.6)*, and* (3.35) *satisfies* (3.19)*. Moreover,* $\lim_{t\to\infty} \|\frac{\partial^2 y(t,.)}{\partial x^2}\|^2_{L^2(0,1)} + \|y'(t,.)\|^2_{L^2(0,1)} = 0$ *if and only if*

$$\int_0^1 g\psi_n dx \neq 0 \quad \textit{for all } n. \tag{3.36}$$

Before stating a stabilization result which improves the theorem above, let us mention the following regularity result.

PROPOSITION 3.7. *Suppose that, for some positive constant C_g,*

$$|\langle g, \psi_n\rangle| = \left|\int_0^1 g\psi_n dx\right| \leq \frac{C_g}{n^2} \quad \textit{for all } n. \tag{3.37}$$

Then, for any $\{y_0, y_1\} \in D(A) \times D(A^{\frac{1}{2}})$, the system (1.5)–(1.6) *admits a unique solution such that*

$$y \in C(0,T;D(A)) \cap C^1(0,T;D(A^{\frac{1}{2}})). \tag{3.38}$$

Proof. If we set

$$\widetilde{b} = A^{\frac{1}{2}} g \tag{3.39}$$

and consider the change of variables

$$\begin{cases} \widetilde{y}(t) = A^{\frac{1}{2}} y(t), \\ \widetilde{y}_0 = A^{\frac{1}{2}} y_0, \\ \widetilde{y}_1 = A^{\frac{1}{2}} y_1, \end{cases} \tag{3.40}$$

then $\widetilde{b} \in D(A^{\frac{1}{2}})'$, and, for some positive constant $\widetilde{C}_g$,

$$|\langle \widetilde{b}, \psi_n\rangle| \leq \widetilde{C}_g \quad \text{for all } n \tag{3.41}$$

so that, by Proposition 1.1, we have

$$\widetilde{y} \in C(0,T;D(A^{\frac{1}{2}})) \cap C^1(0,T;L^2(0,1)). \tag{3.42}$$

This yields (3.38). ∎

Then we have the following strong stabilization result.

THEOREM 3.8. *Let f be a monotone continuous function satisfying* (1.27)*, and consider the feedback*

$$u(t) = -f\left[\frac{d}{dt}\left(\int_0^1 g(x)\frac{\partial^4 y(t,x)}{\partial x^4}dx\right)\right]. \tag{3.43}$$

Then, for any $\{y_0, y_1\} \in D(A)\times D(A^{\frac{1}{2}})$, the feedback system defined by (1.5)–(1.6) *and* (3.43) *admits a unique solution satisfying* (3.38)*. Moreover,* $\lim_{t\to\infty} \|\frac{\partial^4 y(t,.)}{\partial x^4}\|^2_{L^2(0,1)} + \|\frac{\partial^2 y'(t,.)}{\partial x^2}\|^2_{L^2(0,1)} = 0$ *if and only if* (3.36) *holds.*

Proof. Let us consider the change of variables defined by (3.40). Then the solution of the feedback system

$$\begin{cases} \widetilde{y}''(t) + A\widetilde{y}(t) + f(\frac{d}{dt}\langle\widetilde{b}, \widetilde{y}(t)\rangle)\widetilde{b} = 0, \\ \widetilde{y}(0) = \widetilde{y}_0,\ \widetilde{y}'(0) = \widetilde{y}_1, \end{cases} \tag{3.44}$$

satisfies (3.42). Moreover, $\lim_{t\to\infty} \|\widetilde{y}(t)\|^2_{D(A^{\frac{1}{2}})} + \|\widetilde{y}'(t)\|^2_{L^2(0,1)} = 0$ if and only if

$$\langle\widetilde{b}, \psi_n\rangle = \omega_n \int_0^1 g\psi_n dx \neq 0 \quad \text{for all } n. \tag{3.45}$$

This amounts to saying that there exists a unique solution to the feedback system defined by (1.5), (1.6), and (3.43) and that this solution satisfies the stability of the theorem if and only if (3.36) holds. □

From Theorem 1.8, we can easily deduce the following nonuniform stabilization result.

THEOREM 3.9. *Let f satisfy the assumptions of Theorem* 1.8*. Suppose that, for some positive constant c_g and some $\frac{1}{2} < \mu \leq 4$, we have*

$$\left|\int_0^1 g\psi_n dx\right| \geq \frac{c_g}{n^\mu} \quad \textit{for all } n. \tag{3.46}$$

Then, for any $\{y_0, y_1\} \in D(A) \times D(A^{\frac{1}{2}})$, the solution of the feedback system defined by (1.5)*,* (1.6)*, and* (3.35) *satisfies*

(3.47)

$$\begin{cases} \left\|\frac{\partial^2 y(t,.)}{\partial x^2}\right\|^2_{L^2(0,1)} + \|y'(t,.)\|^2_{L^2(0,1)} = d(t)\left\{\left\|\frac{d^4 y_0}{dx^4}\right\|^2_{L^2(0,1)} + \left\|\frac{d^2 y_1}{dx^2}\right\|^2_{L^2(0,1)}\right\}, \\ d(t) = O(t^{-\frac{2}{\mu}}) \ \ (t \to \infty). \end{cases}$$

Furthermore, Theorem 1.7 enables us to get uniform stabilization with explicit decay estimate for the feedback system defined by (1.5), (1.6), and (3.43).

THEOREM 3.10. *Let f satisfy the assumptions of Theorem* 1.7*. Suppose that, for some positive constants c_g, C_g,*

$$0 < \frac{c_g}{n^2} \leq \left|\int_0^1 g\psi_n dx\right| \leq \frac{C_g}{n^2} \quad \textit{for all } n. \tag{3.48}$$

Then, for some positive constants ω, K, T *and any* $\{y_0, y_1\} \in D(A) \times D(A^{\frac{1}{2}})$, *the solution of the feedback system defined by* (1.5), (1.6), *and* (3.43) *satisfies*

(3.49)

$$\left\| \frac{\partial^4 y(t,.)}{\partial x^4} \right\|^2_{L^2(0,1)} + \left\| \frac{\partial^2 y'(t,.)}{\partial x^2} \right\|^2_{L^2(0,1)} \leq K \left\{ \left\| \frac{d^4 y_0}{dx^4} \right\|^2_{L^2(0,1)} + \left\| \frac{d^2 y_1}{dx^2} \right\|^2_{L^2(0,1)} \right\} e^{-\omega t}$$
$$\text{for all } t > T$$

if $\alpha = \beta = 1$ *and*

$$\left\| \frac{\partial^4 y(t,.)}{\partial x^4} \right\|^2_{L^2(0,1)} + \left\| \frac{\partial^2 y'(t,.)}{\partial x^2} \right\|^2_{L^2(0,1)} = O(t^{-\frac{2\beta}{\alpha+1-2\beta}}) \ (t \to \infty) \tag{3.50}$$

if $\alpha + 1 > 2\beta$.

4. Further extensions and related questions. As mentioned in the introduction, our results deduced from the study of the abstract model may be applied to various hyperbolic-like systems. To illustrate this, we shall consider the plate and the membrane equations with point controls. In what follows, Ω will denote an open bounded domain in IR^2 with sufficiently smooth boundary Γ. For each operator A defined in the examples below, we shall use the same notation relative to the eigenvalues and the corresponding eigenfunctions introduced in subsection 1.3. Every eigenvalue will be supposed simple. Here a is a given point in Ω.

4.1. Plate equation with internal point control. Consider the system

$$\begin{cases} y'' + \Delta^2 y = u(t)\delta(x-a) \ \text{ in } (0,\infty) \times \Omega, \\ y = \Delta y = 0 \ \text{ on } (0,\infty) \times \Gamma, \\ y(0,x) = y_0(x), \ y'(0,x) = y_1(x) \ \text{ in } \Omega. \end{cases} \tag{4.1}$$

This system has the form (1.8) if we set

$$A = \Delta^2, \ D(A) = \left\{ v \in H^4(\Omega) \ / \ y = \Delta y = 0 \text{ on } \Gamma \right\}, \tag{4.2}$$

$$H = L^2(\Omega), \ V = D(A^{\frac{1}{2}}) = H^2(\Omega) \cap H_0^1(\Omega), \tag{4.3}$$

$$b = \delta(x-a) \in V'. \tag{4.4}$$

Then, for f satisfying the assumptions of Theorem 1.5 and any $\{y_0, y_1\} \in H^2(\Omega) \cap H_0^1(\Omega) \times L^2(\Omega)$, the solution of the feedback system defined by (4.1) and

$$u(t) = -f\left(\frac{d}{dt} y(t,a) \right) \tag{4.5}$$

satisfies $\lim_{t\to\infty} \|\Delta y(t,.)\|^2_{L^2(\Omega)} + \|y'(t,.)\|^2_{L^2(\Omega)} = 0$ if and only if

$$\psi_n(a) \neq 0 \quad \text{for all } n. \tag{4.6}$$

4.2. Wave equation with internal point control. Consider the system

$$\begin{cases} y'' - \Delta y = u(t)\delta(x-a) \ \text{ in } (0,\infty)\times\Omega, \\ y = 0 \ \text{ on } (0,\infty)\times\Gamma, \\ y(0,x) = y_0(x),\ y'(0,x) = y_1(x) \ \text{ in } \Omega. \end{cases} \tag{4.7}$$

If we consider the setting defined by

$$A = -\Delta,\ D(A) = H^2(\Omega)\cap H_0^1(\Omega), \tag{4.8}$$

$$H = L^2(\Omega),\ V = D(A^{\frac{1}{2}}) = H_0^1(\Omega), \tag{4.9}$$

then we have $\delta(x-a)\notin D(A^{\frac{1}{2}})'$. Furthermore, we cannot readily apply the results of subsection 1.3 since, from [18], [24], and [25], we have the following sharp regularity result:

$$y\in C(0,T;D(A^{\frac{1}{4}}))\cap C^1(0,T;D(A^{\frac{1}{4}})').$$

However, exploiting the fact that, for $\epsilon>0$ arbitrarily small, $A^{-\epsilon}\delta(x-a)\in D(A^{\frac{1}{2}})'$, we can consider the change of variables

$$\begin{cases} \widetilde{y}(t) = A^{-\epsilon}y(t), \\ \widetilde{y}_0 = A^{-\epsilon}y_0, \\ \widetilde{y}_1 = A^{-\epsilon}y_1 \end{cases} \tag{4.10}$$

and the auxiliary feedback system

$$\begin{cases} \widetilde{y}''(t) + A\widetilde{y}(t) + f(\frac{d}{dt}\langle\widetilde{b},\widetilde{y}(t)\rangle)\widetilde{b} = 0, \\ \widetilde{y}(0) = \widetilde{y}_0,\ \widetilde{y}'(0) = \widetilde{y}_1, \end{cases} \tag{4.11}$$

where $\widetilde{b} = A^{-\epsilon}\delta(x-a)$. Then Theorem 1.5 may be applied to obtain that, for any $\{y_0,y_1\}\in D(A^{\frac{1}{2}-\epsilon})\times D(A^{\epsilon})'$ and for f satisfying the assumptions of Theorem 1.5, the feedback system defined by (4.7) and

$$u(t) = -f\left(\frac{d}{dt}\left\langle A^{-\epsilon}\delta(x-a), A^{-\epsilon}y(t)\right\rangle\right) \tag{4.12}$$

admits a unique solution $y\in C(0,T;D(A^{\frac{1}{2}-\epsilon}))\cap C^1(0,T;D(A^{\epsilon})')$. Moreover, $\lim_{t\to\infty}\|y(t)\|^2_{D(A^{\frac{1}{2}-\epsilon})} + \|y'(t)\|^2_{D(A^{\epsilon})'} = 0$ if and only if the analogous condition to (4.6) holds.

Appendix A.

LEMMA A.1. *For $v\in L^2(0,T)$, consider*

$$v_k = \int_0^T v(t)e^{-i\omega_k t}dt, \quad 1\le k\le m. \tag{A.1}$$

Then, for any $v\in L^2(0,T)$, we have

$$\sum_{k=1}^{m}|v_k|^2 \le C\|v\|^2_{L^2(0,T)} \tag{A.2}$$

for some positive constant C independent of m.

Proof. For $w = (w_1, \dots, w_m) \in IR^m$, we introduce

$$W_m(t) = \sum_{k=1}^{m} w_k e^{i\omega_k t}. \tag{A.3}$$

Then the following holds:

$$\left| \int_0^T v(t)\overline{W_m(t)}dt \right| = \left| \sum_{k=1}^{m} v_k w_k \right| \le \|v\|_{L^2(0,T)} \|W_m\|_{L^2(0,T)}.$$

On the other hand, for some positive constant C independent of m, we have [3]

$$\|W_m\|^2_{L^2(0,T)} \le C \sum_{k=1}^{m} |w_k|^2.$$

This yields (A.2). ∎

Appendix B. Proof of Lemma 2.5. Let us consider the function

$$g(s) = \int_s^\zeta \frac{d\tau}{q(\tau)}, \quad 0 < s < \zeta. \tag{B.1}$$

Then g is a decreasing function and $g(\zeta) = 0$, $g(0+) = +\infty$. Thus $[0, +\infty)$ is in the range of g, and the solution of (2.61) is given by

$$X(t) = g^{-1}(t), \quad t \ge 0. \tag{B.2}$$

Since $g(0+) = +\infty$,

$$\lim_{t\to\infty} X(t) = \lim_{t\to\infty} g^{-1}(t) = 0.$$

Let $0 < \epsilon < 1$. There exists $\delta(\epsilon) > 0$ such that, if $0 < s < \delta(\epsilon)$,

$$|q(s) - C_q s^\gamma| < \epsilon C_q s^\gamma. \tag{B.3}$$

Moreover, there exists $t_0(\epsilon) > 0$ such that $0 < X(t) < \delta(\epsilon)$ for $t \ge t_0(\epsilon)$. Therefore, if $t \ge t_0(\epsilon)$, we have

$$-q(X(t)) \le C_q(\epsilon - 1)(X(t))^\gamma. \tag{B.4}$$

Hence

$$X'(t) + C_q(1 - \epsilon)(X(t))^\gamma \le 0, \quad t \ge t_0(\epsilon). \tag{B.5}$$

This yields (2.69) and completes the proof of the lemma. ∎

Appendix C. Proof of Lemma 2.6. Let us consider the decomposition defined by (2.53), (2.54), and (2.55). Then the assumptions (1.31) and (1.41) give, for some positive constant C [3],

$$\int_0^T \left| \frac{d}{dt} \langle b, \varphi(t) \rangle \right|^2 dt \ge C(\|\varphi_0\|^2_{V_\mu} + \|\varphi_1\|^2_{D(A^\mu)'}). \tag{C.1}$$

On the other hand, Proposition 1.3 implies

$$\int_0^T \left| \frac{d}{dt}(\langle b, \varphi(t)\rangle - \langle b, y(t)\rangle)\right|^2 dt \leq C \int_0^T \left| \frac{d}{dt} \langle b, y(t)\rangle \right|^2 dt \tag{C.2}$$

so that, for some positive constant C,

$$\left\| \frac{d}{dt} \langle b, \varphi(.)\rangle \right\|_{L^2(0,T)} \leq C \left\| \frac{d}{dt} \langle b, y(.)\rangle \right\|_{L^2(0,T)}. \tag{C.3}$$

This yields (2.70). □

REFERENCES

[1] K. Ammari and M. Tucsnak, *Stabilization of Bernoulli–Euler beams by means of a pointwise feedback force*, SIAM J. Control Optim., 39 (2000), pp. 1160–1181.

[2] K. Ammari and M. Tucsnak, *Stabilization of second order evolution equations by a class of unbounded feedback*, ESAIM Control Optim. Calc. Var., 6 (2001), pp. 361–386.

[3] J. Ball and M. Slemrod, *Nonharmonic Fourier series and stabilization of distributed semilinear control systems*, Comm. Pure Appl. Math., 32 (1979), pp. 555–587.

[4] H. T. Banks, W. Fang, R. J. Silcox, and R. C. Smith, *Approximation methods for control of the acoustic/structure interaction with pizoceramic actuators*, J. Intelligent Material Systems and Structures, 4 (1993), pp. 98–116.

[5] A. Bensoussan, G. Da Prato, M. C. Delfour, and S. K. Mitter, *Representation and Control of Infinite-Dimensional Systems*, *Vol.* 1, Birkhäuser Boston, Boston, 1992.

[6] H. Brezis, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, North–Holland Math. Stud. 5, North–Holland, Amsterdam, 1973.

[7] J. W. Cassals, *An Introduction to Diophantine Approximations*, Cambridge University Press, Cambridge, UK, 1966.

[8] G. Chen, M. C. Delfour, A. M. Krall, and G. Payre, *Modelling, stabilization and control of serially connected beams*, SIAM J. Control Optim., 25 (1987), pp. 526–546.

[9] F. Conrad, J. Leblond, and J.-P. Marmorat, *Stabilization of second order evolution equations by unbounded nonlinear feedback*, in Proceedings of the Fifth IFAC Symposium on Control of Distributed Parameter Systems, Perpignan, France, 1989, pp. 111–116.

[10] E. F. Crawley and E. H. Anderson, *Detailed models for piezoceramic actuators for beams*, J. Intelligent Material Systems and Structures, 1 (1990), pp. 79–83.

[11] C. Dafermos and M. Slemrod, *Asymptotic behaviour of nonlinear contraction semigroups*, J. Funct. Anal., 13 (1973), pp. 97–106.

[12] M. C. Delfour, J. Lagnese, and M. P. Polis, *Stabilization of hyperbolic systems using concentrated actuators and sensors*, IEEE Trans. Automat. Control, 31 (1986), pp. 1091–1096.

[13] P. Grisvard, *Caractérisation de quelques espaces d'interpolation*, Arch. Ration. Mech. Anal., 25 (1967), pp. 40–63.

[14] L. F. Ho and D. L. Russel, *Admissible input elements for systems in Hilbert space and a Carleson measure criterion*, SIAM J. Control Optim., 21 (1983), pp. 614–640.

[15] S. Jaffard, M. Tucsnak, and E. Zuazua, *Singular internal stabilization of the wave equation*, J. Differential Equations, 145 (1998), pp. 184–215.

[16] S. Lang, *Introduction to Diophantine Approximations*, Addison–Wesley, New York, 1966.

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

[18] J.-L. Lions, *Pointwise control for distributed systems*, in Control and Estimation in Distributed Parameter Systems, H. T. Banks, ed., SIAM, Philadelphia, 1992, pp. 1–41.

[19] J.-L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, *Vol.* 1, Dunod, Paris, 1967.

[20] R. Rebarber, *Exponential stability of coupled beams with dissipative joints: A frequency domain approach*, SIAM J. Control Optim., 33 (1995), pp. 1–28.

[21] D. L. Russel, *Decay rates for weakly damped systems in Hilbert space obtained with control theoretic methods*, J. Differential Equations, 19 (1975), pp. 344–370.

[22] Y. Sakawa, *Observability and related problems for partial differential equations of parabolic type*, SIAM J. Control, 13 (1975), pp. 14–27.

[23] J. Simon, *Compact sets in the space* $L^p(0,T;B)$, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65–96.

[24] R. Triggiani, *Regularity with interior point control.* I. *Wave and Euler-Bernoulli equations*, in Boundary Control and Boundary Variation (Sophia-Antipolis, 1990), Lecture Notes in Control and Inform. Sci. 178, J.-P. Zolezio, ed., Springer-Verlag, Berlin, 1992, pp. 321–355.

[25] R. Triggiani, *Interior and boundary regularity of the wave equation with interior point control*, Differential Integral Equations, 6 (1993), pp. 111–129.

[26] M. Tucsnak, *Contrôle d'une poutre avec actionneur piézoélectrique*, C. R. Acad. Sci. Paris Sér. I Math., 319 (1994), pp. 697–702.

[27] M. Tucsnak, *Regularity and exact controllability for a beam with piezoelectric actuator*, SIAM J. Control Optim., 34 (1996), pp. 922–930.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 4, pp. 1194–1211

AN INVERSE INITIAL BOUNDARY VALUE PROBLEM FOR THE WAVE EQUATION IN THE PRESENCE OF IMPERFECTIONS OF SMALL VOLUME*

HABIB AMMARI†

Dedicated to Jean-Claude Nédélec for his 60th birthday

Abstract. We consider for the wave equation the inverse problem of identifying locations and certain properties of the shapes of small conductivity inhomogeneities in a homogeneous background medium from dynamic boundary measurements on part of the boundary and for a finite interval in time. Using as weights particular background solutions constructed by a geometrical control method, we develop an asymptotic method based on appropriate averaging of the partial dynamic boundary measurements. Our approach is expected to lead to very effective computational identification algorithms.

Key words. inverse problem, wave equation, reconstruction, geometric control

AMS subject classifications. 35R30, 35B40, 35B37, 35L05

PII. S0363012901384247

1. The inverse initial boundary value problem. Let Ω be a bounded, smooth subdomain of $\mathbf{R}^2$. For simplicity, we take $\partial\Omega$ to be $\mathcal{C}^\infty$, but this condition could be considerably weakened. Let n denote the outward unit normal to $\partial\Omega$. We suppose that Ω contains a finite number of inhomogeneities, each of the form $z_j + \alpha B_j$, where $B_j \subset \mathbf{R}^2$ is a bounded, smooth domain containing the origin. The total collection of inhomogeneities thus takes the form $\mathcal{B}_\alpha = \cup_{j=1}^m (z_j + \alpha B_j)$. The points $z_j \in \Omega, j = 1, \ldots, m$ that determine the location of the inhomogeneities are assumed to satisfy

$$|z_j - z_l| \geq d_0 > 0 \quad \forall\, j \neq l \quad \text{and} \quad \operatorname{dist}(z_j, \partial\Omega) \geq d_0 > 0 \quad \forall\, j.$$

As a consequence of this assumption, it follows immediately that $m \leq \frac{4|\Omega|}{\pi d_0^2}$. We also assume that $\alpha > 0$, the common order of magnitude of the diameters of the inhomogeneities, is sufficiently small so that these are disjoint and their distance to $\mathbf{R}^2 \setminus \overline{\Omega}$ is larger than $d_0/2$. Let γ_0 denote the conductivity of the background medium; for simplicity, we shall assume in this paper that it is constant. Let γ_j denote the constant conductivity of the jth inhomogeneity, $z_j + \alpha B_j$. Using this notation, we introduce the piecewise constant conductivity

$$\gamma_\alpha(x) = \begin{cases} \gamma_0, & x \in \Omega \setminus \overline{\mathcal{B}_\alpha}, \\ \gamma_j, & x \in z_j + \alpha B_j,\ j = 1, \ldots, m. \end{cases}$$

*Received by the editors January 29, 2001; accepted for publication (in revised form) April 12, 2002; published electronically October 29, 2002. This work is partially supported by ACI Jeunes Chercheurs (0693) from the Ministry of Education and Scientific Research, France.
http://www.siam.org/journals/sicon/41-4/38424.html

†Centre de Mathématiques Appliquées, CNRS UMR 7641 & École Polytechnique, 91128 Palaiseau Cedex, France (ammari@cmapx.polytechnique.fr).

Consider the initial boundary value problem for the (scalar) wave equation

$$\text{(1)}\qquad \begin{cases} (\partial_t^2 - \text{ div } \gamma_\alpha \,\text{grad})u_\alpha = 0 \quad \text{in } \Omega \times (0,T), \\ u_\alpha|_{t=0} = \varphi, \partial_t u_\alpha|_{t=0} = \psi \quad \text{in } \Omega, \\ u_\alpha|_{\partial\Omega\times(0,T)} = f. \end{cases}$$

Define u to be the solution of the wave equation in the absence of any inhomogeneities. Thus u satisfies

$$\text{(2)}\qquad \begin{cases} (\partial_t^2 - \gamma_0\Delta)u = 0 \quad \text{in } \Omega \times (0,T), \\ u|_{t=0} = \varphi, \partial_t u|_{t=0} = \psi \quad \text{in } \Omega, \\ u|_{\partial\Omega\times(0,T)} = f. \end{cases}$$

Here $T > 0$ is a final observation time, and the initial conditions $\varphi, \psi \in \mathcal{C}^\infty(\overline{\Omega})$ and the boundary condition $f \in \mathcal{C}^\infty(0,T;\mathcal{C}^\infty(\partial\Omega))$ are subject to the compatibility conditions

$$\partial_t^{2l} f|_{t=0} = (\gamma_0)^l(\Delta^l\varphi)|_{\partial\Omega} \text{ and } \partial_t^{2l+1} f|_{t=0} = (\gamma_0)^l(\Delta^l\psi)|_{\partial\Omega}, \quad l = 1, 2, \ldots.$$

From the above compatibility conditions on φ, ψ and f, it follows that the initial boundary value problem (2) has a unique solution in $\mathcal{C}^\infty([0,T]\times\overline{\Omega})$; see [14]. It is also classical to prove that the transmission problem for the wave equation (1) has a unique weak solution $u_\alpha \in \mathcal{C}^0(0,T;H^1(\Omega)) \cap \mathcal{C}^1(0,T;L^2(\Omega))$; see, for example, [19]. Indeed, Lions proved in [19, Chapter VI, Theorem 4.1, p. 369] that $\frac{\partial u_\alpha}{\partial n}|_{\partial\Omega}$ belongs to $L^2(0,T;L^2(\partial\Omega))$. His proof is based on an extension of the multiplier method.

Throughout this paper, we shall use quite standard L^2-based Sobolev spaces to measure regularity. The notation H^s is used to denote those functions which, along with all their derivatives of order less than or equal to s, are in L^2. H_0^1 denotes the closure of $\mathcal{C}_0^\infty$ in the norm of H^1. Sobolev spaces with negative indices are in general defined by duality, using an L^2-inner product. We shall only need two such spaces, namely, H^{-1}, which is defined as the dual of H_0^1, and H^{-2}, which is defined as the dual of H_0^2 that is the closure of $\mathcal{C}_0^\infty$ in the norm of H^2.

Define ν_j to be the outward unit normal to $\partial(z_j + \alpha B_j)$ for $j = 1, \ldots, m$. Let $\Gamma \subset \Omega$ be a given part of the boundary $\partial\Omega$. The aim of this paper is to identify the location and certain properties of the shapes of the inhomogeneities $\mathcal{B}_\alpha$ from only knowledge of boundary measurements of

$$\frac{\partial u_\alpha}{\partial n} \quad \text{on } \Gamma \times (0,T),$$

i.e., on the part Γ of the boundary $\partial\Omega$ and on the finite interval in time $(0,T)$. For this purpose, we develop an asymptotic method based on appropriate averaging, using particular background solutions as weights. These particular solutions are constructed by a control method as was done in the original work [32].

The first fundamental step in the design of our reconstruction method is the derivation of an asymptotic formula for $\frac{\partial u_\alpha}{\partial \nu_j}|_{\partial(z_j+\alpha B_j)^+}$ in terms of the reference solution u, the location z_j of the imperfection $z_j + \alpha B_j$, and the geometry of B_j. The second step consists of the use of this asymptotic formula to derive integral boundary formulae with a convenient choice of test functions, which is based on a geometrical control method and solving Volterra-type integral equations. We expect that these

boundary integral formulae will form the basis of very effective computational identifying algorithms. A similar approach may be applied to the full (time-dependent) Maxwell equations with small inhomogeneities of different electric permittivity or magnetic permeability (or both). This will be discussed in a forthcoming paper. The elastodynamic inverse problem will also be considered.

Whereas the determination of conductivity profiles from knowledge of boundary measurements has received a great deal of attention (see, for example, [1], [4], [9], [12], [15], and [34]), the reconstruction of imperfections within dynamics is much less investigated. To the best of our knowledge, the present paper is the first attempt to design an effective method to determine the location and the size of small conductivity imperfections inside a homogeneous medium from the dynamical measurements on part of the boundary.

The inverse problem considered in this paper is more complicated from the mathematical point of view and more interesting in applications than the one solved in [4] and [34] because, in many applications, one cannot get measurements for all t or on the whole boundary, and so one cannot, by taking a Fourier transform in the time variable, reduce our dynamic inverse problem to the inverse problem for the Helmholtz equation considered in [4] and [34].

The general approach we will take to recuperate the locations and shapes of the imperfections is to integrate the solution against special test functions. Our method is quite similar to the ideas used (in the time-independent case) by Calderón [11] in his proof of uniqueness of the linearized conductivity problem and later by Sylvester and Uhlmann in their important work [29] on uniqueness of the three-dimensional inverse conductivity problem (see Nachman [21] for the two-dimensional problem). It is also closely related to ideas used by Yamamoto in his original work [32] on inverse source hyperbolic problems and by Rakesh and Symes [27]. For discussions on other interesting inverse source hyperbolic problems, the reader is referred, for example, to Isakov [17], Belishev and Kurylev [8], Romanov and Kabanikhin [28], Yamamoto [31], [33], Puel and Yamamoto [23], [24], [25], [26], Grasselli and Yamamoto [16], Bruckner and Yamamoto [10], Nicaise [22], and Sun [30].

2. An energy estimate. We start the derivation of the asymptotic formula for $\frac{\partial u_\alpha}{\partial \nu_j}|_{\partial(z_j+\alpha B_j)^+}$ with the following energy estimate of $u_\alpha - u$.

PROPOSITION 2.1. *There exist constants $0 < \alpha_0$ and C such that, for $0 < \alpha < \alpha_0$, the following energy estimate holds:*

$$||\partial_t(u_\alpha - u)||_{L^\infty(0,T;H^{-1}(\Omega))} + ||u_\alpha - u||_{L^\infty(0,T;L^2(\Omega))} \leq C\alpha. \tag{3}$$

The constants α_0 and C depend on the domains $\{B_j\}_{j=1}^m$, the domain Ω, d_0, T, γ_0, $\{\gamma_j\}_{j=1}^m$, the data φ, ψ, and f but are otherwise independent of the points $\{z_j\}_{j=1}^m$.

Proof. Since $u_\alpha - u \in H_0^1(\Omega)$, we have, for any $v \in H_0^1(\Omega)$,

$$\int_\Omega \partial_t^2(u_\alpha - u)v + \int_\Omega \gamma_\alpha \operatorname{grad}(u_\alpha - u) \cdot \operatorname{grad} v = \sum_{j=1}^m (\gamma_0 - \gamma_j) \int_{z_j+\alpha B_j} \operatorname{grad} u \cdot \operatorname{grad} v. \tag{4}$$

Let v_α be defined by

$$\begin{cases} v_\alpha \in H_0^1(\Omega), \\ \operatorname{div} \gamma_\alpha \operatorname{grad} v_\alpha = \partial_t(u_\alpha - u) \quad \text{in } \Omega. \end{cases}$$

Then

$$\int_\Omega \gamma_\alpha \operatorname{grad}(u_\alpha - u)\cdot \operatorname{grad} v_\alpha = -\int_\Omega \partial_t(u_\alpha - u)(u_\alpha - u) = -\frac{1}{2}\partial_t \int_\Omega (u_\alpha - u)^2$$

and

$$\begin{aligned}\int_\Omega \partial_t^2(u_\alpha - u)v_\alpha &= \int_\Omega \operatorname{div}\gamma_\alpha \operatorname{grad}\partial_t v_\alpha\ v_\alpha \\ &= -\int_\Omega \gamma_\alpha \operatorname{grad}\partial_t v_\alpha \cdot \operatorname{grad} v_\alpha \\ &= -\frac{1}{2}\partial_t \int_\Omega \gamma_\alpha |\operatorname{grad} v_\alpha|^2.\end{aligned}$$

Thus it follows that

$$\partial_t \int_\Omega \gamma_\alpha |\operatorname{grad} v_\alpha|^2 + \partial_t \int_\Omega (u_\alpha - u)^2 = -2\sum_{j=1}^m (\gamma_0 - \gamma_j)\int_{z_j+\alpha B_j} \operatorname{grad} u\cdot \operatorname{grad} v_\alpha.$$

Next

$$\left|\sum_{j=1}^m (\gamma_0 - \gamma_j)\int_{z_j+\alpha B_j} \operatorname{grad} u\cdot \operatorname{grad} v_\alpha\right| \le C||\operatorname{grad} u||_{L^2(\mathcal{B}_\alpha)}||\operatorname{grad} v_\alpha||_{L^2(\Omega)}.$$

Since $u \in \mathcal{C}^\infty([0,T]\times\overline{\Omega})$, we have

$$||\operatorname{grad} u||_{L^2(\mathcal{B}_\alpha)} \le ||\operatorname{grad} u||_{L^\infty(\mathcal{B}_\alpha)}\alpha\left(\sum_{j=1}^m |B_j|\right)^{\frac{1}{2}} \le C\alpha,$$

which gives

$$\left|\sum_{j=1}^m (\gamma_0 - \gamma_j)\int_{z_j+\alpha B_j} \operatorname{grad} u\cdot \operatorname{grad} v_\alpha\right| \le C\alpha||\operatorname{grad} v_\alpha||_{L^2(\Omega)},$$

and so

$$\partial_t \int_\Omega \gamma_\alpha |\operatorname{grad} v_\alpha|^2 + \partial_t \int_\Omega (u_\alpha - u)^2 \le C\alpha\left(\int_\Omega \gamma_\alpha |\operatorname{grad} v_\alpha|^2 + \int_\Omega (u_\alpha - u)^2\right)^{1/2}.$$

From the Gronwall lemma, it follows that

$$\left(\int_\Omega \gamma_\alpha |\operatorname{grad} v_\alpha|^2\right)^{1/2} + \left(\int_\Omega (u_\alpha - u)^2\right)^{1/2} \le C\alpha.$$

Combining this last estimate with the fact that

$$||\partial_t(u_\alpha - u)||_{L^\infty(0,T;H^{-1}(\Omega))} \le C||\operatorname{grad} v_\alpha||_{L^\infty(0,T;L^2(\Omega))},$$

we obtain the desired estimate (3). We remark that, taking (at least formally) $v = \partial_t(u_\alpha - u)$ in (4), we arrive at

$$\partial_t \int_\Omega \left[|\partial_t(u_\alpha - u)|^2 + \gamma_\alpha |\operatorname{grad}(u_\alpha - u)|^2\right] = 2\sum_{j=1}^m (\gamma_0 - \gamma_j)\int_{z_j+\alpha B_j} \operatorname{grad} u\cdot \operatorname{grad}\partial_t(u_\alpha - u).$$

Using now the regularity of u in Ω and estimate (3) given above, we see that

$$\left|\sum_{j=1}^{m}(\gamma_0-\gamma_j)\int_{z_j+\alpha B_j}\operatorname{grad}u\cdot\operatorname{grad}\partial_t(u_\alpha-u)\right| \le C||\operatorname{grad}u||_{H^2(\mathcal{B}_\alpha)}||\partial_t(u_\alpha-u)||_{H^{-1}(\Omega)} \le C\alpha^2,$$

where C is independent of t and α, and so we obtain

$$\partial_t\int_\Omega\left[|\partial_t(u_\alpha-u)|^2+\gamma_\alpha|\operatorname{grad}(u_\alpha-u)|^2\right]\le C\alpha^2,$$

which yields the estimate

$$||\partial_t(u_\alpha-u)||_{L^\infty(0,T;L^2(\Omega))}+||u_\alpha-u||_{L^\infty(0,T;H_0^1(\Omega))}\le C\alpha,$$

where C is independent of α and the points $\{z_j\}_{j=1}^m$. □

3. An asymptotic formula. Before formulating the main result of this section, we need to introduce some additional notation. For any $1\le j\le m$, let Φ_j denote the vector-valued solution to

$$\text{(5)}\qquad\begin{cases}\Delta\Phi_j=0 \text{ in } B_j, \text{ and } \mathbf{R}^2\setminus\overline{B_j},\\ \Phi_j \text{ is continuous across } \partial B_j,\\ \dfrac{\gamma_0}{\gamma_j}\dfrac{\partial\Phi_j}{\partial\nu_j}\Big|_+-\dfrac{\partial\Phi_j}{\partial\nu_j}\Big|_-=-\nu_j,\\ \lim\limits_{|y|\to+\infty}|\Phi_j(y)|=0.\end{cases}$$

The existence and uniqueness of this Φ_j can be established using single layer potentials with suitably chosen densities; see [12]. In terms of this function, we are able to prove the following result about the asymptotic behavior of $\frac{\partial u_\alpha}{\partial\nu_j}|_{\partial(z_j+\alpha B_j)^+}$.

PROPOSITION 3.1. *For $y\in\partial B_j$, we have, in the weak sense,*

$$\text{(6)}\qquad\frac{\partial u_\alpha}{\partial\nu_j}|_{\partial(z_j+\alpha B_j)^+}(z_j+\alpha y,t)=\left[\nu_j+\left(\frac{\gamma_0}{\gamma_j}-1\right)\frac{\partial\Phi_j}{\partial\nu_j}|_+(y)\right]\cdot\operatorname{grad}u(z_j,t)+o(1).$$

The term $o(1)$ depends on the shapes of the domains $\{B_j\}_{j=1}^m$ and Ω, the constants d_0, T, γ_0, $\{\gamma_j\}_{j=1}^m$, the data φ,ψ, and f but is otherwise independent of the points $\{z_j\}_{j=1}^m$.

For simplicity, let us restrict our attention to the case of a single inhomogeneity, i.e., the case $m=1$. The proof, for any fixed number m of well-separated inhomogeneities, follows by iteration of the argument that we will present for the case $m=1$. In order to further simplify notation, we assume that the single inhomogeneity has the form αB; that is, we assume it is centered at the origin. We denote the conductivity inside αB by γ_* and define Φ_* to be the same as Φ_j, defined in (5), but with B_j and γ_j replaced by B and γ_*, respectively. Define ν to be the outward unit normal to ∂B. Let $U_\alpha=\operatorname{grad}u_\alpha(x,t)$ and $U_0=\operatorname{grad}u(x,t)$ in $\Omega\times(0,T)$.

We start with a formal derivation of the asymptotic formula (6). Following a common practice in multiscale expansions, we introduce the local variable $y=\frac{x}{\alpha}$. We expect that $U_\alpha(x,t)$ will differ appreciably from $U_0(x,t)$ for x near the origin,

but it will differ little from $U_0(x,t)$ for x far from the origin. Therefore, in the spirit of matched asymptotic expansions, we shall represent $U_\alpha(x,t)$ by two different expansions: an inner expansion for x near the origin and an outer expansion for x far from the origin. The outer expansion must begin with U_0, so we write

$$U_\alpha(x,t) = U_0(x,t) + \beta_1(\alpha)U_1(x,t) + \beta_2(\alpha)U_2(x,t) + \cdots \quad \text{for } |x| >> O(\alpha),\ t \in (0,T),$$

where the gauge functions $\beta_1(\alpha), \beta_2(\alpha), \ldots$ and the functions $U_1, U_2, \ldots$ are to be found. We write the inner expansion as

$$U_\alpha(z_j + \alpha y, t) = V_0(y,t) + \mu_1(\alpha)V_1(y,t) + \mu_2(\alpha)V_2(y,t) + \cdots \quad \text{for } |y| = O(1),\ t \in (0,T),$$

where the gauge functions $\mu_1(\alpha), \mu_2(\alpha), \ldots$ and the functions $V_0, V_1, V_2, \ldots$ are to be found. Here the gauge functions $\beta_i(\alpha)$ and $\mu_i(\alpha)$ satisfy $\beta_i(\alpha) >> \beta_{i+1}(\alpha)$ and $\mu_i(\alpha) >> \mu_{i+1}(\alpha)$ as α tends to 0.

The inner and outer expansions must be asymptotically equal in some overlap domain within which the stretched variable $|y|$ is large and $|x|$ is small. In this domain, the matching condition is

$$U_0(x,t) + \beta_1(\alpha)U_1(x,t) + \cdots \sim V_0(y,t) + \mu_1(\alpha)V_1(y,t) + \cdots.$$

From the terms of order α^0, we obtain the first matching condition

$$V_0(y,t) \to U_0(0,t) \text{ as } |y| \to +\infty \text{ (for } t \in (0,T)).$$

Since

$$\partial_t^2 u_\alpha - \ \text{div}\ \gamma_\alpha U_\alpha = 0 \text{ and } \ \text{curl}\ U_\alpha = 0, \tag{7}$$

by substituting the inner and outer expansions into these equations and formally equating coefficients of α^{-1}, we get

$$\text{curl}\ {}_y V_0 = 0,\ \text{div}\ {}_y\gamma(y)V_0 = 0 \text{ in } \mathbf{R}^2,$$

where

$$\gamma(y) = \begin{cases} \gamma_0 \text{ in } \mathbf{R}^2 \setminus \overline{B}, \\ \gamma_* \text{ in } B. \end{cases}$$

Therefore,

$$V_0(y) = \text{grad}\left(\left(\frac{\gamma_0}{\gamma_*} - 1\right)\Phi_*(y) + y\right) \cdot \text{grad}\, u(0,t),$$

and so, by multiplying by ν_j, we arrive at

$$\frac{\partial u_\alpha}{\partial \nu}\bigg|_{\partial(\alpha B)^+}(\alpha y, t) = \nu \cdot \text{grad}\, u(0,t) + \left(\frac{\gamma_0}{\gamma_*} - 1\right)\frac{\partial \Phi_*}{\partial \nu}\bigg|_+ (y) \cdot \text{grad}\, u(0,t) + o(1). \tag{8}$$

In the case of m (well-separated) inhomogeneities $z_j + \alpha B_j$, $j = 1, \ldots, m$, we (formally) obtain from (8) that the following asymptotic formula holds for any $y \in \partial B_j$:

$$\frac{\partial u_\alpha}{\partial \nu_j}\bigg|_{\partial(z_j+\alpha B_j)^+}(z_j + \alpha y, t) = \nu_j \cdot \text{grad}\, u(z_j,t) + \left(\frac{\gamma_0}{\gamma_j} - 1\right)\frac{\partial \Phi_j}{\partial \nu_j}\bigg|_+ (y) \cdot \text{grad}\, u(z_j,t) + o(1).$$

Proof of Proposition 3.1. Let θ be given in $\mathcal{C}_0^\infty(]0,T[)$. For any function $v \in L^1(0,T;L^2(\Omega))$, we define

$$\hat{v}(x) = \int_0^T v(x,t)\,\theta(t)\,dt \in L^2(\Omega).$$

We remark that $\widehat{\partial_t v}(x) = -\int_0^T v(x,t)\theta'(t)\,dt$. So we deduce from (7) that $\hat{U}_\alpha$ satisfies

$$\begin{cases} \text{div } \gamma_\alpha \hat{U}_\alpha = \displaystyle\int_0^T u_\alpha\,\theta''(t)\,dt & \text{in } \Omega, \\ \text{curl } \hat{U}_\alpha = 0 & \text{in } \Omega. \end{cases}$$

Analogously, $\hat{U}_0 = \int_0^T U_0(x,t)\,\theta(t)\,dt$ satisfies

$$\begin{cases} \gamma_0 \text{ div } \hat{U}_0 = \displaystyle\int_0^T u\,\theta''(t)\,dt & \text{in } \Omega, \\ \text{curl } \hat{U}_0 = 0 & \text{in } \Omega. \end{cases}$$

Indeed, we have $\hat{U}_\alpha \times n = \hat{U}_0 \times n = \text{grad}_{\partial\Omega}\,\hat{f} \times n$ on the boundary $\partial\Omega$, where $\text{grad}_{\partial\Omega}$ is the tangential gradient. Following [6], we introduce q_α^* as the unique solution to the following problem:

$$\begin{cases} \Delta q_\alpha^* = 0 & \text{in } \widetilde{\Omega} = \left(\dfrac{\Omega}{\alpha}\right) \Big\backslash\, \overline{B} \text{ and in } B, \\ q_\alpha^* \text{ is continuous across } \partial B, \\ \gamma_0 \dfrac{\partial q_\alpha^*}{\partial \nu}\Big|_+ - \gamma_* \dfrac{\partial q_\alpha^*}{\partial \nu}\Big|_- = -(\gamma_0 - \gamma_*)\hat{U}_0(\alpha y)\cdot\nu & \text{on } \partial B, \\ q_\alpha^* = 0 & \text{on } \partial\widetilde{\Omega}. \end{cases}$$

The jump condition

$$\gamma_0 \frac{\partial q_\alpha^*}{\partial \nu}\Big|_+ - \gamma_* \frac{\partial q_\alpha^*}{\partial \nu}\Big|_- = -(\gamma_0 - \gamma_*)\hat{U}_0(\alpha y)\cdot\nu \quad \text{on } \partial B$$

guarantees that $\hat{U}_\alpha(x) - \hat{U}_0(x) - \text{grad}_y\, q_\alpha^*(\frac{x}{\alpha})$ belongs to the functional space

$$Z_\alpha(\Omega) = \{v \in L^2(\Omega),\ \text{div } (\gamma_\alpha v) \in L^2(\Omega),\ \text{curl } v \in L^2(\Omega), v \times n = 0 \text{ on } \partial\Omega\}.$$

Since

$$\text{(9)} \qquad \begin{cases} \text{div } \gamma_\alpha \left(\hat{U}_\alpha - \hat{U}_0 - \text{grad}_y\, q_\alpha^*\left(\dfrac{x}{\alpha}\right)\right) \\ \qquad = \displaystyle\int_0^T \left[u_\alpha - \chi(\Omega \setminus \overline{\alpha B})u - \frac{\gamma_*}{\gamma_0}\chi(\alpha B)u\right]\theta''(t)\,dt & \text{in } \Omega, \\ \text{curl } \left(\hat{U}_\alpha - \hat{U}_0 - \text{grad}_y\, q_\alpha^*\left(\dfrac{x}{\alpha}\right)\right) = 0 & \text{in } \Omega, \\ \left(\hat{U}_\alpha - \hat{U}_0 - \text{grad}_y\, q_\alpha^*\left(\dfrac{x}{\alpha}\right)\right) \times n = 0 & \text{on } \partial\Omega, \end{cases}$$

where $\chi(\omega)$ is the characteristic function of the domain ω, we arrive, as a consequence of the energy estimate (3), at the following:

$$\begin{cases} \left(\hat{U}_\alpha - \hat{U}_0 - \operatorname{grad}_y q_\alpha^*\left(\frac{x}{\alpha}\right)\right) \in Z_\alpha(\Omega), \\ \operatorname{div} \gamma_\alpha \left(\hat{U}_\alpha - \hat{U}_0 - \operatorname{grad}_y q_\alpha^*\left(\frac{x}{\alpha}\right)\right) = 0(\alpha) \quad \text{in } \Omega, \\ \operatorname{curl} \left(\hat{U}_\alpha - \hat{U}_0 - \operatorname{grad}_y q_\alpha^*\left(\frac{x}{\alpha}\right)\right) = 0 \quad \text{in } \Omega, \\ \left(\hat{U}_\alpha - \hat{U}_0 - \operatorname{grad}_y q_\alpha^*\left(\frac{x}{\alpha}\right)\right) \times n = 0 \quad \text{on } \partial\Omega. \end{cases}$$

From [6], we know that this yields the estimate

$$\left\| \operatorname{div} \gamma_\alpha \left(\hat{U}_\alpha - \hat{U}_0 - \operatorname{grad}_y q_\alpha^*\left(\frac{x}{\alpha}\right)\right)\right\|_{L^2(\Omega)} + \left\|\hat{U}_\alpha - \hat{U}_0 - \operatorname{grad}_y q_\alpha^*\left(\frac{x}{\alpha}\right)\right\|_{L^2(\Omega)} \leq C\alpha,$$

and so

$$\left(\hat{U}_\alpha - \hat{U}_0 - \operatorname{grad}_y q_\alpha^*\left(\frac{x}{\alpha}\right)\right) \cdot \nu|_+ = 0(\alpha) \quad \text{on } \partial(\alpha B).$$

Let q_* be the unique (scalar) solution to

$$\begin{cases} \Delta q_* = 0 \quad \text{in } \mathbf{R}^2 \setminus \overline{B} \text{ and in } B, \\ q_* \text{ is continuous across } \partial B, \\ \gamma_0 \dfrac{\partial q_*}{\partial \nu}\bigg|_+ - \gamma_* \dfrac{\partial q_*}{\partial \nu}\bigg|_- = -(\gamma_0 - \gamma_*)\hat{U}_0(0) \cdot \nu \quad \text{on } \partial B, \\ \lim\limits_{|y| \to +\infty} q_* = 0. \end{cases}$$

From [12, Theorem 1], it follows that

$$\left\|(\operatorname{grad}_y q_* - \operatorname{grad}_y q_\alpha^*)\left(\frac{x}{\alpha}\right)\right\|_{L^2(\Omega)} \leq C\alpha^{1/2},$$

which yields

$$\left(\hat{U}_\alpha - \hat{U}_0 - \operatorname{grad}_y q_*\left(\frac{x}{\alpha}\right)\right) \cdot \nu = o(1) \quad \text{on } \partial(\alpha B).$$

Writing q_* in terms of Φ_* gives

$$\int_0^T \left[\frac{\partial u_\alpha}{\partial \nu}\bigg|_{\partial(\alpha B)^+}(\alpha y) - \nu \cdot \operatorname{grad} u(0,t) - \left(\frac{\gamma_0}{\gamma_*} - 1\right)\frac{\partial \Phi_*}{\partial \nu}\bigg|_+ (y) \cdot \operatorname{grad} u(0,t)\right] \theta(t)\, dt = o(1)$$

for any $\theta \in \mathcal{C}_0^\infty(]0,T[)$. In view of (9), the remainder $o(1)$ in the above asymptotic formula is bounded by $C_\alpha ||\theta||_{H^2(0,T)}$, where the constant C_α is independent of θ and goes to zero as $\alpha \to 0$. Therefore,

$$\frac{\partial u_\alpha}{\partial \nu}\bigg|_{\partial(\alpha B)^+}(\alpha y) - \nu \cdot \operatorname{grad} u(0,t) - \left(\frac{\gamma_0}{\gamma_*} - 1\right)\frac{\partial \Phi_*}{\partial \nu}\bigg|_+ (y) \cdot \operatorname{grad} u(0,t) = o(1)$$

holds in a weak sense, and so, by iterating the same argument for the case of m (well-separated) inhomogeneities $z_j + \alpha B_j, j = 1, \ldots, m$, we arrive at the promised asymptotic formula (6). □

4. The identification procedure. Let $\beta(x) \in \mathcal{C}_0^\infty(\Omega)$ be a cutoff function such that $\beta(x) \equiv 1$ in a subdomain Ω' of Ω that contains the inhomogeneities $\mathcal{B}_\alpha$. For an arbitrary $\eta \in \mathbf{R}^2$, we assume that we are in possession of the boundary measurements of

$$\frac{\partial u_\alpha}{\partial n} \quad \text{on } \Gamma \times (0,T)$$

for

$$\varphi(x) = \varphi_\eta(x) = e^{i\eta\cdot x}, \psi(x) = \psi_\eta(x) = -i\sqrt{\gamma_0}|\eta|e^{i\eta\cdot x}$$
$$\text{and } f(x,t) = f_\eta(x,t) = e^{i\eta\cdot x - i\sqrt{\gamma_0}|\eta|t}.$$

This particular choice of data φ, ψ, and f implies that the background solution u of the wave equation (2) in the absence of any inhomogeneity is given by

$$u(x,t) = u_\eta(x,t) = e^{i\eta\cdot x - i\sqrt{\gamma_0}|\eta|t} \quad \text{in } \Omega \times (0,T).$$

Suppose that T and the part Γ of the boundary $\partial\Omega$ are such that they geometrically control Ω, which roughly means that every geometrical optic ray, starting at any point $x \in \Omega$ at time $t = 0$, hits Γ before time T at a nondiffractive point; see [7]. Then, from [19, Theorem 6.4, p. 75] and [7], it follows that, for any $\eta \in \mathbf{R}^2$, we can construct by the Hilbert uniqueness method a unique $g_\eta \in H_0^1(0,T;L^2(\Gamma))$ in such a way that the unique weak solution w_η in $\mathcal{C}^0(0,T;L^2(\Omega)) \cap \mathcal{C}^1(0,T;H^{-1}(\Omega))$ of the wave equation

$$\text{(10)} \qquad \begin{cases} (\partial_t^2 - \gamma_0\Delta)w_\eta = 0 \quad \text{in } \Omega \times (0,T), \\ w_\eta|_{t=0} = \beta(x)e^{i\eta\cdot x} \in H_0^1(\Omega), \\ \partial_t w_\eta|_{t=0} = 0 \quad \text{in } \Omega, \\ w_\eta|_{\Gamma\times(0,T)} = g_\eta, \\ w_\eta|_{\partial\Omega\setminus\overline{\Gamma}\times(0,T)} = 0, \end{cases}$$

satisfies $w_\eta(T) = \partial_t w_\eta(T) = 0$. Let $v_{\alpha,\eta} \in \mathcal{C}^0(0,T;L^2(\Omega)) \cap \mathcal{C}^1(0,T;H^{-1}(\Omega))$ be defined by

$$\begin{cases} (\partial_t^2 - \gamma_0\Delta)v_{\alpha,\eta} = 0 \quad \text{in } \Omega \times (0,T), \\ v_{\alpha,\eta}|_{t=0} = 0 \quad \text{in } \Omega, \\ \partial_t v_{\alpha,\eta}|_{t=0} = \displaystyle\sum_{j=1}^m i\left(1 - \frac{\gamma_0}{\gamma_j}\right)\eta \cdot \left(\nu_j + \left(\frac{\gamma_0}{\gamma_j} - 1\right)\frac{\partial\Phi_j}{\partial\nu_j}|_+\right) e^{i\eta\cdot z_j}\delta_{\partial(z_j+\alpha B_j)} \quad \text{in } \Omega, \\ v_{\alpha,\eta}|_{\partial\Omega\times(0,T)} = 0. \end{cases}$$

Since $\frac{\partial\Phi_j}{\partial\nu_j}|_+(y)\delta_{\partial(z_j+\alpha B_j)} \in H^{-1}(\Omega)$ for $j = 1,\dots,m$, the existence and uniqueness of a solution $v_{\alpha,\eta}$ can be established by transposition; see [20] and [19, Theorem 4.2, p. 46]. Indeed, we can prove that $\frac{\partial v_{\alpha,\eta}}{\partial n}|_\Gamma \in H^{-1}(0,T;L^2(\Gamma))$. To do so, let θ be defined as

$$\begin{cases} \theta \in H_0^1(\Omega), \\ \gamma_0\,\Delta\theta = \displaystyle\sum_{j=1}^m i\left(1 - \frac{\gamma_0}{\gamma_j}\right)\eta \cdot \left(\nu_j + \left(\frac{\gamma_0}{\gamma_j} - 1\right)\frac{\partial\Phi_j}{\partial\nu_j}|_+\right) e^{i\eta\cdot z_j}\delta_{\partial(z_j+\alpha B_j)} \in H^{-1}(\Omega) \quad \text{in } \Omega, \end{cases}$$

and introduce

$$z(x,t) = \int_0^t v_{\alpha,\eta}(x,s)\, ds + \theta(x) \in L^2(\Omega).$$

It is easy to see that z satisfies the initial boundary value problem

$$\begin{cases} (\partial_t^2 - \gamma_0\Delta)z = 0 \quad \text{in } \Omega, \\ z|_{t=0} = \theta \in H_0^1(\Omega), \partial_t z|_{t=0} = 0 \quad \text{in } \Omega, \\ z|_{\partial\Omega\times(0,T)} = 0. \end{cases}$$

Classical regularity results (see [19, Theorem 4.1, p. 44]) yield

$$\frac{\partial z}{\partial n}|_\Gamma \in L^2(0,T;L^2(\Gamma)),$$

and so $\frac{\partial v_{\alpha,\eta}}{\partial n}|_\Gamma = \partial_t(\frac{\partial z}{\partial n}|_\Gamma) \in H^{-1}(0,T;L^2(\Gamma))$.

The following holds.

PROPOSITION 4.1. *Suppose that Γ and T geometrically control Ω. For any $\eta \in \mathbf{R}^2$, we have*

$$\alpha \sum_{j=1}^m i\left(1 - \frac{\gamma_0}{\gamma_j}\right) e^{2i\eta\cdot z_j}\eta \cdot \int_{\partial B_j} \left(\nu_j + \left(\frac{\gamma_0}{\gamma_j} - 1\right)\frac{\partial \Phi_j}{\partial \nu_j}|_+(y)\right) e^{i\alpha\eta\cdot y}\, ds_j(y)$$

$$= -\gamma_0 \int_0^T \int_\Gamma g_\eta \frac{\partial v_{\alpha,\eta}}{\partial n}.$$

Here $\int_0^T \int_\Gamma g_\eta \frac{\partial v_{\alpha,\eta}}{\partial n}$ is in the sense of the duality pairing between $H_0^1(0,T)$ and $H^{-1}(0,T)$. Proposition 4.1 is obtained by multiplying $(\partial_t^2 - \gamma_0\Delta)v_{\alpha,\eta} = 0$ by w_η and integrating by parts over $(0,T)\times\Omega$. In fact, we have

$$-\gamma_0 \int_0^T \int_\Gamma g_\eta \frac{\partial v_{\alpha,\eta}}{\partial n} = \alpha \sum_{j=1}^m i\left(1 - \frac{\gamma_0}{\gamma_j}\right) e^{i\eta\cdot z_j}\eta$$
$$\cdot \int_\Omega \left(\nu_j + \left(\frac{\gamma_0}{\gamma_j} - 1\right)\frac{\partial \Phi_j}{\partial \nu_j}|_+\left(\frac{x - z_j}{\alpha}\right)\right) \delta_{\partial(z_j+\alpha B_j)} e^{i\eta\cdot x}\beta(x)\, dx,$$

where the integral on the right-hand side is in the sense of the duality pairing between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$. Thus

$$-\gamma_0 \int_0^T \int_\Gamma g_\eta \frac{\partial v_{\alpha,\eta}}{\partial n} = \alpha \sum_{j=1}^m i\left(1 - \frac{\gamma_0}{\gamma_j}\right) e^{i\eta\cdot z_j}\eta$$
$$\cdot \int_{\partial(z_j+\alpha B_j)} \left(\nu_j + \left(\frac{\gamma_0}{\gamma_j} - 1\right)\frac{\partial \Phi_j}{\partial \nu_j}|_+\left(\frac{x - z_j}{\alpha}\right)\right) e^{i\eta\cdot x}\, ds_j(x)$$

since $\beta(x) \equiv 1$ in a subdomain Ω' of Ω that contains the inhomogeneities $\mathcal{B}_\alpha$. By a change of variables, the above identity leads to the desired formula.

Taking now Taylor expansion of $e^{i\alpha\eta\cdot y}$ and having in mind that [12]

$$\int_{\partial B_j} \left(\nu_j + \left(\frac{\gamma_0}{\gamma_j} - 1\right)\frac{\partial \Phi_j}{\partial \nu_j}|_+(y)\right) ds_j(y) = 0,$$

we obtain the more convenient asymptotic formula.

PROPOSITION 4.2. *Suppose that Γ and T geometrically control Ω. For any $\eta \in \mathbf{R}^2$, we have*

$$\alpha^2 \sum_{j=1}^{m} \left(1 - \frac{\gamma_0}{\gamma_j}\right) e^{2i\eta\cdot z_j} \eta \cdot \int_{\partial B_j} \left(\nu_j + \left(\frac{\gamma_0}{\gamma_j} - 1\right) \frac{\partial \Phi_j}{\partial \nu_j}|_+(y)\right) \eta \cdot y \, ds_j(y) = \gamma_0 \int_0^T \int_\Gamma g_\eta \frac{\partial v_{\alpha,\eta}}{\partial n} + o(\alpha^2).$$

Next, for any $\eta \in \mathbf{R}^2$, let θ_η denote the solution to the Volterra equation of the second kind:

$$\text{(11)} \quad \begin{cases} \partial_t \theta_\eta(x,t) + \displaystyle\int_t^T e^{-i\sqrt{\gamma_0}|\eta|(s-t)} (\theta_\eta(x,s) - i\sqrt{\gamma_0}|\eta| \partial_t \theta_\eta(x,s)) \, ds = g_\eta(x,t) \\ \quad \text{for } x \in \Gamma, t \in (0,T), \\ \theta_\eta(x,0) = 0 \quad \text{for } x \in \Gamma. \end{cases}$$

The existence and uniqueness of this θ_η in $H^1(0,T;L^2(\Gamma))$ for any $\eta \in \mathbf{R}^2$ can be established using the resolvent kernel. Since $g_\eta \in H_0^1(0,T;L^2(\Gamma))$, the solution θ_η belongs, in fact, to $H^2(0,T;L^2(\Gamma))$. Note that it was Yamamoto [32] who first conceived the idea of using such a Volterra equation to apply the geometrical control for solving inverse source problems. We also note from differentiation of (11) with respect to t that θ_η is the unique solution of the ODE

$$\text{(12)} \quad \begin{cases} \partial_t^2 \theta_\eta - \theta_\eta = e^{i\sqrt{\gamma_0}|\eta|t} \partial_t (e^{-i\sqrt{\gamma_0}|\eta|t} g_\eta) \quad \text{for } x \in \Gamma, \ t \in (0,T), \\ \theta_\eta(x,0) = 0, \partial_t \theta_\eta(x,T) = 0 \quad \text{for } x \in \Gamma. \end{cases}$$

Therefore, the function θ_η may be found in practice explicitly with variation of parameters. It also immediately follows from this observation that θ_η belongs to $H^2(0,T;L^2(\Gamma))$ since $g_\eta \in H_0^1(0,T;L^2(\Gamma))$.

To identify the locations and certain properties of the small inhomogeneities $\mathcal{B}_\alpha$, let us view the averaging of the boundary measurements $\frac{\partial u_\alpha}{\partial n}|_{\Gamma\times(0,T)}$, using the solution θ_η to the Volterra equation (11) or, equivalently, the ODE (12) as a function of η. The following holds.

THEOREM 4.3. *Let $\eta \in \mathbf{R}^2$. Let u_α be the unique solution in $\mathcal{C}^0(0,T;H^1(\Omega)) \cap \mathcal{C}^1(0,T;L^2(\Omega))$ to the wave equation* (1) *with*

$$\varphi(x) = e^{i\eta\cdot x}, \psi(x) = -i\sqrt{\gamma_0}|\eta| e^{i\eta\cdot x}, \text{ and } f(x,t) = e^{i\eta\cdot x - i\sqrt{\gamma_0}|\eta|t}.$$

Suppose that Γ and T geometrically control Ω; then we have

$$\text{(13)} \quad \begin{aligned} &\int_0^T \int_\Gamma \left[\theta_\eta \left(\frac{\partial u_\alpha}{\partial n} - \frac{\partial u}{\partial n}\right) + \partial_t \theta_\eta \partial_t \left(\frac{\partial u_\alpha}{\partial n} - \frac{\partial u}{\partial n}\right)\right] \\ &= -\int_0^T \int_\Gamma e^{i\sqrt{\gamma_0}|\eta|t} \partial_t (e^{-i\sqrt{\gamma_0}|\eta|t} g_\eta) \left(\frac{\partial u_\alpha}{\partial n} - \frac{\partial u}{\partial n}\right) \\ &= \alpha^2 \sum_{j=1}^{m} \left(\frac{\gamma_0}{\gamma_j} - 1\right) e^{2i\eta\cdot z_j} \left[M_j(\eta)\cdot\eta - |\eta|^2 |B_j|\right] \\ &\quad + o(\alpha^2), \end{aligned}$$

where θ_η is the unique solution to the ODE (12)*, with g_η defined as the boundary control in* (10)*, and M_j is the polarization tensor of B_j, defined by*

$$(M_j)_{k,l} = e_k \cdot \left(\int_{\partial B_j} \left(\nu_j + \left(\frac{\gamma_0}{\gamma_j} - 1 \right) \frac{\partial \Phi_j}{\partial \nu_j}|_+(y) \right) y \cdot e_l \, ds_j(y) \right). \tag{14}$$

Here (e_1, e_2) is an orthonormal basis of $\mathbf{R}^2$.

Proof. The first identity in (13) follows from integration by parts and use of the fact that θ_η is the solution to the ODE (12).

From $\partial_t \theta_\eta(T) = 0$ and $(\frac{\partial u_\alpha}{\partial n} - \frac{\partial u}{\partial n})|_{t=0} = 0$, the term $\int_0^T \int_\Gamma \partial_t \theta_\eta \partial_t (\frac{\partial u_\alpha}{\partial n} - \frac{\partial u}{\partial n})$ has to be interpreted as follows:

$$\int_0^T \int_\Gamma \partial_t \theta_\eta \partial_t \left(\frac{\partial u_\alpha}{\partial n} - \frac{\partial u}{\partial n} \right) = - \int_0^T \int_\Gamma \partial_t^2 \theta_\eta \left(\frac{\partial u_\alpha}{\partial n} - \frac{\partial u}{\partial n} \right). \tag{15}$$

Next, introducing

$$\widetilde{u}_{\alpha,\eta}(x,t) = u(x,t) - \gamma_0 \int_0^t e^{-i\sqrt{\gamma_0}|\eta|s} v_{\alpha,\eta}(x, t-s) \, ds, x \in \Omega, t \in (0,T),$$

we rewrite

$$\begin{aligned}
&\int_0^T \int_\Gamma \left[\theta_\eta \left(\frac{\partial u_\alpha}{\partial n} - \frac{\partial u}{\partial n} \right) + \partial_t \theta_\eta \partial_t \left(\frac{\partial u_\alpha}{\partial n} - \frac{\partial u}{\partial n} \right) \right] \\
&= \int_0^T \int_\Gamma \left[\theta_\eta \left(\frac{\partial u_\alpha}{\partial n} - \frac{\partial \widetilde{u}_{\alpha,\eta}}{\partial n} \right) + \partial_t \theta_\eta \partial_t \left(\frac{\partial u_\alpha}{\partial n} - \frac{\partial \widetilde{u}_{\alpha,\eta}}{\partial n} \right) \right] \\
&- \gamma_0 \int_0^T \int_\Gamma \left[\theta_\eta \int_0^t e^{-i\sqrt{\gamma_0}|\eta|s} \frac{\partial v_{\alpha,\eta}}{\partial n}(x, t-s) \, ds + \partial_t \theta_\eta \partial_t \int_0^t e^{-i\sqrt{\gamma_0}|\eta|s} \frac{\partial v_{\alpha,\eta}}{\partial n}(x, t-s) \, ds \right].
\end{aligned}$$

Since θ_η satisfies the Volterra equation (11) and

$$\begin{aligned}
&\partial_t \left(\int_0^t e^{-i\sqrt{\gamma_0}|\eta|s} \frac{\partial v_{\alpha,\eta}}{\partial n}(x, t-s) \, ds \right) = \partial_t \left(e^{-i\sqrt{\gamma_0}|\eta|t} \int_0^t e^{i\sqrt{\gamma_0}|\eta|s} \frac{\partial v_{\alpha,\eta}}{\partial n}(x,s) \, ds \right) \\
&= -i\sqrt{\gamma_0}|\eta| e^{-i\sqrt{\gamma_0}|\eta|t} \int_0^t e^{i\sqrt{\gamma_0}|\eta|s} \frac{\partial v_{\alpha,\eta}}{\partial n}(x,s) \, ds + \frac{\partial v_{\alpha,\eta}}{\partial n}(x,t),
\end{aligned}$$

we obtain by integrating by parts over $(0,T)$ that

$$\begin{aligned}
&\int_0^T \int_\Gamma \left[\theta_\eta \int_0^t e^{-i\sqrt{\gamma_0}|\eta|s} \frac{\partial v_{\alpha,\eta}}{\partial n}(x, t-s) \, ds + \partial_t \theta_\eta \partial_t \int_0^t e^{-i\sqrt{\gamma_0}|\eta|s} \frac{\partial v_{\alpha,\eta}}{\partial n}(x, t-s) \, ds \right] \\
&= \int_0^T \int_\Gamma \frac{\partial v_{\alpha,\eta}}{\partial n}(x,t) \left(\partial_t \theta_\eta + \int_t^T \theta_\eta(s) e^{i\sqrt{\gamma_0}|\eta|(t-s)} \, ds \right) \\
&\quad - i\sqrt{\gamma_0}|\eta| (e^{-i\sqrt{\gamma_0}|\eta|t} \partial_t \theta_\eta(t)) \int_0^t e^{i\sqrt{\gamma_0}|\eta|s} \frac{\partial v_{\alpha,\eta}}{\partial n}(x,s) \, ds \, dt \\
&= \int_0^T \int_\Gamma \frac{\partial v_{\alpha,\eta}}{\partial n}(x,t) \left(\partial_t \theta_\eta + \int_t^T (\theta_\eta(s) - i\sqrt{\gamma_0}|\eta| \partial_t \theta_\eta(s)) e^{i\sqrt{\gamma_0}|\eta|(t-s)} \, ds \right) dt \\
&= \int_0^T \int_\Gamma g_\eta(x,t) \frac{\partial v_{\alpha,\eta}}{\partial n}(x,t) \, dt,
\end{aligned}$$

and so, from Proposition 4.2, we obtain that

$$\begin{aligned}
&\int_0^T \int_\Gamma \left[\theta_\eta \left(\frac{\partial u_\alpha}{\partial n} - \frac{\partial u}{\partial n}\right) + \partial_t \theta_\eta \partial_t \left(\frac{\partial u_\alpha}{\partial n} - \frac{\partial u}{\partial n}\right)\right] \\
(16) \quad &= -\alpha^2 \sum_{j=1}^m \left(1 - \frac{\gamma_0}{\gamma_j}\right) e^{2i\eta\cdot z_j} \eta \cdot \int_{\partial B_j} \left(\nu_j + \left(\frac{\gamma_0}{\gamma_j} - 1\right) \frac{\partial \Phi_j}{\partial \nu_j}|_+(y)\right) \eta \cdot y \, ds_j(y) \\
&+ \int_0^T \int_\Gamma \left[\theta_\eta \left(\frac{\partial u_\alpha}{\partial n} - \frac{\partial \widetilde{u}_{\alpha,\eta}}{\partial n}\right) + \partial_t \theta_\eta \partial_t \left(\frac{\partial u_\alpha}{\partial n} - \frac{\partial \widetilde{u}_{\alpha,\eta}}{\partial n}\right)\right] + o(\alpha^2).
\end{aligned}$$

In order to prove Theorem 4.1, it suffices then to find the leading order term in the asymptotic expansion of

$$\int_0^T \int_\Gamma \left[\theta_\eta \left(\frac{\partial u_\alpha}{\partial n} - \frac{\partial \widetilde{u}_{\alpha,\eta}}{\partial n}\right) + \partial_t \theta_\eta \partial_t \left(\frac{\partial u_\alpha}{\partial n} - \frac{\partial \widetilde{u}_{\alpha,\eta}}{\partial n}\right)\right].$$

Let $h_{\alpha,\eta} \in \mathcal{C}^0(0,T;H_0^1(\Omega)) \cap \mathcal{C}^1(0,T;L^2(\Omega))$ be the solution to

$$\begin{cases}
(\partial_t^2 - \gamma_0 \Delta) h_{\alpha,\eta} = 0 \quad \text{in } \Omega \times (0,T), \\
h_{\alpha,\eta}|_{t=0} = 0 \quad \text{in } \Omega, \\
\partial_t h_{\alpha,\eta}|_{t=0} = -\gamma_0 |\eta|^2 \displaystyle\sum_{j=1}^m \left(1 - \frac{\gamma_0}{\gamma_j}\right) e^{i\eta\cdot x} \chi(z_j + \alpha B_j) \quad \text{in } \Omega, \\
h_{\alpha,\eta}|_{\partial\Omega\times(0,T)} = 0,
\end{cases}$$

where $\chi(z_j + \alpha B_j)$ denotes the characteristic function of the inhomogeneity $z_j + \alpha B_j$. Since

$$\begin{cases}
(\partial_t^2 - \gamma_0 \Delta) \left(\displaystyle\int_0^t e^{-i\sqrt{\gamma_0}|\eta| s} v_{\alpha,\eta}(x, t-s)\, ds\right) \\
= \displaystyle\sum_{j=1}^m i \left(1 - \frac{\gamma_0}{\gamma_j}\right) \eta \cdot \left(\nu_j + \left(\frac{\gamma_0}{\gamma_j} - 1\right) \frac{\partial \Phi_j}{\partial \nu_j}|_+(y)\right) e^{i\eta\cdot z_j} \delta_{\partial(z_j+\alpha B_j)} e^{-i\sqrt{\gamma_0}|\eta| t} \\
\quad \text{in } \Omega \times (0,T), \\
\left(\displaystyle\int_0^t e^{-i\sqrt{\gamma_0}|\eta| s} v_{\alpha,\eta}(x, t-s)\, ds\right)|_{t=0} = 0, \\
\partial_t \left(\displaystyle\int_0^t e^{-i\sqrt{\gamma_0}|\eta| s} v_{\alpha,\eta}(x, t-s)\, ds\right)|_{t=0} = 0 \quad \text{in } \Omega, \\
\left(\displaystyle\int_0^t e^{-i\sqrt{\gamma_0}|\eta| s} v_{\alpha,\eta}(x, t-s)\, ds\right)|_{\partial\Omega\times(0,T)} = 0
\end{cases}$$

and

$$\begin{cases} (\partial_t^2 - \gamma_0\Delta)\left(\displaystyle\int_0^t e^{-i\sqrt{\gamma_0}|\eta|s} h_{\alpha,\eta}(x,t-s)\,ds\right) \\ = -\gamma_0|\eta|^2 \displaystyle\sum_{j=1}^m \left(1-\frac{\gamma_0}{\gamma_j}\right) e^{i\eta\cdot x}\chi(z_j+\alpha B_j) e^{-i\sqrt{\gamma_0}|\eta|t} \quad \text{in } \Omega\times(0,T), \\ \left(\displaystyle\int_0^t e^{-i\sqrt{\gamma_0}|\eta|s} h_{\alpha,\eta}(x,t-s)\,ds\right)|_{t=0} = 0, \\ \partial_t\left(\displaystyle\int_0^t e^{-i\sqrt{\gamma_0}|\eta|s} h_{\alpha,\eta}(x,t-s)\,ds\right)|_{t=0} = 0 \quad \text{in } \Omega, \\ \left(\displaystyle\int_0^t e^{-i\sqrt{\gamma_0}|\eta|s} h_{\alpha,\eta}(x,t-s)\,ds\right)|_{\partial\Omega\times(0,T)} = 0, \end{cases}$$

setting $\widetilde{h}_{\alpha,\eta} = \int_0^t e^{-i\sqrt{\gamma_0}|\eta|s} h_{\alpha,\eta}(x,t-s)\,ds$, it follows that

$$\begin{aligned} &(\partial_t^2 - \gamma_0\Delta)(u_\alpha - \widetilde{u}_{\alpha,\eta} - \widetilde{h}_{\alpha,\eta}) \\ &= \gamma_0 \sum_{j=1}^m i\left(1-\frac{\gamma_0}{\gamma_j}\right)\left[-\frac{\partial u_\alpha}{\partial n}|_+ + \eta\cdot\left(\nu_j + \left(\frac{\gamma_0}{\gamma_j}-1\right)\frac{\partial \Phi_j}{\partial \nu_j}|_+(y)\right) e^{i\eta\cdot z_j} e^{-i\sqrt{\gamma_0}|\eta|t}\right]\delta_{\partial(z_j+\alpha B_j)} \\ &+ \sum_{j=1}^m \left(1-\frac{\gamma_0}{\gamma_j}\right)(\partial_t^2 u_\alpha + |\eta|^2\gamma_0 e^{i\eta\cdot x - i\sqrt{\gamma_0}|\eta|t})\chi(z_j+\alpha B_j), \end{aligned}$$

and, therefore, by Propositions 3.1 and 2.1, we readily get that

$$\begin{cases} (\partial_t^2 - \gamma_0\Delta)(u_\alpha - \widetilde{u}_{\alpha,\eta} - \widetilde{h}_{\alpha,\eta}) = o(\alpha^2) \quad \text{in } \Omega\times(0,T), \\ (u_\alpha - \widetilde{u}_{\alpha,\eta} - \widetilde{h}_{\alpha,\eta})|_{t=0} = 0, \partial_t(u_\alpha - \widetilde{u}_{\alpha,\eta} - \widetilde{h}_{\alpha,\eta})|_{t=0} = 0 \quad \text{in } \Omega, \\ (u_\alpha - \widetilde{u}_{\alpha,\eta} - \widetilde{h}_{\alpha,\eta})|_{\partial\Omega\times(0,T)} = 0, \end{cases} \tag{17}$$

where the right-hand side in the first equation in (17) is of order $o(\alpha^2)$ in the $H^{-2}(0,T;H^{-1}(\Omega))$ norm.

Following the proof of Proposition 2.1, we immediately obtain that

$$||u_\alpha - \widetilde{u}_{\alpha,\eta} - \widetilde{h}_{\alpha,\eta}||_{L^2(\Omega)} = o(\alpha^2), \quad t\in(0,T),\ x\in\Omega,$$

where the remainder $o(\alpha^2)$ is independent of the points $\{z_j\}_{j=1}^m$.

We now show that the estimate

$$\left\|\frac{\partial}{\partial n}(u_\alpha - \widetilde{u}_{\alpha,\eta} - \widetilde{h}_{\alpha,\eta})\right\|_{L^2(0,T;L^2(\Gamma))} = o(\alpha^2) \tag{18}$$

holds, which will immediately imply that

$$\begin{aligned} &\int_0^T\int_\Gamma \left[\theta_\eta\left(\frac{\partial u_\alpha}{\partial n} - \frac{\partial \widetilde{u}_{\alpha,\eta}}{\partial n}\right) + \partial_t\theta_\eta\partial_t\left(\frac{\partial u_\alpha}{\partial n} - \frac{\partial \widetilde{u}_{\alpha,\eta}}{\partial n}\right)\right] \\ &= \int_0^T\int_\Gamma \theta_\eta\frac{\partial \widetilde{h}_{\alpha,\eta}}{\partial n} + \partial_t\theta_\eta\partial_t\frac{\partial \widetilde{h}_{\alpha,\eta}}{\partial n} + o(\alpha^2). \end{aligned} \tag{19}$$

Let θ be given in $\mathcal{C}_0^\infty(]0,T[)$, and define

$$\hat{\widetilde{u}}_{\alpha,\eta}(x) = \int_0^T \widetilde{u}_{\alpha,\eta}(x,t)\theta(t)\,dt,$$

$$\hat{\widetilde{h}}_{\alpha,\eta}(x) = \int_0^T \widetilde{h}_{\alpha,\eta}(x,t)\theta(t)\,dt,$$

and

$$\hat{u}_\alpha(x) = \int_0^T u_\alpha(x,t)\theta(t)\,dt.$$

We have

$$\begin{cases} \operatorname{div}\gamma_\alpha \operatorname{grad}(\hat{u}_\alpha - \hat{\widetilde{u}}_{\alpha,\eta} - \hat{\widetilde{h}}_{\alpha,\eta}) = o(\alpha^2) \in L^2(\Omega\setminus\overline{\Omega'}) \cap H^{-1}(\Omega), \\ (\hat{u}_\alpha - \hat{\widetilde{u}}_{\alpha,\eta} - \hat{\widetilde{h}}_{\alpha,\eta}) = 0 \quad \text{on } \partial\Omega, \end{cases}$$

which implies from [34] that

$$\left\| \frac{\partial}{\partial n}(\hat{u}_\alpha - \hat{\widetilde{u}}_{\alpha,\eta} - \hat{\widetilde{h}}_{\alpha,\eta}) \right\|_{L^2(\Gamma)} = o(\alpha^2)$$

for all $\theta \in \mathcal{C}_0^\infty(]0,T[)$, whence

$$\left\| \frac{\partial}{\partial n}(u_\alpha - \widetilde{u}_{\alpha,\eta} - \widetilde{h}_{\alpha,\eta}) \right\|_{L^2(\Gamma)} = o(\alpha^2) \text{ a.e. in } t \in (0,T),$$

and so the desired estimate (18) holds.

On the other hand, analogously to Proposition 4.1, by integration by parts and taking the Taylor expansion of $e^{i\eta\cdot x}$ in $z_j + \alpha B_j$, the following holds:

$$\int_0^T \int_\Gamma g_\eta(x,t) \frac{\partial h_{\alpha,\eta}}{\partial n}(x,t)\,dt = \alpha^2 \sum_{j=1}^m \left(1 - \frac{\gamma_0}{\gamma_j}\right) e^{2i\eta\cdot z_j} |\eta|^2 |B_j| + o(\alpha^2). \tag{20}$$

However,

$$\int_0^T \int_\Gamma g_\eta(x,t) \frac{\partial h_{\alpha,\eta}}{\partial n}(x,t)\,dt = \int_0^T \int_\Gamma \theta_\eta \frac{\partial \widetilde{h}_{\alpha,\eta}}{\partial n} + \partial_t\theta_\eta \partial_t \frac{\partial \widetilde{h}_{\alpha,\eta}}{\partial n}, \tag{21}$$

and so, combining (16), (19), (20), and (21), we arrive at our promised asymptotic formula (13). The proof of Theorem 4.1 is then over. □

We are now in position to describe our identification procedure, which is based on Theorem 4.1. Let us neglect the asymptotically small remainder in the asymptotic formula (13) and define $\Lambda_\alpha(\eta)$ by

$$\Lambda_\alpha(\eta) = \int_0^T \int_\Gamma \left[\theta_\eta \left(\frac{\partial u_\alpha}{\partial n} - \frac{\partial u}{\partial n} \right) + \partial_t\theta_\eta \partial_t \left(\frac{\partial u_\alpha}{\partial n} - \frac{\partial u}{\partial n} \right) \right].$$

The function $\Lambda_\alpha(\eta)$ is computed in the following way. First, we construct the control g_η in (10) for given $\eta \in \mathbf{R}^2$. Then we solve the ODE (12) to find the auxiliary test

function θ_η. From the boundary measurements $\frac{\partial u_\alpha}{\partial n}|_{\Gamma\times(0,T)}$, we form the integrals that come in the expression of $\Lambda_\alpha(\eta)$.

Recall that the function $e^{2i\eta\cdot z_j}$ is exactly the Fourier transform (up to a multiplicative constant) of the Dirac function δ_{-2z_j} (a point mass located at $-2z_j$). From Theorem 4.3, it follows that the function $\Lambda_\alpha(\eta)$ is (approximately) the Fourier transform of a linear combination of derivatives of point masses, or

$$\breve{\Lambda}_\alpha(\eta) \approx \alpha^2 \sum_{j=1}^{m} L_j \delta_{-2z_j},$$

where L_j is a second order constant coefficient, differential operator whose coefficients depend on the polarization tensor M_j defined by (14) (see [12] for its properties) and $\breve{\Lambda}_\alpha(\eta)$ represents the inverse Fourier transform of $\Lambda_\alpha(\eta)$. The reader is referred to [12] for properties of the tensor polarization M_j.

The method of reconstruction we propose here consists, as in [4], in sampling values of $\breve{\Lambda}_\alpha(\eta)$ at some discrete set of points and then calculating the corresponding discrete inverse Fourier transform. After a rescaling by $-\frac{1}{2}$, the support of this discrete inverse Fourier transform yields the location of the small inhomogeneities $\mathcal{B}_\alpha$. This procedure generalizes the approach that we developed in [4] for the two-dimensional (time-independent) inverse conductivity problem. On other terms, once $\Lambda_\alpha(\eta)$ is computed from dynamic boundary measurements on Γ, we calculate its inverse Fourier transform. The asymptotic formula (13) in Theorem 4.1 asserts that this inverse Fourier transform is a distribution supported at the locations $(z_j)_{j=1}^m$.

Once the locations are known, we may calculate the polarization tensors $(M_j)_{j=1}^m$ by solving an appropriate linear system arising from (13). These polarization tensors give ideas on the orientation and relative size of the inhomogeneities [18]. We wish to point out that, from the leading order term of $\Lambda_\alpha(\eta)$ given by (13), we cannot reconstruct more details of the shapes of the domains B_j. Higher order terms in the asymptotic expansion of $\Lambda_\alpha(\eta)$, with respect to α, are needed to reconstruct the domains B_j with high resolution.

The number of data (sampling) points needed for an accurate discrete Fourier inversion of $\Lambda_\alpha(\eta)$ follows from Shannon's sampling theorem [13]. We need (conservatively) order $(\frac{h}{\delta})^2$ sampled values of $\Lambda_\alpha(\eta)$ to reconstruct, with resolution δ, a collection of inhomogeneities that lie inside a square of side h. In order to simulate errors in the measurements of $\frac{\partial u_\alpha}{\partial n}$ on $\Gamma\times(0,T)$, as well as the errors inherent in the approximation (13) and in the calculations of g_η, θ_η, and $\Lambda_\alpha(\eta)$ (by some quadrature rules), we should add random noise to the values of $\Lambda_\alpha(\eta)$. Numerical experiments in [4] for the two-dimensional (time-independent) inverse conductivity problem seem to suggest that the method is quite stable with respect to noise in measurements and errors in the different approximations.

We are convinced that the use of approximate formulae such as (13) represents a very promising approach to the dynamical identification of small inhomogeneities that are embedded in a homogeneous medium. In particular, our method can be extended to solve the dynamical identification problem of small incompressible or rigid inclusions. Formally, we can recover these two cases by letting γ_j tend to $+\infty$ or 0 in (5) and the asymptotic formula (13). Rigorously, to assert that (13) is still valid for incompressible or rigid inclusions, we should prove that the term $o(\alpha^2)$ is uniform in γ_j as $\gamma_j \to +\infty$ or 0. We also believe that our method yields a good approximation to small amplitude perturbations in the conductivity ($\gamma_\alpha(x) = \gamma_0 + \alpha\gamma_1(x)$) from the measurements of $\frac{\partial u_\alpha}{\partial n}$ on $\Gamma\times(0,T)$. Our method may yield the Fourier transform of

the perturbation $\gamma_1(x)$. This inverse problem is considered in [2].

Finally, we wish to emphasize the fact that, in the algorithm described in this paper, the locations z_j, $j = 1, \ldots, m$, of the inhomogeneities are found with an error $O(\alpha)$, and only the polarization tensors of the domains B_j can be reconstructed. Making use of higher order terms in the asymptotic expansion of $\frac{\partial u_\alpha}{\partial \nu_j}|_{\partial(z_j+\alpha B_j)^+}$, we certainly would be able to reconstruct the small inhomogeneities with higher resolution from dynamical boundary measurements on part of the boundary and capture more details of the geometries of the domains B_j. Perhaps, more importantly, this would also allow us to identify quite general conductivity inhomogeneities without restrictions on their sizes. Results in this direction are now available for the conductivity problem. In [3], based on layer potential techniques, high order terms in the asymptotic expansions of the steady-state voltage potentials in the presence of a finite number of diametrically small inhomogeneities with conductivities different from the background conductivity are rigorously derived. In [5], similar accurate asymptotic formulae are applied for the purpose of identifying the location and certain properties of the shape of the conductivity inhomogeneities. A real-time algorithm with a very high resolution and accuracy that makes use of constant current sources is designed. We believe that the results and techniques of [3] and [5] could be combined with the approach developed in this paper for recovering the small electromagnetic inhomogeneities from dynamic boundary measurements with higher resolution and accuracy. This very important issue will be considered in a forthcoming work.

Acknowledgments. The author expresses his thanks to M. Vogelius for various interesting discussions. He is also very grateful to the referees for their comments, which enabled him to make many improvements to the presentation.

REFERENCES

[1] C. Alves and H. Ammari, *Boundary integral formulae for the reconstruction of imperfections of small diameter in an elastic medium,* SIAM J. Appl. Math., 62 (2001), pp. 94–106.

[2] H. Ammari, *Identification of small amplitude perturbations in the electromagnetic parameters from partial dynamic boundary measurements,* J. Math. Anal. Appl., submitted; also available online from http://www.cmap.polytechnique/~ammari/~preprints.

[3] H. Ammari and H. Kang, *High-order terms in the asymptotic expansions of the steady-state voltage potentials in the presence of conductivity inhomogeneities of small diameter,* SIAM J. Math. Anal., submitted; also available online from http://www.cmap.polytechnique/~ammari/~preprints.

[4] H. Ammari, S. Moskow, and M. Vogelius, *Boundary integral formulas for the reconstruction of electromagnetic imperfections of small diameter,* ESAIM Control Optim. Calc. Var., to appear; also available online from http://www.cmap.polytechnique/~ammari/~preprints.

[5] H. Ammari and J. K. Seo, *A new algorithm for the reconstruction of conductivity inhomogeneities,* J. Amer. Math. Soc., submitted; also available online from http://www.cmap.polytechnique/~ammari/~preprints.

[6] H. Ammari, M. Vogelius, and D. Volkov, *Asymptotic formulas for perturbations in the electromagnetic fields due to the presence of inhomogeneities of small diameter* II. *The full Maxwell equations*, J. Math. Pures Appl. (9), 80 (2001), pp. 769–814.

[7] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary,* SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[8] M. I. Belishev and Ya Kurylev, *Boundary control, wave field continuation and inverse problems for the wave equation,* Comput. Math. Appl., 22 (1991), pp. 27–52.

[9] E. Beretta, A. Mukherjee, and M. Vogelius, *Asymptotic formuli for steady state voltage potentials in the presence of conductivity imperfection of small area,* Z. Angew. Math. Phys., 52 (2001), pp. 543–572.

[10] G. Bruckner and M. Yamamoto, *Determination of point wave sources by pointwise observations: Stability and reconstruction,* Inverse Problems, 16 (2000), pp. 723–748.
[11] A. P. Calderón, *On an inverse boundary value problem,* in Proceedings of a Seminar on Numerical Analysis and its Applications to Continuum Physics, Sociedade Brasileira de Matemática, Rio de Janeiro, Brazil, 1980, pp. 65–73.
[12] D. J. Cedio-Fengya, S. Moskow, and M. Vogelius, *Identification of conductivity imperfections of small diameter by boundary measurements. Continuous dependence and computional reconstruction,* Inverse Problems, 14 (1998), pp. 553–595.
[13] I. Daubechies, *Ten Lectures on Wavelets,* CBMS-NSF Regional Conf. Ser. in Appl. Math. 61, SIAM, Philadelphia, 1992.
[14] L. C. Evans, *Partial Differential Equations,* Grad. Stud. Math. 19, AMS, Providence, RI, 1998.
[15] A. Friedman and M. Vogelius, *Identification of small inhomogeneities of extreme conductivity by boundary measurements: A theorem on continuous dependence*, Arch. Ration. Mech. Anal., 105 (1989), pp. 299–326.
[16] M. Grasselli and M. Yamamoto, *Identifying a spatial body force in linear elastodynamics via traction measurements,* SIAM J. Control Optim., 36 (1998), pp. 1190–1206.
[17] V. Isakov, *Inverse Source Problems,* AMS, Providence, RI, 1990.
[18] R. E. Kleinman and T. B. A. Senior, *Rayleigh scattering*, in Low and High Frequency Asymptotics, V. K. Varadan and V. V. Varadan, eds., North–Holland, Amsterdam, 1986, pp. 1–70.
[19] J.-L. Lions, *Contrôlabilité exacte, Perturbations et Stabilisation de Systèmes Distribués, Tome* 1*, Contrôlabilité Exacte,* Masson, Paris, 1988.
[20] J.-L. Lions and E. Magenes, *Nonhomogeneous Boundary Value Problems and Applications, Vol.* 1, Springer-Verlag, New York, 1972.
[21] A. I. Nachman, *Global uniqueness for a two-dimensional inverse boundary value problem,* Ann. of Math. (2), 143 (1996), pp. 71–96.
[22] S. Nicaise, *Exact boundary controllability of Maxwell's equations in heteregeneous media and an application to an inverse source problem,* SIAM J. Control Optim., 38 (2000), pp. 1145–1170.
[23] J.-P. Puel and M. Yamamoto, *Applications de la contrôlabilité exacte à quelques problèmes inverses hyperboliques,* C. R. Acad. Sci. Paris Sér. I Math., 320 (1995), pp. 1171–1176.
[24] J.-P. Puel and M. Yamamoto, *On a global estimate in a linear inverse hyperbolic problem,* Inverse Problems, 12 (1996), pp. 995–1002.
[25] J.-P. Puel and M. Yamamoto, *Smoothing property in multidimensional inverse hyperbolic problems: Applications to uniqueness and stability,* J. Inverse Ill-Posed Prob., 4 (1996), pp. 283–296.
[26] J.-P. Puel and M. Yamamoto, *Generic well-posedness in a multidimensional inverse hyperbolic problem,* J. Inverse Ill-Posed Prob., 5 (1997), pp. 55–83.
[27] Rakesh and W. Symes, *Uniqueness for an inverse problem for the wave equation,* Comm. Partial Differential Equations, 13 (1988), pp. 87–96.
[28] V. G. Romanov and S. I. Kabanikhin, *Inverse Problems for Maxwell's Equations,* Inverse and Ill-Posed Problems Series, VSP, Utrecht, 1994.
[29] J. Sylvester and G. Uhlmann, *A global uniqueness theorem for an inverse boundary value problem,* Ann. of Math. (2), 125 (1987), pp. 153–169.
[30] Z. Sun, *On the continuous dependence for an inverse initial boundary value problem for the wave equation,* J. Math. Anal. Appl., 150 (1990), pp. 188–204.
[31] M. Yamamoto, *Well-posedness of some inverse hyperbolic problems by the Hilbert uniqueness method,* J. Inverse Ill-Posed Prob., 2 (1994), pp. 349–368.
[32] M. Yamamoto, *Stability, reconstruction formula and regularization for an inverse source hyperbolic problem by a control method,* Inverse Problems, 11 (1995), pp. 481–496.
[33] M. Yamamoto, *Determination of forces in vibrations of beams and plates by pointwise and line observations,* J. Inverse Ill-Posed Prob., 4 (1996), pp. 437–457.
[34] M. Vogelius and D. Volkov, *Asymptotic formulas for perturbations in the electromagnetic fields due to the presence of inhomogeneities of small diameter*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 723–748.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 4, pp. 1212–1228

NECESSARY AND SUFFICIENT CONDITIONS FOR OPTIMAL OFFERS IN ELECTRICITY MARKETS*

EDWARD J. ANDERSON† AND HUIFU XU†

Abstract. In this paper, we consider the optimal policy for a generator offering power into a wholesale electricity market operating under a pool arrangement. Anderson and Philpott [*Math. Oper. Res.*, 27 (2002), pp. 82–100] recently discussed necessary conditions for an optimal offer curve when there is uncertainty in the demand and in the behavior of other participants in the market. They show that the objective function in these circumstances can be expressed as a line integral along the offer curve of a profit function integrated with respect to a market distribution function. In this paper, we prove the existence of an optimal offer stack, and we extend the analysis of [*Math. Oper. Res.*, 27 (2002), pp. 82–100] to include necessary conditions of a higher order in the presence of horizontal and/or vertical sections in an offer curve. Finally, we establish sufficient conditions for an offer curve to be locally optimal.

Key words. electricity markets, optimal offer, necessary conditions, sufficient conditions

AMS subject classifications. 90C46, 65K10, 49K30

PII. S0363012900367801

1. Introduction.

In the past few years, there has been an enormous change in the way that wholesale prices for electricity are determined in many parts of the world. Increasingly, market mechanisms are being set up in which the clearing price for electricity is determined by some sort of auction process. The book by Chao and Huntington [3] gives a useful overview of the nature of electricity markets, and the working paper by von der Fehr and Harbord [4] is another useful starting point as it reviews the form of a variety of markets as they existed in 1998. In this paper, we consider a model of the operation of such a market which captures some important features of the electricity markets which operate in the UK, Australia, New Zealand, and parts of the US. The model was introduced in a recent paper by Anderson and Philpott [1]. We begin by reviewing this model before going on to extend the results of Anderson and Philpott on the form of the optimal solutions for generators operating in a market of this sort.

Generators in an electricity market offer energy into the market at prices that they determine. We take this offer as having the form of an offer curve linking quantity and price. We can write the quantity of electricity offered as a nondecreasing function of price, $S(p)$. $S(p)$ is the quantity of electricity that will be delivered by the generator in question if the clearing price is p. In some markets, power is offered in blocks, which will imply that $S(p)$ has the form of a step function. The clearing price is determined by a market mechanism that incorporates consideration of the offers made by all the generators, transmission constraints operating within the electricity network, and the demand (which may have price dependence for some large consumers).

The model we consider represents a market operating under pool arrangements. In some markets, there are more bilateral trading arrangements, and our model will

*Received by the editors February 11, 2000; accepted for publication (in revised form) December 4, 2001; published electronically October 29, 2002. This work was supported by Australian Research Council grant RMG1965.

http://www.siam.org/journals/sicon/41-4/36780.html

†Australian Graduate School of Management, The University of New South Wales, Sydney, NSW 2052, Australia (eddiea@agsm.edu.au, huifux@agsm.edu.au).

not apply in these cases. In a pool arrangement, some form of independent system operator (ISO) is responsible for determining which generators are dispatched and will do this in a way which satisfies demand at least cost, using the generator bids as proxies for cost. If the spot market consists of a pool located at a single node, then the price of electricity can be computed by successively dispatching generation from the offers with the lowest price until all of the demand is met. The price of the marginal offer, the system marginal price, is then the price that is paid for all the electricity dispatched. Thus a low cost generator can choose to offer power at its real marginal cost and will be very likely dispatched and paid a substantial premium over its marginal generating cost.

Following Anderson and Philpott [1], we will be interested in finding an optimal offer for a generator when there is uncertainty about the demand and the behavior of the other generators. This takes a different approach than that considered by Gross and Finlay [6], who assume perfect competition so that the clearing price is unaffected by any single generator's offer. In our model, the offer that a generator makes has a direct influence on the clearing price, but we will not consider the equilibrium framework that would arise if we were to consider the optimal response of other generators to our offer curve. In fact, the majority of papers dealing with electricity markets have studied equilibria, often with the aim of assessing the degree of market power implied by different market structures. See, for example, the papers by Hobbs [7], Bolle [2], Green and Newbery [5], Rudkevich, Duckworth, and Rosen [8], and Wei and Smeers [9].

Our work is more directly concerned with the problem faced by a generator in deciding on an offer. In this case, the generator might take an equilibrium solution as an indication of where things may end up in the long term, but the immediate problem will be to respond to the current environment, which will involve uncertainty in demand and in other generators' offers. Such uncertainty arises partly because of possible outages and partly because generators' bidding behavior is not stable. Concentration only on an equilibrium solution is problematic: this assumes that other generators are behaving in a fully rational way and that we have access to all relevant information. Moreover, in many circumstances, there will be more than one equilibrium possible; which one should guide the offer behavior for a generator? Finally, we should observe that the actual computation of equilibria in these markets with many participants is extremely hard. Of course, since offers are repeated many times a day and there is often considerable similarity between outcomes at the same hour on different days, we should expect electricity markets to move toward some sort of equilibrium behavior. Thus we are not arguing here against the importance of understanding equilibria but merely that the problem we address is of importance. Indeed, a good understanding of this problem will be helpful in moving forward in the analysis of equilibrium models.

Anderson and Philpott have explored the problem of finding an offer curve that maximizes the expected value of the profit made by an individual generator. The offer curve is simply a monotonic continuous curve in the two-dimensional (quantity, price) space. This curve need not be smooth; indeed, it will often in practice take the form of a series of steps. Anderson and Philpott show how the problem of maximizing expected profit is, in some circumstances, equivalent to maximizing the line integral along the offer curve of a market distribution function defined in the (quantity, price) space. The market distribution function captures all the elements of uncertainty in either demand or the other players' behavior.

Anderson and Philpott give necessary conditions for an optimal offer curve in this framework. In this paper, we make a number of contributions. First, in section 2, we demonstrate the existence of an optimal solution. Second, in section 3, we extend the necessary conditions given previously. In section 4, we give sufficient conditions for an offer curve to be locally optimal. Finally, in section 5, we give an example to demonstrate the application of these conditions.

The analysis we give is quite general and has some interest apart from its electricity market context. The necessary and sufficient conditions apply to the problem of finding a choice of monotonic curve within a bounded region in order to maximize a line integral along the curve. In the absence of monotonicity constraints, this is a problem in the calculus of variations. However, the requirement for monotonicity is fundamental and implies additional conditions that apply on sections of the curve which are either horizontal or vertical (and hence at points where monotonicity acts as a binding constraint). The results we give are proved by relatively direct and elementary methods involving consideration of perturbations of an offer curve. It is interesting that sufficient conditions for optimality can be obtained in this way.

2. Problem formulation and fundamentals. In this section, we will introduce some notation and formulate the problem that we shall consider. Let $R(q,p)$ be the return function: that is, $R(q,p)$ denotes the profit we make if we are dispatched an amount q at a clearing price p. We will assume that R has continuous partial derivatives. This function captures not only the cost of generating an amount q and the proceeds pq which arise from the sale of this electricity but also the effects of any hedging contracts the generator holds which depend on the market clearing price.

Rather than dealing with a supply function $S(\cdot)$ directly, it is convenient to model the offer using a continuous curve $\mathbf{s} = \{(x(t), y(t)), 0 \le t \le T\}$, in which the components $x(t)$ and $y(t)$ are continuous monotonic increasing functions of t, and $x(t)$ and $y(t)$ trace, respectively, the quantity and price components. Without loss of generality we may take $x(0) = y(0) = 0$ and $y(T) \le p_M$, where p_M is a bound on the price of any offer. We also assume that q_m is a bound on the generation capacity of the generator, so $x(T) \le q_M$.

We use a single *market distribution function* $\psi(q,p)$ to describe the uncertainty in the market. $\psi(q,p)$ is defined as the probability of not being fully dispatched if we offer generation q at price p. It turns out that knowledge of the single function $\psi(q,p)$ is enough to determine the expected profit for a generator. In practice, a generator will estimate the market distribution function from knowledge of the distribution of demand and from repeated observations of the behavior of other generators. This estimation problem will depend on the information released to market participants about other players' bids (something which varies between different markets). Another issue will be to decide the class of functions from which an estimate is to be chosen. However, subject to these considerations, either Bayesian or maximum likelihood estimation techniques can be used.

Since $\psi(q,p)$ is a probability, it takes values between 0 and 1. Let $\Psi = \{(q,p), 0 < \psi(q,p) < 1\}$. Throughout, we assume that ψ is continuously differentiable on $\Psi \cap \{(q,p), 0 \le q \le q_M, 0 \le p \le p_M\}$. The first key result, which we repeat from Anderson and Philpott [1], demonstrates that the expected return if a generator offers in a supply curve $\mathbf{s}$ can be expressed as a line integral along $\mathbf{s}$. This can be established by showing that a generator can only be dispatched at price-quantity points lying on its offer curve and observing that the derivative of the market distribution function ψ captures the appropriate probability density.

LEMMA 1 (see [1, Theorem 2]). *If a generator offers in a curve* $\mathbf{s}$ *and the market distribution function* ψ *is continuous, then the expected return is the line integral*

$$v(\mathbf{s}) = \int_{\mathbf{s}} R(q,p)d\psi(q,p).$$

Anderson and Philpott [1] treat $v(\mathbf{s})$ as an objective function and investigate the necessary conditions for an offer curve $\mathbf{s}$ to be a local maximizer. However, there is an important question that they do not address explicitly: does there exist a maximum over the set of curves which are considered? We begin by answering this question before discussing optimality conditions.

A generator need not offer all of its generation capacity into the market; the offer curve will start at some point $(0, y(0))$ and finish at $(x(T), y(T))$. However, the clearing price is determined as though the offer curve began with a vertical segment from the origin to $(0, y(0))$ and finished with a vertical segment from $(x(T), y(T))$ to $(x(T), p_M)$. Hence we assume that Ω, the set of curves, has these characteristics.

LEMMA 2. *Let* Ω *be the set of monotonic continuous curves starting at the origin and ending on the closed line segment,* $\mathcal{L}$*, from* $(0, p_M)$ *to* (q_M, p_M)*. Then* Ω *is compact under the Hausdorff metric:*

$$|\mathbf{s}_1 - \mathbf{s}_2|_H = \max_{(x_1,y_1)\in\mathbf{s}_1} \min_{(x_2,y_2)\in\mathbf{s}_2} \sqrt{(x_1-x_2)^2 + (y_1-y_2)^2}.$$

Proof. Let $\mathbf{s} \in \Omega$ and $L(\mathbf{s})$ be the arc length of $\mathbf{s}$. By the monotonicity of $\mathbf{s}$,

$$p_M \le L(\mathbf{s}) \le p_M + q_M.$$

For any $\mathbf{s} \in \Omega$, we may use the arc length measured from the origin as a parameter and write $\mathbf{s} = \{(x(t), y(t)), 0 \le t \le L(\mathbf{s})\}$. For $0 \le t_1 < t_2 \le L(\mathbf{s})$, we have

$$t_2 - t_1 \ge \max\left(x(t_2) - x(t_1), y(t_2) - y(t_1)\right).$$

Thus both x and y are Lipschitz with respect to t. By replacing t with $\frac{T}{L(\mathbf{s})}t$, we obtain a representation of $\mathbf{s}$ with both x and y defined on $[0, T]$. The scaled x and y are still Lipschitz and monotonic. It is well known that the class of monotonic Lipschitz functions defined on $[0, T]$ forms a compact set. Thus the set of curves Ω, when represented as pairs of (Lipschitz) functions, is compact with the metric

$$|\mathbf{s}_1 - \mathbf{s}_2| = \max\left(\sup_{0\le t\le T} |x_1(t) - x_2(t)|, \sup_{0\le t\le T} |y_1(t) - y_2(t)|\right).$$

Since $|s_1 - s_2|_H \le \sqrt{2}|s_1 - s_2|$, this implies compactness of Ω with the Hausdorff metric. □

Anderson and Philpott [1] introduced the notation

$$Z(q,p) \equiv \begin{cases} R_q\psi_p - R_p\psi_q, & (q,p) \in \Psi, \\ 0 & \text{otherwise} \end{cases}$$

and observed that

$$\int\int_{\mathcal{S}} Z(q,p)dpdq = \int_{\mathcal{C}} R(q,p)d\psi(q,p), \tag{1}$$

where $\mathcal{S}$ is a region enclosed by a curve $\mathcal{C}$. Relation (1) follows immediately from Green's theorem and plays an important role in the investigation of optimality conditions. The scalar function Z can be thought of as indicating the direction in which the offer curve needs to move to produce an improvement in expected profit. If $Z > 0$, then a move of the offer curve down and to the left in the (q, p) plane will produce an improvement, while $Z < 0$ indicates that the offer curve should move up and to the right. Consequently, when monotonicity constraints are not binding, it will be optimal to follow a $Z = 0$ curve.

Using Lemma 2 and (1), we are able to obtain the following result.

THEOREM 3. *Let Ω be defined as above, and let v be the expected return function given in Lemma* 1. *Then v achieves its maximum on Ω.*

Proof. By Lemma 2, Ω is a compact set. It suffices to prove that v is continuous with the Hausdorff metric. Let $\tilde{\mathbf{s}} \in \Omega$, and assume without loss of generality that $\mathcal{S}$ is the region surrounded by $\tilde{\mathbf{s}}$, $\mathbf{s}$. Thus, as $\tilde{\mathbf{s}}$ is close to $\mathbf{s}$, the area $\mathcal{S}$ is small. By (1), the boundedness of Ψ, and the continuous differentiability ψ and R, we know that the integral of Z over $\mathcal{S}$ will be arbitrarily small when the area of $\mathcal{S}$ shrinks to zero. On the other hand, the line integral at the right side of (1) over the segment of $\mathcal{L}$ between $\mathbf{s}$ and $\tilde{\mathbf{s}}$ tends to zero as $\tilde{\mathbf{s}} \to \mathbf{s}$. This implies that $v(\tilde{\mathbf{s}})$ tends to $v(\mathbf{s})$. □

Given this theorem, the maximization problem

$$\text{maximize } v(\mathbf{s}), \quad \text{subject to } \mathbf{s} \in \Omega$$

is well defined. In the rest of this paper, we will discuss the necessary and sufficient conditions for an offer curve $\mathbf{s}$ to be a local maximizer.

3. Necessary conditions. We turn now to necessary conditions for optimality. Anderson and Philpott [1] establish a set of conditions which we will extend. Such conditions are important in the development of algorithms to solve the generator's maximization problem. Later we will illustrate their use by considering a simple example. In general, more complete optimality conditions will serve to eliminate more potential candidate optimal offer curves. It is convenient to restate the result of [1] in a slightly different form in order to show how it is related to the results that we prove in this paper.

Throughout, we need to use the line integral of Z along a curve $\{(x(t), y(t)) : 0 \leq t \leq T\}$ which is defined by

$$w(t) = \int_0^t Z(x(\tau), y(\tau))(x'(\tau) + y'(\tau))d\tau.$$

In some cases, we will write $w(x(t), y(t))$ to denote $w(t)$ for clarity.

THEOREM 4. *Suppose that $\mathbf{s} = \{x(t), y(t), 0 \leq t \leq T\}$ is an increasing continuous offer curve. Suppose that there exist m numbers $0 \leq t_1 < t_2 < \cdots < t_m \leq T$ with $0 < x(t) < q_M$ and $0 < y(t) < p_M$ for $t_1 < t < t_m$ and such that, on each section (t_{i-1}, t_i), $i = 2, \ldots, m$, $\mathbf{s}$ is either strictly increasing in both components or horizontal or vertical, with different characteristics in successive segments. If $\mathbf{s}$ is optimal, then each of the $w(t_i)$, $i = 1, 2, \ldots, m$, takes the same value, say, w_0, and, for each interval I, being one of (t_{i-1}, t_i), $i = 2, \ldots, m$, $(0, t_1)$, or (t_m, T), one of the following holds:*

(i) $\mathbf{s}$ *is strictly increasing in both components, and $Z(x(t), y(t)) = 0$ for $t \in I$;*

(ii) $\mathbf{s}$ *is horizontal on I, and $w(t) \leq w_0$ for $t \in I$;*

(iii) $\mathbf{s}$ *is vertical on I, and $w(t) \geq w_0$ for $t \in I$.*

Anderson and Philpott [1] show, in addition, that, when $\mathbf{s}$ is neither horizontal nor vertical, then there are sign constraints on the partial derivatives of Z if they exist. In fact, the existence of partial derivatives for Z will allow us to extend the results of Theorem 4 provided that Z is well behaved enough.

We will need to make an assumption about the partial derivatives of Z. We give the weakest form of this required for our results; it is stronger than continuous differentiability, requiring also a uniformity condition on horizontal and vertical sections of $\mathbf{s}$. This condition will be implied, for example, by Z having continuous second derivatives.

Assumption 1. Z is continuously differentiable on Ψ. If $\mathbf{s}$ is horizontal on $[t_{i-1}, t_i]$, then, for every $\eta > 0$, there is a $\tau_0 > 0$ with

$$|Z(x(t), y(t_i) + \tau) - Z(x(t), y(t_i)) - Z_p(x(t), y(t_i))\tau| \leq \eta|\tau|$$

for every $t \in [t_{i-1}, t_i]$ and $|\tau| < \tau_0$. Similarly, if $\mathbf{s}$ is vertical on $[t_{i-1}, t_i]$, then, for every $\eta > 0$, there is a $\tau_0 > 0$ with

$$|Z(x(t_i) + \tau, y(t)) - Z(x(t_i), y(t)) - Z_q(x(t_i), y(t))\tau| \leq \eta|\tau|$$

for every $t \in [t_{i-1}, t_i]$ and $|\tau| < \tau_0$.

The theorem below extends Theorem 4, but there is a key difference that is worth pointing out before stating the result. In the previous theorem, the values t_i are defined as corresponding to points where the curve changes characteristic, say, from horizontal to vertical. In the theorem we give next, we will define the values t_i in terms of the w values instead. Thus suppose we have a horizontal segment within which $w(t) \leq w_0$: then, in Theorem 4, the t_i values mark either end of this segment, but in the next theorem we add to this any other values of t at which $w(t) = w_0$, between the end points. Thus we may have a point t_i, with $\mathbf{s}$ horizontal on either side of it. It will be convenient to distinguish those values of t_i such that the curve $\mathbf{s}$ changes its characteristics at $(x(t_i), y(t_i))$, for instance, from vertical to horizontal or strictly increasing in both components. For convenience, we call both the parameter t_i and the point $(x(t_i), y(t_i))$ a *turning point.*

THEOREM 5. *Suppose that* $\mathbf{s} = \{x(t), y(t), 0 \leq t \leq T\}$ *is an increasing continuous offer curve. Suppose that there exist* m *numbers* $0 \leq t_1 < t_2 < \cdots < t_m \leq T$ *such that each of the* $w(t_i)$, $i = 1, 2, \ldots, m$, *takes the same value, say,* w_0, *and, on each section* $[t_{i-1}, t_i]$, $i = 2, \ldots, m$, $\mathbf{s}$ *is either strictly increasing in both components or horizontal or vertical. Suppose that* $\mathbf{s}$ *is an optimal offer stack. Then, under Assumption* 1, *for each section* (t_{i-1}, t_i), *the following hold:*

(i) *if* $\mathbf{s}$ *is strictly increasing on* (t_{i-1}, t_i), *then* $Z_p(x(t), y(t)) \geq 0$ *and* $Z_q(x(t), y(t)) \leq 0$ *for* $t \in (t_{i-1}, t_i)$;

(ii) *if* $\mathbf{s}$ *is horizontal on* (t_{i-1}, t_i) *and one of* t_{i-1} *or* t_i *is a turning point, then*

$$\int_{x(t_{i-1})}^{x(t_i)} Z_p(x, y(t_i))dx \geq 0; \tag{2}$$

further, if $\mathbf{s}$ *turns from horizontal to vertical at* t_i, *then*

$$\int_{x(t_{i-1})}^{x(t_i)} Z_p(x, y(t_i))dx - \int_{y(t_i)}^{y(t_{i+1})} Z_q(x(t_i), y)dy \geq 2Z(x(t_i), y(t_i)); \tag{3}$$

(iii) *if* $\mathbf{s}$ *is vertical on* (t_{i-1}, t_i) *and one of* t_{i-1} *or* t_i *is a turning point, then*

$$\int_{y(t_{i-1})}^{y(t_i)} Z_q(x(t_i), y)dy \leq 0; \tag{4}$$

further, if $\mathbf{s}$ *turns from vertical to horizontal at* t_i*, then*

$$\int_{y(t_{i-1})}^{y(t_i)} Z_q(x(t_i), y)dy - \int_{x(t_i)}^{x(t_{i+1})} Z_p(x, y(t_i))dx \leq 2Z(x(t_i), y(t_i)). \tag{5}$$

Before giving a proof of this result, it may be helpful to discuss the conditions (2), (3), (4), and (5).

If $\mathbf{s}$ is horizontal on (t_{i-1}, t_i), then, since $w(t_{i-1}) = w(t_i) = w_0$,

$$\int_{x(t_{i-1})}^{x(t_i)} Z(x, y(t_i))dx = 0.$$

This integral captures the first order effect of a move of the horizontal segment up or down by a small amount. Given that the integral is zero, we can look at the second order effects of the same move, and these are given by the integral in (2). The same argument applied to a vertical segment leads to the integral in (4).

The stronger conditions (3) and (5), which apply when there is a turn from horizontal to vertical (or vice versa), arise from considering a move of a horizontal segment at the same time as a vertical segment. Notice that, if $\mathbf{s}$ turns from horizontal to vertical at t_i, then $Z(x(t_i), y(t_i)) \geq 0$. This follows from the observation that Theorem 4 implies that $w(t)$ is increasing at $t = t_i$. In the same way, $Z(x(t_i), y(t_i)) \leq 0$ if there is a turn from vertical to horizontal at t_i.

Proof. Part (i) was already established by Anderson and Philpott [1]. We will prove part (ii); part (iii) follows in exactly the same way. For the sake of contradiction, assume that (2) does not hold and $\mathbf{s}$ is horizontal on (t_{i-1}, t_i) but not horizontal on (t_i, t_{i+1}). Let $\delta > 0$ be sufficiently small, and consider a vertical perturbation of $\mathbf{s}$ by an amount δ between t_{i-1} and t_i. We can make this explicit as follows. Define $t_i(+\delta) = y^{-1}(y(t_i) + \delta)$, and let

$$\tilde{\mathbf{s}}(t) = \begin{cases} (x(t), y(t)), & t \leq t_{i-1}, \\ (x(t_{i-1}), t - t_{i-1} + y(t_{i-1})), & t_{i-1} \leq t \leq t_{i-1} + \delta, \\ (x(t-\delta), y(t_{i-1}) + \delta), & t_{i-1} + \delta \leq t \leq t_i(+\delta) + \delta, \\ (x(t-\delta), y(t-\delta)), & t_i(+\delta) + \delta \leq t \leq T + \delta, \end{cases}$$

be a perturbation of $\mathbf{s}$. This is illustrated in Figure 1. For $0 \leq \tau \leq \delta$, we consider the line integral

$$I(\tau) = \int_{x(t_{i-1})}^{x(t_i)} Z(x, y(t_i) + \tau)dx.$$

Let $\epsilon > 0$ be such that $\int_{x(t_{i-1})}^{x(t_i)} Z_p(x, y(t_i))dx < -\epsilon < 0$. By Assumption 1, for τ sufficiently small and $x \in [x(t_{i-1}), x(t_i)]$,

$$Z(x, y(t_i) + \tau) \leq Z(x, y(t_i)) + Z_p(x, y(t_i))\tau + \frac{\tau\epsilon}{x(t_i) - x(t_{i-1})}.$$

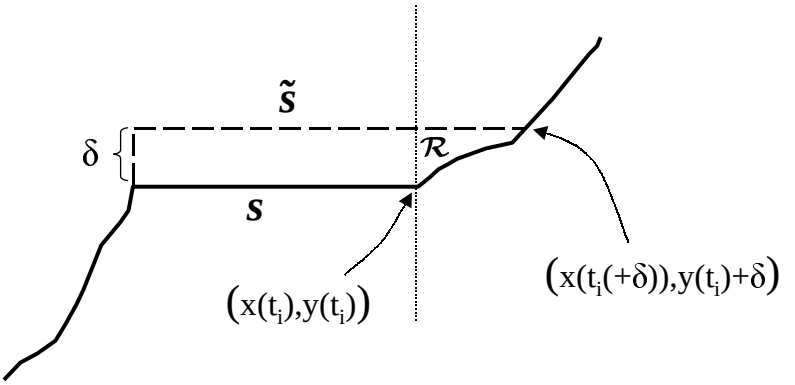


FIG. 1. *Perturbation of a horizontal section of* **s**.

Thus

$$I(\tau) \le w(t_i) - w(t_{i-1}) + \tau \int_{x(t_{i-1})}^{x(t_i)} Z_p(x, y(t_i))dx + \tau\epsilon.$$

Consequently, we have, for all τ, $I(\tau) < 0$ when δ is sufficiently small. The area integral of Z over the region surrounded by $\mathbf{s}$, $\tilde{\mathbf{s}}$, $x = x(t_{i-1})$, and $x = x(t_i)$ can be written as $\int_0^\delta I(\tau)d\tau$, which is not larger than $\frac{\delta^2}{2}[\int_{x(t_{i-1})}^{x(t_i)} Z_p(x, y(t_i))dx + \epsilon]$.

Now we need to consider the region, namely, $\mathcal{R}$, to the right of $x = x(t_i)$ surrounded by $\mathbf{s}$ and $\tilde{\mathbf{s}}$. This area exists only when $\mathbf{s}$ is strictly increasing in each component on the interval (t_i, t_{i+1}). In this case, the area is of order $o(\delta)$. On the other hand, since $Z(x(t), y(t)) = 0$ along the lower boundary of the region,

$$Z(x(t), y(t_i) + \tau) = Z_p(x(t), y(t))(y(t_i) + \tau - y(t)) + o(\tau) < Z_p(x(t), y(t))\tau + o(\tau)$$

for all $t_i \le t \le t_i(+\delta)$, $0 \le \tau \le \delta$, such that $(x(t), y(t_i) + \tau) \in \mathcal{R}$. Note that Z_p is continuous, and the distance between any point $(x, y) \in \mathcal{R}$ and $(x(t_i), y(t_i))$ tends to 0 as $\delta \to 0$. Consequently, the integral of Z over $\mathcal{R}$ is at most of order $o(\delta^2)$. Thus we have

$$v(\mathbf{s}) - v(\tilde{\mathbf{s}}) \le \int_0^\delta I(\tau)d\tau + o(\delta^2) = \frac{\delta^2}{2} \int_{x(t_{i-1})}^{x(t_i)} Z_p(x, y(t_i))dx + \frac{\delta^2}{2}\epsilon + o(\delta^2)$$

for δ small enough. From the choice of ϵ, this contradicts the optimality of $\mathbf{s}$ for δ small enough.

In the case that $\mathbf{s}$ is not horizontal on (t_{i-2}, t_{i-1}), we can obtain the same result by considering a vertical perturbation of $\mathbf{s}$ between t_{i-1} and t_i downward by an amount δ.

Now suppose that t_i is a corner where $\mathbf{s}$ turns from horizontal to vertical. We will assume that (3) does not hold and derive a contradiction. Let η be a scalar such that

$$(6)\quad \int_{x(t_{i-1})}^{x(t_i)} Z_p(x, y(t_i))dx - \int_{y(t_i)}^{y(t_{i+1})} Z_q(x(t_i), y)dy - 2Z(x(t_i), y(t_i)) < -2\eta < 0.$$

We consider a perturbation that moves the horizontal section between t_{i-1} and t_i upward by a small amount δ and the vertical section between t_i and t_{i+1} to the

left by the same amount δ. We can make this explicit as follows. Define $t_i(-\delta) = x^{-1}(x(t_i) - \delta)$, and let

$$\tilde{\mathbf{s}}(t) = \begin{cases} (x(t), y(t)), & t \le t_{i-1}, \\ (x(t_{i-1}), t - t_{i-1} + y(t_{i-1})), & t_{i-1} \le t \le t_{i-1} + \delta, \\ (x(t-\delta), y(t_{i-1}) + \delta), & t_{i-1} + \delta \le t \le t_i(-\delta) + \delta, \\ (x(t_i) - \delta, y(t + k - \delta)), & t_i(-\delta) + \delta \le t \le t_{i+1} - k + \delta, \\ (x(t_{i+1}) + t - 2\delta + k - t_{i+1}, y(t_{i+1})), & t_{i+1} - k + \delta \le t \le t_{i+1} - k + 2\delta, \\ (x(t + k - 2\delta), y(t + k - 2\delta))), & t_{i+1} - k + 2\delta \le t \le T - k + 2\delta, \end{cases}$$

be a perturbation of $\mathbf{s}$, where $k = t_i(+\delta) - t_i(-\delta)$.

First observe that

$$\begin{aligned} w(t_i(-\tau)) - w(t_i(+\tau)) &= -\int_{x(t_i)-\tau}^{x(t_i)} Z(x, y(t_i))dx - \int_{y(t_i)}^{y(t_i)+\tau} Z(x(t_i), y)dy \\ &\le -2\tau Z(x(t_i), y(t_i)) + \eta\tau/2 \end{aligned} \tag{7}$$

for τ sufficiently small, using the mean value theorem and the continuity of Z. On the other hand, under Assumption 1, for the given η, there exists $\delta > 0$ sufficiently small such that, for $0 < \tau \le \delta$, the line integral

$$\begin{aligned} I(\tau) &\equiv \int_{x(t_{i-1})}^{x(t_i)-\tau} Z(x, y(t_i) + \tau)dx + \int_{y(t_i)+\tau}^{y(t_{i+1})} Z(x(t_i) - \tau, y)dy \\ &\le w(t_i(-\tau)) - w(t_{i-1}) + w(t_{i+1}) - w(t_i(+\tau)) \\ &\quad + \tau \int_{x(t_{i-1})}^{x(t_i)-\tau} Z_p(x, y(t_i))dx - \tau \int_{y(t_i)+\tau}^{y(t_{i+1})} Z_q(x(t_i), y)dy + \eta\tau/2. \end{aligned} \tag{8}$$

Since Z_p and Z_q are continuous, it follows from (6) that, for δ sufficiently small and $\tau \le \delta$,

$$\int_{x(t_{i-1})}^{x(t_i)-\tau} Z_p(x, y(t_i))dx - \int_{y(t_i)+\tau}^{y(t_{i+1})} Z_q(x(t_i), y)dy - 2Z(x(t_i), y(t_i)) < -\eta. \tag{9}$$

Combining (7)–(9) and noticing that $w(t_{i-1}) = w(t_{i+1}) = w_0$, we have $I(\tau) < 0$ for δ sufficiently small and all $0 \le \tau \le \delta$.

Now the area integral of Z over the region surrounded by $\mathbf{s}$ and $\tilde{\mathbf{s}}$ can be written as $\int_0^\delta I(\tau)d\tau$ and hence is strictly negative. Thus we have constructed a perturbed curve which leads to a larger value for v, which is a contradiction. This completes the proof. □

4. Sufficient conditions. In this section, we discuss sufficient conditions for an offer curve $\mathbf{s}$ to be locally optimal. This involves a more complex set of conditions than were required for the necessary conditions. To establish this result, we will have to consider all possible monotonic perturbations around the offer curve $\mathbf{s}$. As we shall see, considerable care is needed in the argument required to prove this result.

THEOREM 6. *Let $\mathbf{s} = \{x(t), y(t), 0 \le t \le T\}$ be an increasing and continuous offer stack. Suppose that there exist finite numbers $0 = t_0 < t_1 < t_2 < \cdots < t_M = T$ such that, for $i = 1, \ldots, M-1$, $w(t_i)$ takes a common value, say, w_0, and on each*

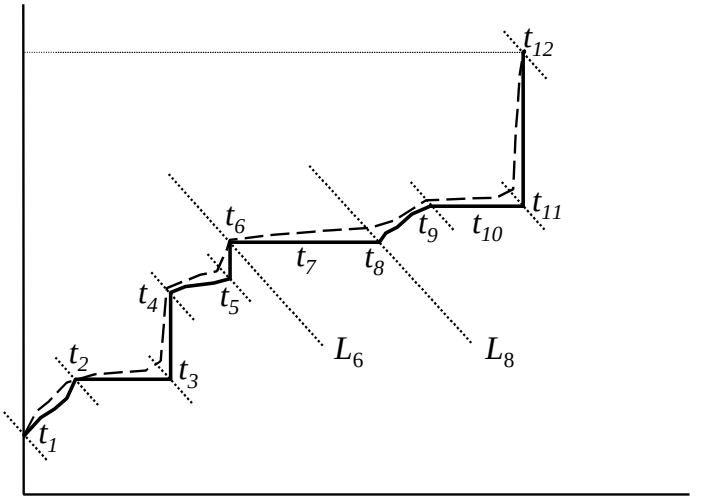


FIG. 2. *Area of perturbation split into subregions.*

section (t_{i-1}, t_i), $i = 1, \ldots, M$, $\mathbf{s}$ *is either strictly increasing in both components or horizontal or vertical. Suppose also that Assumption* 1 *and the following hold:*

(i) *if* $\mathbf{s}$ *is increasing in both components on* (t_{i-1}, t_i), *then, for* $t \in (t_{i-1}, t_i)$, $Z(x(t), y(t)) = 0$, $Z_p(x(t), y(t)) > 0$, *and* $Z_q(x(t), y(t)) < 0$;

(ii) *if* $\mathbf{s}$ *is horizontal on* (t_{i-1}, t_i), *then, for* $t \in (t_{i-1}, t_i)$, $w(t) < w_0$; *moreover, for any* $j < k$ *such that* $\mathbf{s}$ *is horizontal from* t_j *to* t_k *with at least one of* t_j *and* t_k *a turning point,*

$$\int_{x(t_j)}^{x(t_k)} Z_p(x, y(t_j))dx > Z(x(t_k), y(t_k)) - Z(x(t_j), y(t_j)); \tag{10}$$

(iii) *if* $\mathbf{s}$ *is vertical on* (t_{i-1}, t_i), *then, for* $t \in (t_{i-1}, t_i)$, $w(t) > w_0$; *moreover, for any* $j < k$ *such that* $\mathbf{s}$ *is vertical from* t_j *to* t_k *with at least one of* t_j *and* t_k *a turning point,*

$$\int_{y(t_j)}^{y(t_k)} Z_q(x(t_j), y)dy < Z(x(t_k), y(t_k)) - Z(x(t_j), y(t_j)); \tag{11}$$

(iv) *if* $(x(t_i), y(t_i))$ *is a point where the curve turns from horizontal to vertical, then either* $Z(x(t_i), y(t_i)) > 0$ *or* $Z(x(t_i), y(t_i)) = 0$ *and* $Z_p(x(t_i), y(t_i)) > 0$, $Z_q(x(t_i), y(t_i)) < 0$; *if* $(x(t_i), y(t_i))$ *is a point where the curve turns from vertical to horizontal, then either* $Z(x(t_i), y(t_i)) < 0$ *or* $Z(x(t_i), y(t_i)) = 0$ *and* $Z_p(x(t_i), y(t_i)) > 0$, $Z_q(x(t_i), y(t_i)) < 0$.

Then s *is a locally optimal offer stack.*

Proof. In order to prove that $\mathbf{s}$ is locally optimal, it suffices to prove that, for any local perturbation $\tilde{\mathbf{s}}$ which is sufficiently close to $\mathbf{s}$, $v(\tilde{\mathbf{s}}) < v(\mathbf{s})$. By Green's theorem, this is equivalent to proving that the area integral of Z over any region surrounded by $\tilde{\mathbf{s}}$ and $\mathbf{s}$ is positive if the region is above or to the left of the curve $\mathbf{s}$ and negative if under or to the right of $\mathbf{s}$. We discuss only the case that the perturbed region is above or to the left of the curve $\mathbf{s}$, and the other case can be dealt with similarly.

We define a line $y = y(t_i) - (x - x(t_i))$ at every turning point and call this line L_i. Figure 2 illustrates this with $\mathbf{s}$ shown bold and $\tilde{\mathbf{s}}$ as a dashed line. Assume the

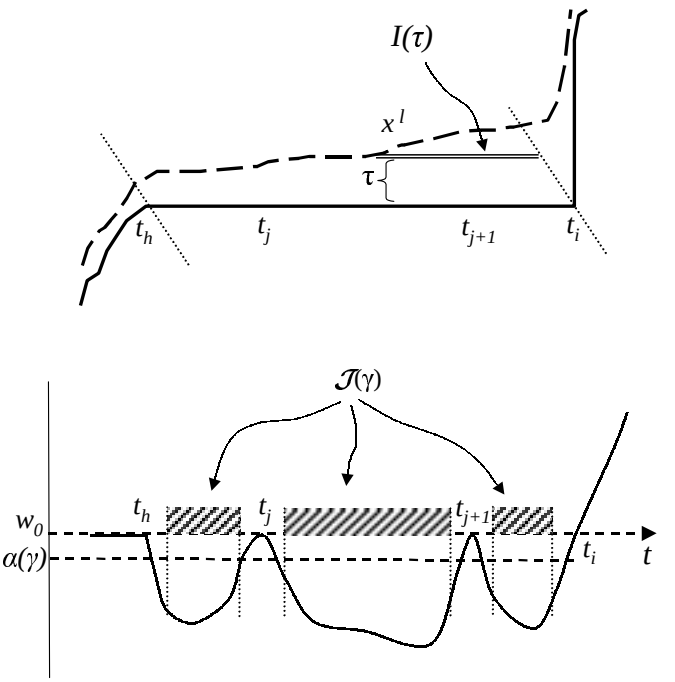


FIG. 3. *Comparison between* w *and* $\mathbf{s}$.

maximum distance between $\mathbf{s}$ and $\tilde{\mathbf{s}}$ is no larger than δ_0. The idea of the proof is to deal separately with different parts of the region between $\mathbf{s}$ and $\tilde{\mathbf{s}}$. In fact, we divide this region into subregions using the lines L_i and show that the area integral of Z is positive over each nonempty subregion. To accomplish this, we will consider integrals along horizontal (or vertical) segments with an end point on one of the lines L_i and show that each of these has positive value. One of these integrals $I(\tau)$ is illustrated in Figure 3. Since the perturbation is monotonic, the horizontal line segments will finish on an L_i, while the vertical line segments will start on an L_i.

Suppose first that $\mathbf{s}$ is strictly increasing in both components on (t_{i-1}, t_i). By assumption (i), it is easy to verify that $Z(x, y) > 0$ for any (x, y) within the region surrounded by $\mathbf{s}$, $\tilde{\mathbf{s}}$, and lines L_{i-1} and L_i provided δ_0 is chosen small enough. Consequently, the area integral of Z over the region surrounded by $\mathbf{s}$, $\tilde{\mathbf{s}}$, and these lines, if not empty, is positive.

Now we turn to the central part of the proof, and we consider the case that $\mathbf{s}$ is horizontal between t_h and t_i, and $(x(t_h), y(t_h))$ and $(x(t_i), y(t_i))$ are turning points. We observe first from (i) and (iv) that

$$Z(x(t_h), y(t_h)) \leq 0 \tag{12}$$

and

$$Z(x(t_i), y(t_i)) \geq 0. \tag{13}$$

We will need to keep track of the parameter t at points on $\mathbf{s}$ which are a distance γ to the right or left of one of the t_i values in this horizontal segment. We let $t_k(-\gamma)$ denote $x^{-1}(x(t_k) - \gamma)$ for $k = h + 1, \ldots, i$, and $t_j(+\gamma) = x^{-1}(x(t_j) + \gamma)$ for $j = h, \ldots, i - 1$.

The proof proceeds in two major steps. In step 1, we construct a value of δ sufficiently small for the $I(\tau)$ integrals to be positive when $\delta_0 < \delta$; then, in step 2, we demonstrate this inequality.

Step 1. To define δ appropriately, we need first to define some intermediate quantities γ and ϵ. Choose $\gamma > 0$ small enough so that, for all $x \in (x(t_i) - \gamma, x(t_i))$,

$$Z(x, y(t_i)) > 0,\ Z_p(x, y(t_i)) > 0, \tag{14}$$

and for $x^u \in [x(t_i) - \gamma, x(t_i)]$, $x^l \in [x(t_j) - \gamma, x(t_j) + \gamma]$, $j = h+1, \ldots, i-1$,

$$\int_{x^l}^{x^u} Z_p(x, y(t_i))dx - Z(x(t_i), y(t_i)) > 0, \tag{15}$$

and for $x^l \in [x(t_h), x(t_h) + \gamma]$,

$$\int_{x^l}^{x^u} Z_p(x, y(t_i))dx + Z(x(t_h), y(t_h)) - Z(x(t_i), y(t_i)) > 0. \tag{16}$$

The existence of a γ satisfying (14) is guaranteed by conditions (i) and (iv) and (13), while the existence of a γ satisfying (15) and (16) follows from (10) after observing that $Z(x(t_j), y(t_j)) = 0$ since $w(t_j)$ is a local maximum of $w(\cdot)$ for $j = h+1, \ldots, i-1$.

Given such a γ, choose $\epsilon > 0$ small enough so that

$$w_0 - \alpha(\gamma) > \epsilon,$$

where

$$\alpha(\gamma) = \sup_{x(t) \in \mathcal{J}(\gamma)} w(t),$$

and

$$\mathcal{J}(\gamma) = \bigcup_{h \le j \le i-1} [x(t_j) + \gamma, x(t_{j+1}) - \gamma];$$

and for $x^u \in [x(t_i) - \gamma, x(t_i)]$, $x^l \in [x(t_j) - \gamma, x(t_j) + \gamma]$, $j = h+1, \ldots, i-1$,

$$\int_{x^l}^{x^u} Z_p(x, y(t_i))dx - Z(x(t_i), y(t_i)) > 2\epsilon, \tag{17}$$

and for $x^l \in [x(t_h), x(t_h) + \gamma]$,

$$\int_{x^l}^{x^u} Z_p(x, y(t_i))dx + Z(x(t_h), y(t_h)) - Z(x(t_i), y(t_i)) > 3\epsilon. \tag{18}$$

The definitions of $\alpha(\gamma)$ and $\mathcal{J}(\gamma)$ are illustrated in Figure 3.

Having chosen ϵ and γ, now let $\delta > 0$ be chosen small enough such that the following six conditions (a)–(f) hold.

(a)

$$\delta < \frac{1}{2\epsilon} \left(w_0 - \alpha(\gamma) - \epsilon \right). \tag{19}$$

(b) For $x^l \in \mathcal{J}(\gamma)$ and $x^u \in [x(t_i) - \gamma, x(t_i)]$,

$$\delta \left| \int_{x^l}^{x^u} Z_p(x, y(t_i))dx - Z(x(t_i), y(t_i)) \right| < \epsilon. \tag{20}$$

This follows from the fact that Z_p is continuous.

(c) For $0 \le \tau \le \delta$,

$$w_0 - w(t_i(-\tau)) \le \tau Z(x(t_i), y(t_i)) + \epsilon\tau. \tag{21}$$

This can be proved using continuity of Z and the mean value theorem.

(d) For $0 \le \tau \le \delta$ and $t \in [t_h, t_i]$,

$$Z(x(t), y(t_i) + \tau) \ge Z(x(t), y(t_i)) + Z_p(x(t), y(t_i))\tau - \frac{\epsilon\tau}{(x(t_i) - x(t_h))}. \tag{22}$$

This follows from Assumption 1.

(e) For $x \in (x(t_i) - \gamma, x(t_i))$, $0 \le \tau \le \delta$,

$$Z(x, y(t_i) + \tau) > 0. \tag{23}$$

This follows from (14) and Assumption 1.

(f) For $0 \le \tau_1, \tau_2 \le \delta$,

$$Z(x(t_h) - \tau_1, y(t_h) + \tau_2) > Z(x(t_h), y(t_h)) - \epsilon. \tag{24}$$

This follows from the continuity of Z.

Step 2. For $0 < \tau \le \delta$, we consider the integral

$$I(\tau) = \int_{x^l}^{x(t_i)-\tau} Z(x, y(t_i) + \tau)dx.$$

We shall prove that $I(\tau) > 0$ for all $x^l \in [x(t_h) - \tau, x(t_i) - \tau]$. To do this, we need to consider four cases depending on the position of x^l. When $x^l \in \mathcal{J}(\gamma)$, we will show that the value of the integral $I(\tau)$ is dominated by $I(0)$, which is a similar integral of Z but shifted down by an amount τ. Now $I(0) = w_0 - w(x^l, y(t_i))$, which is positive. However, when x^l is near t_h, t_i, or one of the intermediate t_j, then $I(0)$ is near zero, and we need to consider more precisely the difference between $I(\tau)$ and $I(0)$. This difference will be determined by the integral of Z_p along the line segment.

Suppose first that $x^l \ge x(t_h)$. By (22), we have

$$I(\tau) \ge w(x(t_i) - \tau, y(t_i)) - w(x^l, y(t_i)) + \tau \int_{x^l}^{x(t_i)-\tau} Z_p(x, y(t_i))dx - \epsilon\tau. \tag{25}$$

Case A. For $x^l \in [x(t_h), x(t_i) - \gamma] \backslash \mathcal{J}(\gamma)$, since $w(x^l, y(t_i)) \le w(x(t_i), y(t_i)) = w_0$, it follows from (21) and (25) that

$$I(\tau) \ge \tau \left(\int_{x^l}^{x(t_i)-\tau} Z_p(x, y(t_i))dx - Z(x(t_i), y(t_i)) \right) - 2\epsilon\tau. \tag{26}$$

Thus (using (17), (18), and (12)), $I(\tau) > 0$.

Case B. For $x^l \in \mathcal{J}(\gamma)$, combining (19), (20), (21), and (25), we have

$$\begin{aligned} I(\tau) &\geq w_0 - \alpha(\gamma) + \tau \left(\int_{x^l}^{x(t_i)-\tau} Z_p(x, y(t_i))dx - Z(x(t_i), y(t_i)) \right) - 2\epsilon\tau \\ &\geq w_0 - \alpha(\gamma) - \epsilon - 2\tau\epsilon \\ &> 0. \end{aligned}$$

Case C. For $x^l \in [x(t_i) - \gamma, x(t_i) - \tau]$, $0 \leq \tau \leq \delta$, it follows immediately from (23) that $I(\tau)$ is positive.

Case D. For $x^l < x(t_h)$, we must have $x^l \in [x(t_h) - \tau, x(t_h)]$, $0 \leq \tau \leq \delta$, and

$$I(\tau) = \int_{x^l}^{x(t_h)} Z(x, y(t_i) + \tau)dx + \int_{x(t_h)}^{x(t_i)-\tau} Z(x, y(t_i) + \tau)dx.$$

By (26),

$$\int_{x(t_h)}^{x(t_i)-\tau} Z(x, y(t_i) + \tau)dx \geq \tau \left(\int_{x(t_h)}^{x(t_i)-\tau} Z_p(x, y(t_i))dx - Z(x(t_i), y(t_i)) \right) - 2\epsilon\tau.$$

On the other hand, it follows from (24) that

$$\int_{x^l}^{x(t_h)} Z(x, y(t_i) + \tau)dx \geq (Z(x(t_h), y(t_h)) - \epsilon)\tau.$$

Thus

$$I(\tau) \geq \tau \left(\int_{x(t_h)}^{x(t_i)-\tau} Z_p(x, y(t_i))dx - Z(x(t_i), y(t_i)) + Z(x(t_h), y(t_h)) - 3\epsilon \right) > 0.$$

The last inequality is due to (18).

Now that we have established that $I(\tau)$ is positive, we are almost done. Let $\mathcal{L}$ be any horizontal line segment with a distance less than δ_0 above $\mathbf{s}$ and lying between the lines L_h and L_i, and with its right-hand end on the line L_i. The above discussion shows that, when $\delta_0 \leq \delta$, the line integral of Z along $\mathcal{L}$ is positive. Let $\tilde{\mathcal{R}}$ denote the region surrounded by $\tilde{\mathbf{s}}$, $\mathbf{s}$, L_h, and L_i. Since $\tilde{\mathbf{s}}$ is monotonic increasing and the maximum distance between $\tilde{\mathbf{s}}$ and $\mathbf{s}$ is not larger than δ_0, any intersection of a horizontal line and $\tilde{\mathcal{R}}$ will take the form $\mathcal{L}$. This implies that the area integral of Z over $\tilde{\mathcal{R}}$ is positive.

The argument for a vertical section is similar, but the integral $I(\tau)$ takes the form

$$I(\tau) = \int_{y(t_i)+\tau}^{y^u} Z(x(t_i) - \tau, y)dy.$$

This completes the proof. □

The conditions of Theorem 6 are stronger than the necessary conditions of Theorems 4 and 5, and it is worth discussing the differences. First observe that the necessary conditions of Theorem 4 (i) and Theorem 5 (i) carry over as we would expect after a change to strict inequalities for Z_p and Z_q.

When we come to consider conditions (ii) and (iii) of Theorems 4 and 5, the position is more complex. We replace the condition $w(t) \leq w(t_1)$ on a horizontal

section with the condition $w(t) < w(t_1)$ except at identified points among the t_j, $j = 1, \ldots,$ where $w(t_j) = w(t_1)$. The same thing happens on a vertical section. However, there is no direct equivalence of conditions (3) and (5) (though these inequalities can be derived from adding (10) and (11) in an appropriate way). Consider the inequality (10). As we have already observed, $Z(x(t_k), y(t_k)) \geq 0$, and $Z(x(t_j), y(t_j)) \leq 0$. So inequality (10) is strictly stronger than inequality (2). Similarly, inequality (11) is stronger than inequality (4).

Condition (iv) of Theorem 6 strengthens the inequality $Z(x(t_i), y(t_i)) \geq 0$ (or $Z(x(t_i), y(t_i)) \leq 0$), which can be derived from the necessary conditions at a turning point from horizontal to vertical (or vertical to horizontal).

5. An example. To illustrate the application of these necessary and sufficient conditions, we consider a small example based on one given in [1]. We define the market distribution function ψ via an intermediate function ϕ, which is defined as

$$\phi(q,p) = ((q-p)^2 - 1)((q-p)^2 - 0.7) - 1.59p^2 - 1.11q^2.$$

Then we set $\psi\,(q,p) = pq + 0.045\phi(q,p) - 0.1$. We suppose that the cost function is given by the quadratic $C(q) = 0.08q^2$. We also suppose that the generator has a two-way hedging contract for a quantity 0.15 and thus makes payments under these contracts of $0.15(f-p)$ where f is the contract price. Since f is fixed, we can ignore this term in seeking an optimal solution, and so we can take the profit function as $R(q,p) = (q - 0.15)p - C(q)$.

The first step in understanding the behavior of this example is to look at the values of the function Z over the region Ψ. This is shown in Figure 4, where the dashed lines show that $Z = 0$ and divide Ψ into regions where Z is either positive or negative. Also shown in the figure are three solid lines AB, CD, and EF, which connect the lower boundary of Ψ, $\psi = 0$, with its upper boundary, $\psi = 1$. These are candidate offer curves. AB runs along a $Z = 0$ line, and EF runs along another $Z = 0$ line for most of its length. It is clear that AB will satisfy all of the conditions used in this paper and is a local optimum, but the position is less clear for the other two curves.

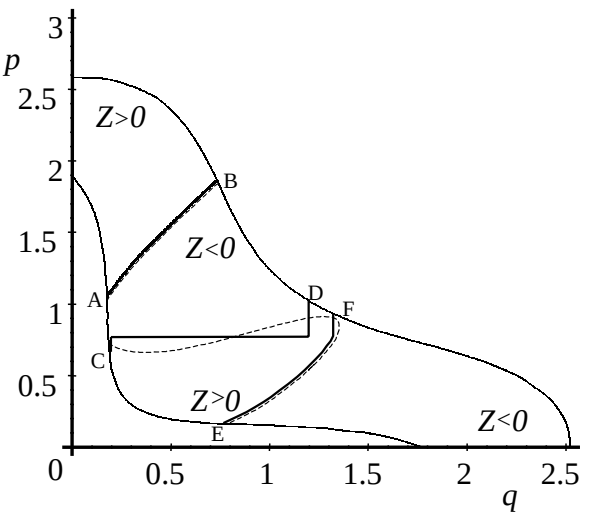


FIG. 4. *Candidate supply curves for the example.*

First we look at the CD offer curve. This starts with a vertical section from $(0.1858, 0.68244)$ to $(0.1858, 0.75685)$, then has a horizontal section to the point $(1.1972, 0.75685)$, and then finishes with a vertical section to hit the boundary of Ψ at the point $(1.1972, 1.0245)$. These points have been chosen so that the solution satisfies all of the conditions of Theorem 4. Each of the three sections has the property that the integral of Z along the section is zero, which is what is required for w to take the same value w_0 at the endpoints of each section. Moreover, the fact that the two vertical sections move from $Z > 0$ to $Z < 0$, while the horizontal section does the reverse, will ensure that w is no less than w_0 on the vertical sections and no more than w_0 on the horizontal section. Once it is decided to search for an offer curve of this general form, these conditions can be used to find the exact curve. Starting from different points on the $\psi = 0$ curve, we can let the w condition determine when to switch from vertical to horizontal and then back to vertical. We then iterate amongst possible starting positions to search for a solution which achieves a zero Z integral on the final vertical section; i.e., it makes $w = w_0$ at the point where the vertical section crosses the $\psi = 1$ curve. (All of the numerical calculations for this example were performed using Maple.)

The next step is to check the second order conditions of Theorem 5. We require that the integral of Z_q on the first vertical section be no greater than zero, and, in fact,

$$\int_{0.68244}^{0.75685} Z_q(0.1858, y)dy = -5.408 \times 10^{-3},$$

so this condition is satisfied. We also require that the integral of Z_p along the horizontal section be no less than zero, but

$$\int_{0.1858}^{1.1972} Z_p(x, 0.75685)dx = -0.16526,$$

so this condition fails. Moreover, the integral of Z_q on the last vertical section is greater than zero so that this condition fails as well. Finally, both the conditions (3) and (5) involving the value of Z at the corner points fail. So we know from the theorem that this solution is not a local optimum.

Next we consider the EF solution. The final vertical section of this is chosen in such a way that the integral of Z on this vertical section is zero. It starts at the point $(1.365, 0.82561)$ and moves vertically until meeting the $\psi = 1$ boundary at $(1.365, 0.90056)$ Since the rest of the curve is on the $Z = 0$ curve, the conditions of Anderson and Philpott will be satisfied. We check the conditions of Theorem 5. We have

$$\int_{0.82561}^{0.90056} Z_q(1.365, y)dy = -9.4287 \times 10^{-3} < 0$$

as required.

The next step is to check the sufficient conditions of Theorem 6. Most of the conditions of this theorem will hold trivially, but we need to check that

$$\int_{0.82561}^{0.90056} Z_q(1.365, y)dy < Z(1.365, 0.90056) - Z(1.365, 0.82561).$$

Now $Z(1.365, 0.82561) = 0$, and $Z(1.365, 0.90056) = -1.3136 \times 10^{-3}$, so this condition will hold.

Hence both the curves AB and EF are locally optimal: to choose between them, we must evaluate the objective function for each. In fact, the objective function value along the curve AB is 0.5183, while the value along EF is 0.4857. So the curve AB is the (global) optimum for this problem.

REFERENCES

[1] E. J. Anderson and A. B. Philpott, *Optimal offer construction in electricity markets*, Math. Oper. Res., 27 (2002), pp. 82–100.

[2] F. Bolle, *Supply function equilibria and the danger of tacit collusion: The case of spot markets for electricity*, Energy Economics, 14 (1992), pp. 94–102.

[3] H.-P. Chao and H. G. Huntington, *Designing Competitive Electricity Markets*, Kluwer Academic, Boston, 1998.

[4] N.-H. von der Fehr and D. Harbord, *Competition in Electricity Spot Markets, Economic Theory and International Experience*, Memorandum, Department of Economics, University of Oslo, Oslo, Norway, 1998.

[5] R. J. Green and D. M. Newbery, *Competition in the British electricity spot market*, J. Political Economy, 100 (1992), pp. 929–953.

[6] G. Gross and D. J. Finlay, *Optimal bidding strategies in competetive electricity markets*, in Proceedings of the 12th Power Systems Computation Conference, Dresden, Germany, August 1996, pp. 815–823.

[7] B. F. Hobbs, *Network models of spatial oligopoly with an application to deregulation of electricity generation*, Oper. Res., 34 (1986), pp. 395–409.

[8] A. Rudkevich, M. Duckworth, and R. Rosen, *Modelling electricity pricing in a deregulated generation industry: The potential for oligopoly pricing in a poolco*, The Energy Journal, 19 (1998), pp. 19–48.

[9] J.-Y. Wei and Y. Smeers, *Spatial oligopolistic electricity models with Cournot generators and regulated transmission prices*, Oper. Res., 47 (1999), pp. 102–112.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 4, pp. 1229–1253

AN OPTIMAL CONTROL PROBLEM GOVERNED BY QUASI-LINEAR VARIATIONAL INEQUALITIES*

HONGWEI LOU†

Abstract. An optimal control problem governed by quasi-linear variational inequality is considered. The cost functional to be minimized contains the solution of a quasi-linear variational inequality. To get the existence, regularity, and necessary condition for the optimal pair, a new related control problem is introduced. By proving the existence of an optimal pair to such a new problem, the existence and regularity of the optimal pair to the original problem are obtained. It turns out that the regularity obtained is sharp in general. Some necessary conditions of the optimal pair are also obtained.

Key words. optimal control, quasi-linear, variational inequality, existence, regularity, necessary condition

AMS subject classifications. 35J70, 49J20

PII. S0363012900375032

1. Introduction.

In this paper, we consider the following optimal control problem.

Problem (C). Find a $\bar{y} \in W_0^{1,p}(\Omega)$ such that

$$I(\bar{y}) = \inf_{y \in W_0^{1,p}(\Omega)} I(y), \tag{1.1}$$

where

$$I(y) = I(y; z, p, \Omega) \stackrel{\triangle}{=} \frac{1}{p} \int_\Omega \{|T(y) - z|^p + |\nabla y|^p\} dx, \quad y \in W_0^{1,p}(\Omega), \tag{1.2}$$

$1 < p < +\infty$, $\Omega \subset \mathbb{R}^n$ is a bounded domain with $C^{1,1}$ boundary $\partial\Omega$, $z \in L^p(\Omega)$ is a given target profile, y is a function in Sobolev space $W_0^{1,p}(\Omega)$, ∇y denotes its generalized gradient, and $\psi \equiv T(y)$ is the state corresponding to the control y satisfying the following quasi-linear variational inequality:

$$\begin{cases} \psi \in \mathbb{K}(y) \stackrel{\triangle}{=} \{\varphi \in W_0^{1,p}(\Omega) | \varphi \geq y, \text{ a.e. } \Omega\}, \\ \displaystyle\int_\Omega |\nabla\psi|^{p-2} \nabla\psi \cdot \nabla(\varphi - \psi) dx \geq 0 \quad \forall\, \varphi \in \mathbb{K}(y). \end{cases} \tag{1.3}$$

It is well known that for any $y \in W_0^{1,p}(\Omega)$, (1.3) admits a unique solution (see [5], [24], [25], for example) and ψ is the solution of (1.3) if and only if ψ minimizes the functional $\varphi \mapsto \int_\Omega |\nabla\varphi|^p dx$ over $\mathbb{K}(y)$. Moreover, replacing φ by $\psi + v$ in (1.3), we see that ψ satisfies

$$-\mathrm{div}(|\nabla\psi|^{p-2}\nabla\psi) \geq 0 \quad \text{in } \Omega \tag{1.4}$$

in the weak sense. We denote $\mathcal{H}_+^p(\Omega)$ to be the set of all $\psi \in W_0^{1,p}(\Omega)$ satisfying (1.4). Any $\psi \in \mathcal{H}_+^p(\Omega)$ is called a p-superharmonic function.

*Received by the editors June 23, 2000; accepted for publication (in revised form) April 1, 2002; published electronically October 29, 2002. This work was supported in part by the Science Foundation of Education Ministry of China.

http://www.siam.org/journals/sicon/41-4/37503.html

†Mathematical Department, Fudan University, Shanghai 200433, China (hw-lou@sohu.com).

If one wants to design a membrane having an expected shape, one needs to choose a suitable obstacle. In this case, the obstacle can be looked at as a control, and the membrane can be looked at as the state. Then our aim is to find an optimal obstacle control minimizing some cost functional.

In the literature, many other authors have discussed similar problems concerning different aspects. See [3], [13], [14], [15], [16], [17], [18], [23], [28], for example. They considered optimal control problems for obstacle problems (or variational inequalities). Usually, the obstacle functions are fixed at 0, and the control variables appear in the variational inequality. In other words, controls do not change the obstacle. They change only the functional of obstacle problems. For general cases of the obstacle problem, when the obstacles are smooth enough, they can be reduced to the case of the obstacle being 0 by simple translations (see [30, Chap. 1]). Similarly, by suitable translations, we can reformulate most optimal control problems such that the controls change only the obstacles but not the functionals of the obstacle problems. In case an optimal control does not exist (or we do not know whether it exists or not), we can consider the problem with controls being in a larger space (see [4], for example) as we consider relaxed controls in existence theory of optimal control problems. In these cases, reformulating the obstacle problems and considering obstacles as controls (or depending on controls) will be more convenient.

In mathematical finance, the problem of American option pricing is an obstacle problem (see [20], [21], for example). Researching in optimal obstacle control is also useful in the theory of designing and pricing American-type contingent claims.

In case $p = 2$, the optimal obstacle control problem was studied by Adams, Lenhart, and Yong [1], Chen [8], [9], [10], and Lou [26]. The cases of $p \neq 2$ are related to the so-called non-Newtonian fluids. The quantity p is a characteristic of the medium. Media with $p > 2$ are called dilatant fluids and those with $p < 2$ are called pseudoplastics. For $p = 2$, they are Newtonian fluids (see [11]).

The main purpose of this paper is to establish the existence, regularity, and necessary condition of an optimal pair to Problem (C).

By the fact that an optimal control, if it exists, must be equal to the corresponding optimal state, we can see that finding an optimal pair $(\bar{y}, \bar{\psi})$ to Problem (C) is equivalent to finding a minimizer $\bar{y}$ such that (see the next section for details)

$$\tilde{I}(\bar{y}) = \inf_{y \in \mathcal{H}_+^p(\Omega)} \tilde{I}(y),$$

with

$$\tilde{I}(y) = \frac{1}{p} \int_\Omega \{|y - z|^p + |\nabla y|^p\} dx. \tag{1.5}$$

Thus, it is not very hard to prove that Problem (C) admits a unique optimal pair when $p = 2$. The cases of $p \neq 2$ are quite different. The difficulty is that when $p \neq 2$ and $n \neq 1$, $\mathcal{H}_+^p(\Omega)$ is not a convex set (and we do not know whether it is weakly closed in $W_0^{1,p}(\Omega)$). Thus the standard method of getting the existence of an optimal control is not valid. The main idea of getting the existence and regularity of an optimal pair is to establish the existence theorem for a new related control problem. More precisely, we introduce the following optimal control problem.

Problem (C*). Find a $\bar{u} \in L_+^{p'}(\Omega) \stackrel{\triangle}{=} \{v \in L^{p'}(\Omega) | v \geq 0, \text{ a.e. } \Omega\}$ such that

$$I^*(\bar{u}) = \inf_{u \in L_+^{p'}(\Omega)} I^*(u),$$

where $p' = \frac{p}{p-1}$,

$$I^*(u) = I^*(u; z, p, \Omega) \triangleq \frac{1}{p}\int_\Omega \{|T^*(u) - z|^p + uT^*(u)\}dx, \quad u \in L^{p'}(\Omega), \tag{1.6}$$

with $y \equiv T^*(u) \in W_0^{1,p}(\Omega)$ being the unique solution of the following equation:

$$\begin{cases} -\mathrm{div}(|\nabla y|^{p-2}\nabla y) = u \ \text{ in } \Omega, \\ y|_{\partial\Omega} = 0. \end{cases} \tag{1.7}$$

We will show that $\bar{u}$ is an optimal control to Problem (C*) if and only if $\bar{y} \equiv T^*(\bar{u})$ is an optimal control to Problem (C). Because of this fact, we are able to get $C^{1,\alpha}$-regularity of $\bar{y}$ by the results of [12]. Furthermore, it turns out that such a regularity is the best possible result in general.

The difficulty in getting the necessary condition of optimality $\bar{y}$ is to give a characterization of the singular set $\{x \in \Omega | \nabla\bar{y}(x) = 0\}$ of $\bar{y}$. Without a proper characterization to the singular set, the necessary condition of the optimal control is far from completely determining it. For some special cases, we do get characterizations of optimal controls.

Now we state our main results.

THEOREM 1.1. *Let $1 < p < +\infty$, $z \in L^p(\Omega)$. Then Problem* (C) *admits an optimal control.*

Theorem 1.1 is an existence theorem. The following theorem gives the regularity of an optimal control.

THEOREM 1.2. *Suppose $1 < p < +\infty$, $z \in L^p(\Omega)$, and $z^+ \in L^q(\Omega)$ for some $q \geq p$. Let $\bar{y}$ be an optimal control to Problem* (C). *Then there exists a $\bar{u} \in L_+^{\frac{q}{p-1}}(\Omega)$ such that*

$$\begin{cases} -\mathrm{div}(|\nabla\bar{y}|^{p-2}\nabla\bar{y}) = \bar{u} \ \ \textit{in } \Omega, \\ \bar{y}|_{\partial\Omega} = 0. \end{cases} \tag{1.8}$$

Consequently, if $q > pn$, then $\bar{y} \in C^{1,\alpha}(\bar{\Omega})$ for some $\alpha \in (0,1)$.

In this paper, for $1 \leq q \leq +\infty$, we denote

$$L_+^q(\Omega) \triangleq \{v \in L^q(\Omega) | v \geq 0, \text{ a.e. } \Omega\}. \tag{1.9}$$

For $1 < p < +\infty$, $\varepsilon \geq 0$, $y \in W_0^{1,p}(\Omega)$, define operator $\mathbf{L}$ by

$$\begin{aligned} \mathbf{L}(\varphi; y, p, \varepsilon) \triangleq & -\mathrm{div}[(\varepsilon^2 + |\nabla y|^2)^{\frac{p-2}{2}}\nabla\varphi] \\ & -(p-2)\mathrm{div}[(\varepsilon^2 + |\nabla y|^2)^{\frac{p-4}{2}}(\nabla y \cdot \nabla\varphi)\nabla y] \end{aligned} \tag{1.10}$$

and denote

$$\begin{aligned} \mathbf{L}(\varphi; y, p) \triangleq & \mathbf{L}(\varphi; y, p, 0) \equiv -\mathrm{div}(|\nabla y|^{p-2}\nabla\varphi) \\ & -(p-2)\mathrm{div}[|\nabla y|^{p-4}(\nabla y \cdot \nabla\varphi)\nabla y]. \end{aligned} \tag{1.11}$$

For $\varphi \in W_0^{1,p}(\Omega)$ and $\mu \in W^{-1,p'}(\Omega)$, we always write $\langle \mu, \varphi \rangle$ in the integral form $\int_\Omega \mu\varphi \, dx$.

The positive part and the negative part of a function f will be denoted by f^+, f^-, respectively, i.e., $f^+ = \max(f, 0)$, $f^- = \max(-f, 0)$.

When the domain Ω is clear from the context, the sets $\{x \in \Omega | f(x) < 0\}$ and $\{x \in \Omega | f(x) = 0\}$ will be denoted by $\{f < 0\}$ and $\{f = 0\}$, respectively. The characteristic function of E will be denoted by χ_E. Hereafter, by a solution of a differential equation we mean a weak solution.

The rest of the paper is organized as follows. In section 2, we will transform the original problem to a new related problem. In section 3, we will introduce an approximate problem and give estimates of optimal pairs for the approximate problem. In section 4, we use the results obtained in section 3 to obtain the existence and regularity of the solution to the original problem. An example is presented to show that such regularity is the best possible in general. Finally, in section 5, necessary conditions of optimality in some special cases are derived.

2. Transformation of the problem. As is in the case of $p = 2$ (see [1]), we will prove that if $\bar{y}$ minimizes $I(\cdot)$, then $T(\bar{y})$ must be equal to $\bar{y}$. To see this, let us introduce the following lemma, which will reveal some basic properties of the operator T.

LEMMA 2.1. *Given $1 < p < +\infty$.*

(i) *Suppose $y \in W_0^{1,p}(\Omega)$. Then $T(y) = y$ if and only if $y \in \mathcal{H}_+^p(\Omega)$.*

(ii) $T^2(y) = T(y)$ $\forall y \in W_0^{1,p}(\Omega)$.

(iii) $T[W_0^{1,p}(\Omega)] = \mathcal{H}_+^p(\Omega)$.

Proof. (i) Let $y \in W_0^{1,p}(\Omega)$ satisfy $T(y) = y$. Then by the definition of T, we have $y = T(y) \in \mathcal{H}_+^p(\Omega)$.

On the other hand, let $y \in \mathcal{H}_+^p(\Omega)$. Noting that for any $v \in \mathbb{K}(y)$, we have $v - y \in W_0^{1,p}(\Omega)$ and $v - y \geq 0$, a.e. Ω, thus, it follows that

$$\int_\Omega |\nabla y|^{p-2} \nabla y \cdot \nabla (v - y) dx \geq 0 \quad \forall v \in \mathbb{K}(y).$$

Consequently, $T(y) = y$ by the definition.

(ii) Let $y \in W_0^{1,p}(\Omega)$. We have $T(y) \in \mathcal{H}_+^p(\Omega)$. Consequently, $T^2(y) = T(y)$ by (i).

(iii) First, $T[W_0^{1,p}(\Omega)] \subseteq \mathcal{H}_+^p(\Omega)$. Next, by (i), we have

$$\mathcal{H}_+^p(\Omega) = T[\mathcal{H}_+^p(\Omega)] \subseteq T[W_0^{1,p}(\Omega)],$$

proving the result. □

Now we can establish the following proposition, which shows that an optimal control to Problem (C) must be equal to the corresponding optimal state.

PROPOSITION 2.2. *Given $1 < p < +\infty$, $z \in L^p(\Omega)$, suppose $\bar{y}$ is an optimal control to Problem* (C)*. Then $T(\bar{y}) = \bar{y}$. Consequently, $\bar{y}$ is an optimal control to Problem* (C) *if and only if $\bar{y}$ minimizes $\tilde{I}(\cdot)$ over $\mathcal{H}_+^p(\Omega)$ (see* (1.5) *for the definition of $\tilde{I}(\cdot)$).*

Proof. By the definition of $\bar{y}$, we have

$$\int_\Omega \{|T(\bar{y}) - z|^p + |\nabla \bar{y}|^p\} dx \leq \int_\Omega \{|T(y) - z|^p + |\nabla y|^p\} dx \quad \forall\, y \in W_0^{1,p}(\Omega).$$

Let $y = T(\bar{y})$. By Lemma 2.1, $T(y) = T(\bar{y})$. Therefore

$$\int_\Omega \{|T(\bar{y}) - z|^p + |\nabla \bar{y}|^p\} dx \leq \int_\Omega \{|T(\bar{y}) - z|^p + |\nabla [T(\bar{y})]|^p\} dx.$$

Thus

$$\int_\Omega |\nabla \bar{y}|^p dx \le \int_\Omega |\nabla (T(\bar{y}))|^p dx \le \int_\Omega |\nabla \varphi|^p dx \quad \forall\ \varphi \in \mathbb{K}(\bar{y}).$$

That is, $T(\bar{y}) = \bar{y}$. □

From (1.2), (1.5), and Lemma 2.1, we see that

$$\tilde{I}(y) = I(y) \quad \forall\ y \in \mathcal{H}_+^p(\Omega).$$

Because of Proposition 2.2, we need only consider $\tilde{I}(\cdot)$ over $\mathcal{H}_+^p(\Omega)$. When $p = 2$, $\mathcal{H}_+^p(\Omega)$ is a closed and convex set. Then by the results in [22, Chap. 1], we can prove that the minimizer uniquely exists. But when $p \neq 2$ and $n \neq 1$, $\mathcal{H}_+^p(\Omega)$ is not convex. Here is a counterexample for Ω being a ball. The general cases are similar.

Example 1. Let $n = 2$, $1 < p < +\infty$, $p \neq 2$, and $\Omega = B_4$ be the ball of radius 4 in $\mathbb{R}^2$, centered at the origin. Define

$$u_1(x_1, x_2) = \begin{cases} -x_1^2 + x_2^2 + 16x_1 + 81 & \text{if } p > 2 \\ x_1^2 - x_2^2 - 16x_1 + 64 & \text{if } p < 2 \end{cases} \quad \text{in } \bar{\Omega},$$

$$u_2(x_1, x_2) = \begin{cases} -16x_1 + 65 & \text{if } p > 2 \\ 16x_1 + 65 & \text{if } p < 2 \end{cases} \quad \text{in } \bar{\Omega},$$

and

$$\varphi(x_1, x_2) = 6(16 - x_1^2 - x_2^2) \ \text{ in } \bar{\Omega}.$$

By a straightforward computation, we have

$$-\mathrm{div}(|\nabla u_i|^{p-2}\nabla u_i) \ge 0 \quad \text{in } \Omega, \quad i = 1, 2,$$

$$-\mathrm{div}(|\nabla \varphi|^{p-2}\nabla \varphi) \ge 0 \quad \text{in } \Omega.$$

Let $\omega_i = \min(u_i, \varphi)$. Then $\omega_i \in W_0^{1,\infty}(\Omega) \subset W_0^{1,p}(\Omega)$ and

$$\omega_i = \begin{cases} u_i & \text{in } \Omega_1^i, \\ \varphi & \text{in } \Omega_2^i, \end{cases}$$

where Ω_1^i is an ellipse in Ω and $\Omega_2^i = \Omega \setminus \Omega_1^i$.

Let ν_i be the outer normal of $\partial\Omega_1^i$. We have

$$\nu_i = \frac{-\nabla\varphi + \nabla u_i}{|-\nabla\varphi + \nabla u_i|} \quad \text{on } \partial\Omega_1^i, \quad i = 1, 2.$$

Therefore (see Lemma 3.3),

$$(-|\nabla\varphi|^{p-2}\nabla\varphi + |\nabla u_i|^{p-2}\nabla u_i) \cdot \nu_i \ge 0 \quad \text{on } \partial\Omega_1^i, \quad i = 1, 2.$$

Thus, for any $v \in W_0^{1,p'}(\Omega)$ satisfying $v \ge 0$, a.e. in Ω,

$$\begin{aligned} &\int_\Omega |\nabla\omega_i|^{p-2}\nabla\omega_i \cdot \nabla v dx \\ &= \int_{\Omega_1^i} [-\mathrm{div}(|\nabla u_i|^{p-2}\nabla u_i)] v dx + \int_{\Omega_2^i} [-\mathrm{div}(|\nabla\varphi|^{p-2}\nabla\varphi)] v dx \\ &\quad + \int_{\partial\Omega_1^i} v(-|\nabla\varphi|^{p-2}\nabla\varphi + |\nabla u_i|^{p-2}\nabla u_i) \cdot \nu_i ds \ge 0. \end{aligned}$$

Therefore $\omega_i \in \mathcal{H}_+^p(\Omega)$, $i = 1, 2$.

Since

$$\omega_1+\omega_2=\begin{cases} x_2^2-x_1^2+146 & \text{if } p>2 \\ x_1^2-x_2^2+129 & \text{if } p<2 \end{cases} \quad \text{in } B_{\frac{1}{2}}(0)$$

in this neighborhood, therefore $\frac{1}{2}(\omega_1+\omega_2)\notin\mathcal{H}_+^p(\Omega)$. Thus, $\mathcal{H}_+^p(\Omega)$ is not convex.

Let us rewrite $\tilde{I}(\cdot)$ as follows:

$$\begin{aligned}\tilde{I}(y)&=\frac{1}{p}\int_\Omega\{|y-z|^p+|\nabla y|^p\}dx \\ &=\frac{1}{p}\int_\Omega\{|y-z|^p+[-\mathrm{div}(|\nabla y|^{p-2}\nabla y)]\cdot y\}dx.\end{aligned} \tag{2.1}$$

Next, we introduce Problem (C*) stated in the introduction. It is clear that

$$T^*[L_+^{p'}(\Omega)]\subset\mathcal{H}_+^p(\Omega). \tag{2.2}$$

By (2.1) and (1.7), we have

$$\tilde{I}[T^*(u)]=I^*(u)\quad\forall\ u\in L^{p'}(\Omega). \tag{2.3}$$

Then the existence of an optimal control to Problem (C*) means not only the existence of an optimal control to Problem (C) but also that the optimal control is in the set

$$\{y\in W_0^{1,p}(\Omega)|\ -\mathrm{div}(|\nabla y|^{p-2}\nabla y)\in L_+^{p'}(\Omega)\}.$$

Thus, its regularity is better than $W^{1,p}$-regularity. For example, when $p=2$, it means that an optimal control to Problem (C) belongs to $H_0^1(\Omega)\cap H^2(\Omega)$. To reveal the relation between Problems (C) and (C*), we give the following proposition.

PROPOSITION 2.3. *Given $1<p<+\infty$, $z\in L^p(\Omega)$, suppose $\bar{y}$ is an optimal control to Problem* (C) *and $\bar{u}\equiv-\mathrm{div}(|\nabla\bar{y}|^{p-2}\nabla\bar{y})\in L^{p'}(\Omega)$. Then $\bar{u}\in L_+^{p'}(\Omega)$ and $\bar{u}$ is an optimal control to Problem* (C*).

On the other hand, suppose $\bar{u}$ is an optimal control to Problem (C*). *Then $T^*(\bar{u})$ is an optimal control to Problem* (C).

Proof. Suppose $\bar{y}$ is an optimal control to Problem (C) and $\bar{u}=-\mathrm{div}(|\nabla\bar{y}|^{p-2}\nabla\bar{y})\in L^{p'}(\Omega)$. Then $\bar{u}\in L_+^{p'}(\Omega)$ by $\bar{y}\in\mathcal{H}_+^p(\Omega)$. Now, for any $u\in L_+^{p'}(\Omega)$, we have $T^*(u)\in\mathcal{H}_+^p(\Omega)$. Hence

$$\begin{aligned}&I^*(\bar{u})=I(\bar{y})\le I(T^*(u))=\frac{1}{p}\int_\Omega\{|T(T^*(u))-z|^p+|\nabla[T^*(u)]|^p\}dx \\ &=\frac{1}{p}\int_\Omega\{|T^*(u)-z|^p+uT^*(u)\}dx=I^*(u)\quad\forall\ u\in L_+^{p'}(\Omega).\end{aligned}$$

Thus $\bar{u}$ is an optimal control to Problem (C*).

Similarly, let $\bar{u}$ be an optimal control to Problem (C*). Then for any $u\in L_+^{p'}(\Omega)$,

$$\tilde{I}(T^*(\bar{u}))=I^*(\bar{u})\le I^*(u)=\tilde{I}(T^*(u)).$$

This means $T^*(\bar{u})$ minimizes $\tilde{I}(\cdot)$ over $\{T^*(u)|u\in L_+^{p'}(\Omega)\}$. By density (see Corollary 3.5 for details), $T^*(\bar{u})$ minimizes $\tilde{I}(\cdot)$ over $\mathcal{H}_+^p(\Omega)$. Hence $T^*(\bar{u})$ is an optimal control to Problem (C). ∎

3. Approximate problem. To get the existence, regularity, and necessary condition of an optimal control, we establish the following theorem for an approximate problem. Let us denote $\mathcal{U}_{M,\delta} \equiv \{v : \Omega \to [\delta, M] | v \text{ measurable } \}$ for $0 \le \delta < M < +\infty$.

THEOREM 3.1. *Let $z \in L^\infty(\Omega)$, $1 < p < +\infty$, $0 < \delta < 1$, and $p\|z\|^{p-1}_{L^\infty(\Omega)} + 2p \le M < +\infty$. Then there exists at least one $\bar{u} \in \mathcal{U}_{M,\delta}$ such that*

$$I^*(\bar{u}) = \inf_{u \in \mathcal{U}_{M,\delta}} I^*(u). \tag{3.1}$$

Moreover, if $\bar{u} \in \mathcal{U}_{M,\delta}$ satisfies (3.1), *then*

$$|\bar{u}(x)| \le |z^+(x)|^{p-1} + \delta, \quad \text{a.e. } \Omega. \tag{3.2}$$

It is crucial that in (3.2), the estimate of $\bar{u}$ is independent of $\delta > 0$ (we can replace δ by 1 in (3.2)) and $M \gg 1$. By this fact, we finally get the existence and regularity of an optimal pair for Problem (C).

To prove Theorem 3.1, we need some preliminary lemmas.

LEMMA 3.2. *Let C be a constant. If $\varphi \in W^{m,p}(\Omega)$, $p \ge 1, m \ge 1$, then*

$$\partial^\beta \varphi(x) = 0, \quad \text{a.e. in } \{\varphi = C\} \quad \forall\ 1 \le |\beta| \le m,$$

where $\beta = (\beta_1, \dots, \beta_n)$ is an n-tuple of nonnegative integers β_i, $|\beta| = \sum_{i=1}^n \beta_i$.

The above lemma can be found in ([26, p. 69]; see also [29]). It tells us that for any element z in $W^{m,1}(\Omega)$, on its level set, we can calculate its mth generalized derivatives just as we calculate classical derivatives.

LEMMA 3.3. *Let $1 < p < +\infty$. Then*

(i) *for any $a, b \in \mathbb{R}^m, m \in \mathcal{N}$,*

$$(|a|^{p-2}a - |b|^{p-2}b) \cdot (a - b) \ge 0, \tag{3.3}$$

and the equality holds if and only if $a = b$;

(ii) *for any $\varepsilon > 0$, $a \in \mathbb{R}^m, m \in \mathcal{N}$,*

$$(\varepsilon^2 + |a|^2)^{\frac{p-2}{2}} |a|^2 \ge |a|^p - \varepsilon^p. \tag{3.4}$$

Proof. (i) We omit the proof since it is straightforward.

(ii) If $p \ge 2$ or $|a| \le \varepsilon$, then (3.4) holds obviously. Now, we suppose that $1 < p < 2$ and $|a| > \varepsilon$. Then

$$\begin{aligned}(\varepsilon^2 + |a|^2)^{p/2}|a|^2 + (\varepsilon^2 + |a|^2)\varepsilon^p &\ge |a|^{p+2} + |a|^2\varepsilon^p \\ &\ge |a|^{p+2} + |a|^p\varepsilon^2 = |a|^p(\varepsilon^2 + |a|^2).\end{aligned}$$

That is (3.4). □

Let $\varepsilon \ge 0$. For $\mu \in W^{-1,p'}(\Omega)$, define $y_\varepsilon \equiv T^*_\varepsilon(\mu)$ to be the unique solution of the following quasi-linear elliptic equation:

$$\begin{cases} -\mathrm{div}[(\varepsilon^2 + |\nabla y_\varepsilon|^2)^{\frac{p-2}{2}} \nabla y_\varepsilon] = \mu & \text{in } \Omega, \\ y_\varepsilon|_{\partial\Omega} = 0. \end{cases} \tag{3.5}$$

The following two lemmas give some basic properties of the operator T^*_ε (recall that $T^*_0 = T^*$).

LEMMA 3.4. *Let $\varepsilon \geq 0$, $p \in (1, +\infty)$, $\mu \in W^{-1,p'}(\Omega)$.*

(i) *Equation (3.5) admits a unique solution $T^*_\varepsilon(\mu)$ in $W^{1,p}_0(\Omega)$. Moreover, there exists a positive constant $C = C(p, \Omega)$, independent of $\varepsilon \geq 0$, such that*

$$\|T^*_\varepsilon(\mu)\|_{W^{1,p}_0(\Omega)} \leq C \left(\|\mu\|^{\frac{1}{p-1}}_{W^{-1,p'}(\Omega)} + \varepsilon \right). \tag{3.6}$$

(ii) *Suppose (as $\varepsilon \to 0^+$)*

$$\mu_\varepsilon \to \mu \quad \textit{strongly in } W^{-1,p'}(\Omega).$$

Then

$$T^*_\varepsilon(\mu_\varepsilon) \to T^*(\mu) \quad \textit{strongly in } W^{1,p}_0(\Omega). \tag{3.7}$$

In particular, if $\mu_\varepsilon \to \mu$ weakly in $L^{p'}(\Omega)$, then (3.7) holds.

(iii) *Fix $\varepsilon \geq 0$. Suppose (as $k \to +\infty$)*

$$\mu^k \to \mu \quad \textit{strongly in } W^{-1,p'}(\Omega) \textit{ or } \textit{ weakly in } L^{p'}(\Omega).$$

Then

$$T^*_\varepsilon(\mu^k) \to T^*_\varepsilon(\mu) \quad \textit{strongly in } W^{1,p}_0(\Omega). \tag{3.8}$$

Proof. (i) The results can be obtained easily since y_ε satisfies (3.5) if and only if y_ε minimizes the functional $\varphi \mapsto \int_\Omega \{\frac{1}{p}(\varepsilon^2 + |\nabla\varphi|^2)^{\frac{p}{2}} - \varphi\mu\} dx$ over $W^{1,p}_0(\Omega)$.

(ii) Denote $y_\varepsilon = T^*_\varepsilon(\mu_\varepsilon)$, $y = T^*(\mu)$. By (3.6), we know that for $\varepsilon \in (0, 1)$, y_ε is bounded uniformly in $W^{1,p}_0(\Omega)$. Consequently, by the Banach–Alaoglu theorem (see [31]) and the Sobolev imbedding theorem (see [2]), y_ε has a subsequence which converges weakly in $W^{1,p}_0(\Omega)$ and strongly in $L^p(\Omega)$. Thus, it is sufficient to prove that $y^* = y$ and

$$\lim_{\varepsilon \to 0^+} \|\nabla y_\varepsilon\|_{L^p(\Omega)} = \|\nabla y\|_{L^p(\Omega)} \tag{3.9}$$

if

$$y_\varepsilon \to y^* \quad \text{weakly in } W^{1,p}_0(\Omega), \quad \text{strongly in } L^p(\Omega).$$

To see this, noting that

$$\int_\Omega \left\{ \frac{1}{p}(\varepsilon^2 + |\nabla y_\varepsilon|^2)^{\frac{p}{2}} - y_\varepsilon \mu_\varepsilon \right\} dx \leq \int_\Omega \left\{ \frac{1}{p}(\varepsilon^2 + |\nabla\varphi|^2)^{\frac{p}{2}} - \varphi\mu_\varepsilon \right\} dx \quad \forall\, \varphi \in W^{1,p}_0(\Omega),$$

$$\int_\Omega |\nabla y^*|^p dx \leq \liminf_{\varepsilon \to 0+} \int_\Omega (\varepsilon^2 + |\nabla y_\varepsilon|^2)^{\frac{p}{2}} dx, \tag{3.10}$$

and

$$\lim_{\varepsilon \to 0+} \int_\Omega y_\varepsilon \mu_\varepsilon dx = \int_\Omega y^* \mu dx, \tag{3.11}$$

we have

$$\int_\Omega \left\{ \frac{1}{p}|\nabla y^*|^p - y^*\mu \right\} dx \leq \int_\Omega \left\{ \frac{1}{p}|\nabla\varphi|^p - \varphi\mu \right\} dx \quad \forall \varphi \in W^{1,p}_0(\Omega).$$

Hence $y^* = y$.

On the other hand, by Lemma 3.3(ii),

$$\int_\Omega y_\varepsilon \mu_\varepsilon dx = \int_\Omega (\varepsilon^2 + |\nabla y_\varepsilon|^2)^{\frac{p-2}{2}} |\nabla y_\varepsilon|^2 dx \geq \int_\Omega (|\nabla y_\varepsilon|^p - \varepsilon^p) dx.$$

Therefore

$$\limsup_{\varepsilon \to 0^+} \int_\Omega |\nabla y_\varepsilon|^p dx \leq \int_\Omega y\mu dx = \int_\Omega |\nabla y|^p dx.$$

Consequently,

$$\lim_{\varepsilon \to 0^+} \int_\Omega |\nabla y_\varepsilon|^p dx = \int_\Omega |\nabla y|^p dx.$$

Thus, (3.9) holds and we get the proof.

(iii). The proof is similar to that of (ii). □

COROLLARY 3.5. *Let $p \in (1, +\infty)$. Then the set $\{T^*(u) | u \in L^{p'}_+(\Omega)\}$ is dense in $\mathcal{H}^p_+(\Omega)$.*

Proof. It is easy to see that

$$\mathcal{H}^p_+(\Omega) = \{T^*(\mu) | \mu \in W^{-1,p'}(\Omega), \mu \geq 0 \text{ in the weak sense}\}.$$

Thus, we need only to prove that for any $\mu \in W^{-1,p'}(\Omega)$, $\mu \geq 0$, there exists a sequence $u_j \in L^{p'}_+(\Omega)$ such that $T^*(u_j) \to T^*(\mu)$ strongly in $W^{1,p}_0(\Omega)$.

To see this, let

$$\Psi(x) = \begin{cases} k \ \exp\left(-\dfrac{1}{1-|x|^2}\right) & \text{if } |x| < 1, \\ 0 & \text{if } |x| \geq 1, \end{cases}$$

where $k > 0$ is chosen to satisfy the condition

$$\int_{\mathbb{R}^n} \Psi(x) dx = 1.$$

For $j \in \mathcal{N}$, denote $\Psi_j(x) \equiv j^{\,n} \Psi(jx)$, and $v_j \equiv \mu * \Psi_j$. Then we can verify that $v_j \in C^\infty(\bar{\Omega}) \bigcap L^{p'}_+(\Omega)$ and

$$v_j \to \mu \ \text{ weakly in } W^{-1,p'}(\Omega).$$

By Mazur's theorem, there exists $\alpha_{j,i} \geq 0$, $\sum_{i=1}^{K_j} \alpha_{j,i} = 1$ such that

$$u_j \equiv \sum_{i=1}^{K_j} \alpha_{j,i} v_{j+i} \to \mu \ \text{ strongly in } W^{-1,p'}(\Omega).$$

We have $u_j \in L^p_+(\Omega)$. By Lemma 3.4(iii), we get

$$T^*(u_j) \to T^*(\mu) \quad \text{strongly in } W^{1,p}_0(\Omega),$$

completing the proof. □

LEMMA 3.6. *Let $\varepsilon \geq 0$, $p \in (1, +\infty)$.*

(i) *Suppose $y_\varepsilon \in W_0^{1,p}(\Omega)$, $-\mathrm{div}[(\varepsilon^2 + |\nabla y_\varepsilon|^2)^{\frac{p-2}{2}} \nabla y_\varepsilon] \geq 0$. Then $y_\varepsilon \geq 0$, a.e. Ω.*

(ii) *Suppose $q > p'n$, $u \in L^q(\Omega)$. Let $y_\varepsilon = T_\varepsilon^*(u)$. Then $y_\varepsilon \in C^{1,\alpha}(\bar{\Omega})$, where $\alpha = \alpha(p, q, \|u\|_{L^q(\Omega)}, \Omega) \in (0, 1)$ is independent of $\varepsilon \geq 0$. Moreover, there exists a positive constant $C = C(p, q, \|u\|_{L^q(\Omega)}, \Omega)$ such that*

$$|D_i y_\varepsilon(x) - D_i y_\varepsilon(x')| \leq C|x - x'|^\alpha \quad \forall\ x, x' \in \bar{\Omega}; \qquad i = 1, 2, \ldots, n. \tag{3.12}$$

Proof. (i) Since $y_\varepsilon^- \in W_0^{1,p}(\Omega)$ and $y_\varepsilon^- \geq 0$, a.e. Ω, we have

$$\begin{aligned} 0 &\leq \int_\Omega (\varepsilon^2 + |\nabla y_\varepsilon|^2)^{\frac{p-2}{2}} \nabla y_\varepsilon \cdot \nabla y_\varepsilon^- dx \\ &= -\int_\Omega (\varepsilon^2 + |\nabla y_\varepsilon|^2)^{\frac{p-2}{2}} |\nabla y_\varepsilon^-|^2 dx. \end{aligned}$$

Therefore $\nabla y_\varepsilon^- = 0$, a.e. Ω. Consequently, $y_\varepsilon^- = 0$, a.e. Ω, i.e., $y_\varepsilon \geq 0$, a.e. Ω.

(ii) The result is an immediate corollary of the interior $C^{1,\alpha}$-regularity of the quasi-linear equations with the homogenous boundary condition (see [12] and [33]).

LEMMA 3.7. *Suppose $0 < \alpha < 1$, $f \in C^\alpha(\bar{\Omega}, \mathbb{R}^n)$. Then*

(i) *$\forall \beta > -1$, there exists a $\gamma \in (0, 1)$ such that $|f|^\beta f \in C^\gamma(\bar{\Omega}, \mathbb{R}^n)$;*

(ii) *$\forall \delta > 0$, $\beta > -1$, there exists a $\gamma \in (0, 1)$ and $C > 0$, independent of $\varepsilon \in [0, 1]$, such that*

$$|h_\varepsilon(\tilde{x}) - h_\varepsilon(x)| \leq C|\tilde{x} - x|^\gamma \quad \forall\ \tilde{x},\ x \in \bar{\Omega},$$

where $h_\varepsilon = (\varepsilon^2 + |f|^\delta)^{\frac{\beta}{\delta}} f$.

(iii) *$\forall k \geq 0$, $2s + k + 1 > 0$, there exists a $\gamma \in (0, 1)$ and $C > 0$, independent of $\varepsilon \in [0, 1]$, such that*

$$|g_\varepsilon(\tilde{x}) - g_\varepsilon(x)| \leq C|\tilde{x} - x|^\gamma \quad \forall\ \tilde{x},\ x \in \bar{\Omega},$$

where $g_\varepsilon = (\varepsilon^2 + |f|^2)^s |f|^k f$.

Proof. The proofs of (i) and (ii) are straightforward, and (i) is in fact a special case of (ii). Finally, (iii) follows from (i) and (ii) since $(\varepsilon^2 + |f|^2)^s |f|^k f = (\varepsilon^2 + |g|^{\frac{2}{k+1}})^s g$ with $g = |f|^k f$. □

The following lemma has a result similar to that of Lemma 3.2, which shows some crucial information about the so-called singular set.

LEMMA 3.8. *Suppose $u \in L^\infty(\Omega)$, and $y \in W_0^{1,p}(\Omega)$ is a solution of the following equation:*

$$\begin{cases} -\mathrm{div}(|\nabla y|^{p-2} \nabla y) = u & \text{in } \Omega, \\ y|_{\partial\Omega} = 0. \end{cases}$$

Then

$$u(x) = 0, \quad \text{a.e. } x \in \{\nabla y = 0\}.$$

For a proof of the above lemma, see [27]. As a consequence of the above lemma, the singular set $\{\nabla y = 0\}$ must have Lebesgue measure zero if $u(x) > 0$, a.e. Ω.

Now, let us give a proof of Theorem 3.1.

Proof of Theorem 3.1. First, by (2.3), for any $u \in \mathcal{U}_{M,\delta}$, $I^*(u) \geq 0$. Thus, we have $u_k \in \mathcal{U}_{M,\delta}$ satisfying

$$\lim_{k\to+\infty} I^*(u_k) = \inf_{u\in\mathcal{U}_{M,\delta}} I^*(u).$$

Moreover, we can suppose that

$$u_k \to \bar{u} \quad \text{weakly in } L^{p'}(\Omega).$$

Since $\mathcal{U}_{M,\delta}$ is convex, $\bar{u} \in \mathcal{U}_{M,\delta}$. On the other hand, by Lemma 3.4(ii) (for the case $\varepsilon = 0$), we can obtain that

$$T^*(u_k) \to \bar{y} \equiv T^*(\bar{u}) \quad \text{strongly in } W_0^{1,p}(\Omega).$$

Thus, by the definition of I^*, we have

$$I^*(\bar{u}) = \lim_{k\to+\infty} I^*(u_k) = \inf_{u\in\mathcal{U}_{M,\delta}} I^*(u).$$

Therefore, there exists at least one $\bar{u} \in \mathcal{U}_{M,\delta}$ satisfying (3.1).

Now, let $\bar{u} \in \mathcal{U}_{M,\delta}$ be an optimal control satisfying (3.1), $\bar{y} = T^*(\bar{u})$. By Lemma 3.6(ii), we have $\bar{y} \in C^{1,\alpha}(\bar{\Omega})$ for some $\alpha \in (0,1)$. We will prove (3.2) in three steps.

Step I: *First approximation.* Let $(\bar{y}, \bar{u})$ be described as above. Fix

$$\eta \in \left(0, \frac{1}{1 + \|\bar{y}\|^{p-1}_{L^{+\infty}(\Omega)}}\right).$$

For $\theta > 0$, consider the equation

$$\begin{cases} -\mathrm{div}(|\nabla y_\theta|^{p-2}\nabla y_\theta) + \theta y_\theta = u \ \text{ in } \Omega, \\ y_\theta|_{\partial\Omega} = 0. \end{cases} \tag{3.13}$$

Similar to the existence of a $\bar{u} \in \mathcal{U}_{M,\delta}$ satisfying (3.1), it is easy to prove that there exists a $(\bar{y}_\theta, \bar{u}_\theta) \in W_0^{1,p}(\Omega) \times \mathcal{U}_{M,\delta}$ satisfying (3.13) and

$$\begin{aligned} &\frac{1}{p}\int_\Omega \{|\bar{y}_\theta - z|^p + \eta|\bar{y}_\theta - \bar{y}|^p + \bar{u}_\theta\bar{y}_\theta\}dx \\ &\leq \frac{1}{p}\int_\Omega \{|y_\theta - z|^p + \eta|y_\theta - \bar{y}|^p + uy_\theta\}dx \end{aligned} \tag{3.14}$$

for all $(y_\theta, u) \in W_0^{1,p}(\Omega) \times \mathcal{U}_{M,\delta}$ satisfying (3.13).

For any $k > 0$, by (3.13), we have

$$\begin{aligned} \theta k \int_\Omega (\bar{y}_\theta - k)^+ dx &\leq \theta \int_\Omega \bar{y}_\theta(\bar{y}_\theta - k)^+ dx \\ &\leq \int_\Omega \{|\nabla[(\bar{y}_\theta - k)^+]|^p + \theta\bar{y}_\theta(\bar{y}_\theta - k)^+\}dx \\ &= \int_\Omega \bar{u}_\theta(\bar{y}_\theta - k)^+ dx \leq M\int_\Omega (\bar{y}_\theta - k)^+ dx. \end{aligned}$$

Therefore,

$$\theta\bar{y}_\theta \leq M, \quad \text{a.e. } \Omega. \tag{3.15}$$

Similarly,

$$\theta\bar{y}_\theta \geq 0, \quad \text{a.e. } \Omega. \tag{3.16}$$

Thus, $\|\bar{u}_\theta - \theta\bar{y}_\theta\|_{L^\infty(\Omega)} \leq M$. By Lemma 3.6, we see that $\bar{y}_\theta$ is uniformly bounded in $C^{1,\alpha}(\bar{\Omega})$ for some $\alpha \in (0,1)$. Then we can suppose that, at least in the sense of a subsequence (as $\theta \to 0^+$),

$$\begin{cases} \bar{u}_\theta \to \hat{u} & \text{weakly in } \ L^q(\Omega) \quad \forall\ 1 < q < +\infty, \\ \bar{y}_\theta \to \hat{y} & \text{uniformly in } \ C^1(\bar{\Omega}). \end{cases}$$

Thus, $\hat{u} \in \mathcal{U}_{M,\delta}$. By Lemma 3.4(ii), $\hat{y} = T^*(\hat{u})$. Moreover, by (3.14),

$$\begin{aligned} &\frac{1}{p}\int_\Omega \{|\hat{y} - z|^p + \eta|\hat{y} - \bar{y}|^p + \hat{u}\hat{y}\}dx \\ \leq\ &\frac{1}{p}\int_\Omega \{|y - z|^p + \eta|y - \bar{y}|^p + uy\}dx \quad \forall\ u \in \mathcal{U}_{M,\delta}, \end{aligned}$$

where $y = T^*(u)$. If $\hat{y} \not\equiv \bar{y}$ (i.e., $\hat{u} \not\equiv \bar{u}$), replacing u by $\bar{u}$ in the above inequality, we get

$$\frac{1}{p}\int_\Omega \{|\hat{y} - z|^p + \hat{u}\hat{y}\}dx < \frac{1}{p}\int_\Omega \{|\bar{y} - z|^p + \bar{u}\bar{y}\}dx;$$

this contradicts the optimality of $\bar{u}$ (see (3.1)). Therefore, we have $\hat{u} \equiv \bar{u}$, $\hat{y} \equiv \bar{y}$. Consequently, not only in the sense of a subsequence, we get (as $\theta \to 0^+$),

$$\begin{cases} \bar{u}_\theta \to \bar{u} & \text{weakly in } \ L^q(\Omega) \quad \forall\ 1 < q < +\infty, \\ \bar{y}_\theta \to \bar{y} & \text{uniformly in } \ C^1(\bar{\Omega}). \end{cases} \tag{3.17}$$

Consequently, there exists a $\theta_0 > 0$ such that

$$\theta\|\bar{y}_\theta\|_{L^\infty(\Omega)} < \frac{\delta}{2}, \quad \|\bar{y}_\theta\|^{p-1}_{L^\infty(\Omega)} < \|\bar{y}\|^{p-1}_{L^\infty(\Omega)} + 1 \quad \forall\ 0 < \theta \leq \theta_0. \tag{3.18}$$

In this case, we have (see (3.13))

$$-\mathrm{div}(|\nabla\bar{y}_\theta|^{p-2}\nabla\bar{y}_\theta) = \bar{u}_\theta - \theta\bar{y}_\theta \geq \frac{\delta}{2} > 0 \quad \text{in } \Omega.$$

Thus, by Lemma 3.8, the set $E_\theta \equiv \{\nabla\bar{y}_\theta = 0\}$ has n-dimensional Lebesgue measure zero. On the other hand, E_θ is closed since $\bar{y}_\theta \in C^{1,\alpha}(\bar{\Omega})$.

Step II: *Second approximation.* For $0 < \theta < \theta_0$, $\varepsilon > 0$, consider the equation

$$\begin{cases} -\mathrm{div}[(\varepsilon^2 + |\nabla y_{\theta,\varepsilon}|^2)^{\frac{p-2}{2}}\nabla y_{\theta,\varepsilon}] + \theta y_{\theta,\varepsilon} = u \ \text{ in } \Omega, \\ y_{\theta,\varepsilon}|_{\partial\Omega} = 0. \end{cases} \tag{3.19}$$

Similar to Step I, there exists a $(\bar{y}_{\theta,\varepsilon}, \bar{u}_{\theta,\varepsilon}) \in W_0^{1,p}(\Omega) \times \mathcal{U}_{M,\delta}$ satisfying (3.19) and

$$\begin{aligned} &\frac{1}{p}\int_\Omega \{|\bar{y}_{\theta,\varepsilon} - z|^p + \eta|\bar{y}_{\theta,\varepsilon} - \bar{y}|^p + \eta|\bar{y}_{\theta,\varepsilon} - \bar{y}_\theta|^p + \bar{u}_{\theta,\varepsilon}\bar{y}_{\theta,\varepsilon}\}dx \\ \leq\ &\frac{1}{p}\int_\Omega \{|y_{\theta,\varepsilon} - z|^p + \eta|y_{\theta,\varepsilon} - \bar{y}|^p + \eta|y_{\theta,\varepsilon} - \bar{y}_\theta|^p + uy_{\theta,\varepsilon}\}dx \end{aligned}$$

for any $(y_{\theta,\varepsilon}, u) \in W_0^{1,p}(\Omega) \times \mathcal{U}_{M,\delta}$ satisfying (3.19). Moreover, there exists a $\bar{\varphi}_{\theta,\varepsilon} \in W_0^{1,2}(\Omega)$ such that (see (1.10) for the definition of $\mathbf{L}$)

$$\begin{cases} \mathbf{L}(\bar{\varphi}_{\theta,\varepsilon}; \bar{y}_{\theta,\varepsilon}, p, \varepsilon) + \theta\bar{\varphi}_{\theta,\varepsilon} = f_{\theta,\varepsilon} - \bar{u}_{\theta,\varepsilon} \ \text{ in } \Omega, \\ \bar{\varphi}_{\theta,\varepsilon}|_{\partial\Omega} = 0, \end{cases} \tag{3.20}$$

and

$$\int_\Omega (\bar{\varphi}_{\theta,\varepsilon} - \bar{y}_{\theta,\varepsilon})(u - \bar{u}_{\theta,\varepsilon})dx \le 0 \quad \forall\ u \in \mathcal{U}_{M,\delta}, \tag{3.21}$$

where

$$\begin{aligned} f_{\theta,\varepsilon} = {} & p|z - \bar{y}_{\theta,\varepsilon}|^{p-2}(z - \bar{y}_{\theta,\varepsilon}) + p\eta|\bar{y} - \bar{y}_{\theta,\varepsilon}|^{p-2}(\bar{y} - \bar{y}_{\theta,\varepsilon}) \\ & + p\eta|\bar{y}_\theta - \bar{y}_{\theta,\varepsilon}|^{p-2}(\bar{y}_\theta - \bar{y}_{\theta,\varepsilon}) \quad \text{ in } \ \Omega. \end{aligned} \tag{3.22}$$

Similarly to (3.15)–(3.16) and (3.17), we have $0 \le \theta\bar{y}_{\theta,\varepsilon} \le M$, and (as $\varepsilon \to 0^+$)

$$\begin{cases} \bar{u}_{\theta,\varepsilon} \to \bar{u}_\theta \quad \text{weakly in } L^q(\Omega) \quad \forall\ 1 < q < +\infty, \\ \bar{y}_{\theta,\varepsilon} \to \bar{y}_\theta \quad \text{uniformly in } C^1(\bar{\Omega}). \end{cases} \tag{3.23}$$

Therefore, there exists a $C_\theta > 0$, independent of $\varepsilon \in (0,1)$, such that (see (3.18) and (3.22))

$$\begin{aligned} & -C_\theta \le f_{\theta,\varepsilon} \le p(z^+)^{p-1} + p\eta(\bar{y})^{p-1} + p\eta(\bar{y}_\theta)^{p-1} \\ & \le p\|z\|^{p-1}_{L^\infty(\Omega)} + p\eta(2\|\bar{y}\|^{p-1}_{L^\infty(\Omega)} + 1) \le p\|z\|^{p-1}_{L^\infty(\Omega)} + 2p \le M \quad \text{ in } \Omega. \end{aligned} \tag{3.24}$$

Thus, as in (3.15), we have

$$-\frac{1}{\theta}C_\theta \le \varphi_{\theta,\varepsilon} \le \frac{1}{\theta}M.$$

By the $W^{2,q}$-estimate for linear elliptic equation (3.20), we obtain the following estimate:

$$\|\bar{\varphi}_{\theta,\varepsilon}\|_{W^{2,q}(\Omega_0)} \le C(\theta, q, \Omega_0) \ \ \forall\ \Omega_0 \subset\subset (\Omega \setminus E_\theta), \quad 1 < q < +\infty,$$

where the constant $C(\theta, q, \Omega_0)$ is independent of $\varepsilon \in (0,1)$ (see [19]). Then we can suppose that (at least for a subsequence)

$$\begin{aligned} \bar{\varphi}_{\theta,\varepsilon} \to \bar{\varphi}_\theta \quad & \text{weakly in } L^q(\Omega) \quad \forall\ 1 < q < +\infty, \\ & \text{uniformly in } C^1(\bar{\Omega}_0) \ \ \forall\ \Omega_0 \subset\subset (\Omega \setminus E_\theta). \end{aligned} \tag{3.25}$$

Passing to the limit and noting that (3.21) still holds by replacing the integral domain Ω by $\Omega_0 \subset \Omega$ (just let $u = \bar{u}_{\theta,\varepsilon}$ on the set $\Omega \setminus \Omega_0$), we get from (3.20)–(3.22) that

$$\mathbf{L}(\bar{\varphi}_\theta; \bar{y}_\theta, p) + \theta\bar{\varphi}_\theta = f_\theta - \bar{u}_\theta \ \text{ in } \Omega \setminus E_\theta, \tag{3.26}$$

and

$$\int_{\Omega_0} (\bar{\varphi}_\theta - \bar{y}_\theta)(u - \bar{u}_\theta)dx \le 0 \ \ \forall\ u \in \mathcal{U}_{M,\delta},\ \Omega_0 \subset\subset (\Omega \setminus E_\theta), \tag{3.27}$$

where

$$f_\theta = p|z - \bar{y}_\theta|^{p-2}(z - \bar{y}_\theta) + p\eta|\bar{y} - \bar{y}_\theta|^{p-2}(\bar{y} - \bar{y}_\theta) \quad \text{ in } \ \Omega.$$

One can easily see that (3.27) still holds if the integral domain Ω_0 is replaced by Ω, since E_θ has Lebesgue measure zero.

Step III: *Applying the maximum principles for approximate problems.* First, we want to prove that $\bar{\varphi}_\theta(x) \le \bar{y}_\theta(x)$, and then we want to get a useful estimate of $\bar{u}_\theta$. Equations (3.20)–(3.21) and (3.26)–(3.27) are maximum principles for approximate problems. We notice that in (3.26)–(3.27), no boundary condition is posed for $\bar{\varphi}_\theta$. Thus, as a maximum principle, (3.26)–(3.27) is incomplete (see similar results in [6], [7]).

By (3.21),

$$\begin{cases} \bar{u}_{\theta,\varepsilon}(x) = M, & \text{a.e. } x \in \{\bar{\varphi}_{\theta,\varepsilon} > \bar{y}_{\theta,\varepsilon}\}, \\ \bar{u}_{\theta,\varepsilon}(x) = \delta, & \text{a.e. } x \in \{\bar{\varphi}_{\theta,\varepsilon} < \bar{y}_{\theta,\varepsilon}\}. \end{cases}$$

Therefore, by (3.24),

$$f_{\theta,\varepsilon}(x) - \bar{u}_{\theta,\varepsilon}(x) \le 0, \quad \text{a.e. } x \in \{\bar{\varphi}_{\theta,\varepsilon} > \bar{y}_{\theta,\varepsilon}\}.$$

Thus, it follows from (3.20) that

$$\begin{aligned} \int_\Omega \Big\{ & (\varepsilon^2 + |\nabla \bar{y}_{\theta,\varepsilon}|^2)^{\frac{p-2}{2}} \nabla \bar{\varphi}_{\theta,\varepsilon} \cdot \nabla \left((\bar{\varphi}_{\theta,\varepsilon} - \bar{y}_{\theta,\varepsilon})^+ \right) \\ & + (p-2)(\varepsilon^2 + |\nabla \bar{y}_{\theta,\varepsilon}|^2)^{\frac{p-4}{2}} (\nabla \bar{y}_{\theta,\varepsilon} \cdot \nabla \bar{\varphi}_{\theta,\varepsilon}) \left[\nabla \bar{y}_{\theta,\varepsilon} \cdot \nabla \left((\bar{\varphi}_{\theta,\varepsilon} - \bar{y}_{\theta,\varepsilon})^+ \right) \right] \\ & + \theta \bar{\varphi}_{\theta,\varepsilon} (\bar{\varphi}_{\theta,\varepsilon} - \bar{y}_{\theta,\varepsilon})^+ \Big\} \le 0. \end{aligned} \tag{3.28}$$

On the other hand, by (3.18) and (3.23), we have $\varepsilon_0 > 0$ such that

$$\theta \|\bar{y}_{\theta,\varepsilon}\|_{L^\infty(\Omega)} < \delta \quad \forall\, \varepsilon \in (0, \varepsilon_0).$$

Consequently, by (3.19),

$$-\mathrm{div}[(\varepsilon^2 + |\nabla \bar{y}_{\theta,\varepsilon}|^2)^{\frac{p-2}{2}} \nabla \bar{y}_{\theta,\varepsilon}] = \bar{u}_{\theta,\varepsilon} - \theta \bar{y}_{\theta,\varepsilon} > 0 \quad \forall\, \varepsilon \in (0, \varepsilon_0).$$

Therefore

$$(p-1) \int_\Omega (\varepsilon^2 + |\nabla \bar{y}_{\theta,\varepsilon}|^2)^{\frac{p-2}{2}} \nabla \bar{y}_{\theta,\varepsilon} \cdot \nabla [(\bar{\varphi}_{\theta,\varepsilon} - \bar{y}_{\theta,\varepsilon})^+] dx \ge 0. \tag{3.29}$$

Noting that $\bar{y}_{\theta,\varepsilon} \ge 0$, we have

$$\theta \int_\Omega \bar{y}_{\theta,\varepsilon} (\bar{\varphi}_{\theta,\varepsilon} - \bar{y}_{\theta,\varepsilon})^+ dx \ge 0. \tag{3.30}$$

Combining (3.28)–(3.30), we get

$$\begin{aligned} & \int_\Omega (\varepsilon^2 + |\nabla \bar{y}_{\theta,\varepsilon}|^2)^{\frac{p-2}{2}} |\nabla [(\bar{\varphi}_{\theta,\varepsilon} - \bar{y}_{\theta,\varepsilon})^+]|^2 dx \\ & + (p-2) \int_\Omega (\varepsilon^2 + |\nabla \bar{y}_{\theta,\varepsilon}|^2)^{\frac{p-4}{2}} |\nabla \bar{y}_{\theta,\varepsilon} \cdot \nabla [(\bar{\varphi}_{\theta,\varepsilon} - \bar{y}_{\theta,\varepsilon})^+]|^2 dx \\ & + \int_\Omega \theta [(\bar{\varphi}_{\theta,\varepsilon} - \bar{y}_{\theta,\varepsilon})^+]^2 dx \\ & \le (p-2) \int_\Omega \varepsilon^2 (\varepsilon^2 + |\nabla \bar{y}_{\theta,\varepsilon}|^2)^{\frac{p-4}{2}} \nabla \bar{y}_{\theta,\varepsilon} \cdot \nabla [(\bar{\varphi}_{\theta,\varepsilon} - \bar{y}_{\theta,\varepsilon})^+] dx. \end{aligned} \tag{3.31}$$

Thus,

$$\min(p-1,1)\int_\Omega (\varepsilon^2+|\nabla\bar y_{\theta,\varepsilon}|^2)^{\frac{p-2}{2}}|\nabla[(\bar\varphi_{\theta,\varepsilon}-\bar y_{\theta,\varepsilon})^+]|^2dx \tag{3.32}$$
$$\le |p-2|\int_\Omega (\varepsilon^2+|\nabla\bar y_{\theta,\varepsilon}|^2)^{\frac{p-2}{2}}|\nabla\bar y_{\theta,\varepsilon}|\ |\nabla[(\bar\varphi_{\theta,\varepsilon}-\bar y_{\theta,\varepsilon})^+]|dx.$$

Consequently, by Hölder's inequality and (3.19),

$$\int_\Omega (\varepsilon^2+|\nabla\bar y_{\theta,\varepsilon}|^2)^{\frac{p-2}{2}}|\nabla[(\bar\varphi_{\theta,\varepsilon}-\bar y_{\theta,\varepsilon})^+]|^2dx \tag{3.33}$$
$$\le C_p\int_\Omega (\varepsilon^2+|\nabla\bar y_{\theta,\varepsilon}|^2)^{\frac{p}{2}}dx\le\tilde C_p$$

for some $C_p,\tilde C_p>0$ independent of $\varepsilon\in(0,\varepsilon_0)$. Therefore, we can suppose that

$$(\varepsilon^2+|\nabla\bar y_{\theta,\varepsilon}|^2)^{\frac{p-2}{4}}\nabla[(\bar\varphi_{\theta,\varepsilon}-\bar y_{\theta,\varepsilon})^+]\to g_\theta\quad\text{ weakly in } L^2(\Omega;\mathbb{R}^n).$$

On the other hand, since $\bar y_{\theta,\varepsilon}$ is bounded uniformly in $C^{1,\alpha}(\bar\Omega)$, by Lemma 3.7, both

$$(\varepsilon^2+|\nabla\bar y_{\theta,\varepsilon}|^2)^{\frac{p-2}{4}}\nabla\bar y_{\theta,\varepsilon}\quad\text{and}\quad(\varepsilon^2+|\nabla\bar y_{\theta,\varepsilon}|^2)^{\frac{p-6}{4}}|\nabla\bar y_{\theta,\varepsilon}|^2\nabla\bar y_{\theta,\varepsilon}$$

are bounded uniformly in $C^\beta(\bar\Omega;\mathbb{R}^n)$ for some $\beta\in(0,1)$ independent of ε. Thus, their difference

$$\varepsilon^2(\varepsilon^2+|\nabla\bar y_{\theta,\varepsilon}|^2)^{\frac{p-6}{4}}\nabla\bar y_{\theta,\varepsilon}$$

is also bounded uniformly in $C^\beta(\bar\Omega;\mathbb{R}^n)$. Then we can suppose that

$$\varepsilon^2(\varepsilon^2+|\nabla\bar y_{\theta,\varepsilon}|^2)^{\frac{p-6}{4}}\nabla\bar y_{\theta,\varepsilon}\to h_\theta\quad\text{ uniformly in } C(\bar\Omega;\mathbb{R}^n).$$

Noting that $\nabla\bar y_{\theta,\varepsilon}(x)\to\nabla\bar y_\theta(x)\ne0$ for $x\in(\Omega\setminus E_\theta)$, we get $h_\theta=0$ in $\Omega\setminus E_\theta$. Moreover, there exists an $\varepsilon_0>0$, and $C_1>C_2>0$, independent of $\varepsilon\in(0,\varepsilon_0)$, such that

$$C_2\le|\nabla\bar y_{\theta,\varepsilon}|\le C_2\quad\text{ in }\Omega_0\subset\subset(\Omega\setminus E_\theta)\quad\forall\ \varepsilon\in(0,\varepsilon_0).$$

Thus, by (3.33) and (3.31), there exists a $C>0$, independent of $\varepsilon\in(0,\varepsilon_0)$, such that

$$C\int_{\Omega_0}|\nabla[(\bar\varphi_{\theta,\varepsilon}-\bar y_{\theta,\varepsilon})^+]|^2dx\le\tilde C_p, \tag{3.34}$$

and

$$C\int_{\Omega_0}|\nabla[(\bar\varphi_{\theta,\varepsilon}-\bar y_{\theta,\varepsilon})^+]|^2dx \tag{3.35}$$
$$\le(p-2)\int_\Omega\varepsilon^2(\varepsilon^2+|\nabla\bar y_{\theta,\varepsilon}|^2)^{\frac{p-4}{2}}\nabla\bar y_{\theta,\varepsilon}\cdot\nabla[(\bar\varphi_{\theta,\varepsilon}-\bar y_{\theta,\varepsilon})^+]dx.$$

Combining (3.34) with (3.23) and (3.25), we have

$$(\bar\varphi_{\theta,\varepsilon}-\bar y_{\theta,\varepsilon})^+\to(\bar\varphi_\theta-\bar y_\theta)^+\quad\text{ weakly in }\ W^{1,2}(\Omega_0).$$

Then, by (3.35), we get

$$C\int_{\Omega_0}|\nabla[(\bar\varphi_\theta-\bar y_\theta)^+]|^2dx\le C\liminf_{\varepsilon\to0^+}\int_{\Omega_0}|\nabla[(\bar\varphi_{\theta,\varepsilon}-\bar y_{\theta,\varepsilon})^+]|^2dx$$
$$\le(p-2)\int_\Omega h_\theta\cdot g_\theta dx=0.$$

Thus,

$$\nabla[(\bar\varphi_\theta-\bar y_\theta)^+]=0,\quad \text{a.e. } \Omega\setminus E_\theta.$$

Therefore, for any ball $B\subset(\Omega\setminus E_\theta)$,

$$(\bar\varphi_\theta-\bar y_\theta)^+=C_B,\quad \text{a.e. } B.$$

We claim that $C_B=0$. Otherwise, suppose that $C_B>0$ for some $B\subset(\Omega\setminus E_\theta)$. Then

$$\bar\varphi_\theta=\bar y_\theta+C_B,\quad \text{a.e. } B.$$

Noting that $\nabla\bar y_\theta\neq0$ in $\Omega\setminus E_\theta$ and $\bar y_\theta\in C^{1,\alpha}(\bar\Omega)$, we can easily prove that $\bar y_\theta,\bar\varphi_\theta\in W^{2,q}(B)$ for some $q>1$ (see (3.13) and (3.26)). Then by Lemma 3.2 and (3.26), we have

$$-(p-1)\mathrm{div}(|\nabla\bar y_\theta|^{p-2}\nabla\bar y_\theta)+\theta\bar y_\theta+\theta C_B$$
$$=p|z-\bar y_\theta|^{p-2}(z-\bar y_\theta)+p\eta|\bar y-\bar y_\theta|^{p-2}(\bar y-\bar y_\theta)-\bar u_\theta\quad\text{in } B.$$

Therefore, by (3.13), and noting that $\bar y_\theta\ge0$,

$$\bar u_\theta=|z-\bar y_\theta|^{p-2}(z-\bar y_\theta)+\eta|\bar y-\bar y_\theta|^{p-2}(\bar y-\bar y_\theta)+\frac{p-2}{p}\theta\bar y_\theta-\frac1p\theta C_B \tag{3.36}$$
$$\le(z^+)^{p-1}+\eta(\bar y)^{p-1}+\theta\bar y_\theta\le(z^+)^{p-1}+2<M\quad\text{in } B.$$

On the other hand, by (3.27),

$$\begin{cases}\bar u_\theta=M, & \text{a.e. } \{\bar\varphi_\theta>\bar y_\theta\}\setminus E_\theta,\\ \bar u_\theta=\delta, & \text{a.e. } \{\bar\varphi_\theta<\bar y_\theta\}\setminus E_\theta.\end{cases} \tag{3.37}$$

This contradicts (3.36) since $B\subset(\{\bar\varphi_\theta>\bar y_\theta\}\setminus E_\theta)$. Thus, we prove that

$$(\bar\varphi_\theta-\bar y_\theta)^+=0,\quad \text{a.e. } \Omega\setminus E_\theta.$$

That is,

$$\bar\varphi_\theta\le\bar y_\theta,\quad \text{a.e. } \Omega\setminus E_\theta. \tag{3.38}$$

Similarly to (3.36), we have

$$\bar u_\theta\le(z^+)^{p-1}+\eta(\bar y)^{p-1}+\theta\bar y_\theta,\quad \text{a.e. } \{\bar\varphi_\theta=\bar y_\theta\}\setminus E_\theta.$$

Combining the above with (3.37)–(3.38) and noting that $|E_\theta|=0$, we get

$$\bar u_\theta\le(z^+)^{p-1}+\eta\bar y^{p-1}+\theta\bar y_\theta+\delta,\quad \text{a.e. } \Omega. \tag{3.39}$$

Let $\theta\to0^+$; by (3.17), we get

$$\bar u\le(z^+)^{p-1}+\eta\bar y^{p-1}+\delta,\quad \text{a.e. } \Omega.$$

Finally, let $\eta\to0^+$; then (3.2) follows and we get the proof. □

4. Existence and regularity of an optimal control. In this section, we will prove the main results of this paper.

Proof of Theorem 1.1. We first suppose that $z \in L^\infty(\Omega)$. Let $M > p\|z^+\|^{p-1}_{L^\infty(\Omega)} + 2p$, $0 < \delta < 1$. By Theorem 3.1, there exists a $\bar{u}^{M,\delta} \in \mathcal{U}_{M,\delta}$ such that (3.1) and (3.2) hold with $\bar{u}$ being replaced by $\bar{u}^{M,\delta}$. Thus, we have a subsequence $\delta_j \to 0^+$ such that

$$\bar{u}^{M,\delta_j} \to \bar{u}^M \quad \text{weakly in } L^q(\Omega) \quad \forall\ 1 < q < +\infty.$$

It is not very hard to see that $\bar{u}^M$ minimizes $I^*(\cdot)$ over $\mathcal{U}_{M,0}$ and by (3.2), $\bar{u}^M \le (z^+)^{p-1}$. Similarly, for any $S > M$, there exists a $\bar{u}^S$ minimizing $I^*(\cdot)$ over $\mathcal{U}_{S,0}$ and satisfying $\bar{u}^S \le (z^+)^{p-1}$. Consequently, $\bar{u}^S \in \mathcal{U}_{M,0}$. Therefore $I^*(\bar{u}^M) \le I^*(\bar{u}^S)$. That is, $\bar{u} \equiv \bar{u}^M$ also minimizes $I^*(\cdot)$ over $\mathcal{U}_{S,0}$ for any $S > M$. Then it must minimize $I^*(\cdot)$ over $L^{p'}_+(\Omega)$.

Now, suppose $z \in L^p(\Omega)$. For $\varepsilon > 0$, denote

$$z_\varepsilon(x) = \begin{cases} z(x) & \text{if } |z(x)| \le \frac{1}{\varepsilon}, \\ 0 & \text{if } |z(x)| > \frac{1}{\varepsilon}. \end{cases} \tag{4.1}$$

Then there exists a $\bar{u}_\varepsilon$ which minimizes $I^*(\cdot; z_\varepsilon, p, \Omega)$ over $L^{p'}_+(\Omega)$ such that $0 \le \bar{u}_\varepsilon \le (z^+_\varepsilon)^{p-1}$. Thus,

$$\|\bar{u}_\varepsilon\|_{L^{p'}(\Omega)} \le \|z^+\|^{p-1}_{L^p(\Omega)}. \tag{4.2}$$

Hence, we can suppose that

$$\bar{u}_\varepsilon \to \bar{u} \quad \text{weakly in } L^{p'}(\Omega). \tag{4.3}$$

Therefore,

$$0 \le \bar{u} \le (z^+)^{p-1}, \quad \text{a.e. } \Omega. \tag{4.4}$$

Consequently, $\bar{u} \in L^{p'}_+(\Omega)$. Denote $\bar{y}_\varepsilon = T^*(\bar{u}_\varepsilon), \bar{y} = T^*(\bar{u})$. By Lemma 3.4(iii),

$$\bar{y}_\varepsilon \to \bar{y} \quad \text{strongly in } W^{1,p}_0(\Omega). \tag{4.5}$$

Thus,

$$\begin{aligned} I^*(\bar{u}) &= \frac{1}{p}\int_\Omega \{|\bar{y} - z|^p + \bar{y}\bar{u}\}dx = \lim_{\varepsilon\to 0^+} \frac{1}{p}\int_\Omega \{|\bar{y}_\varepsilon - z_\varepsilon|^p + \bar{y}_\varepsilon\bar{u}_\varepsilon\}dx \\ &\le \lim_{\varepsilon\to 0^+} \frac{1}{p}\int_\Omega \{|T^*(u) - z_\varepsilon|^p + uT^*(u)\}dx \\ &= \frac{1}{p}\int_\Omega \{|T^*(u) - z|^p + uT^*(u)\}dx = I^*(u) \quad \forall\ u \in L^{p'}_+(\Omega). \end{aligned}$$

That is, $\bar{u}$ is an optimal control to Problem (C*). Consequently, $\bar{y}$ is an optimal control to Problem (C).

Proof of Theorem 1.2. If $\bar{y}$ is the limit of $\bar{y}_\varepsilon$ as in (4.5), then we have proved that $\bar{u} \equiv -\mathrm{div}(|\nabla\bar{y}|^{p-2}\nabla\bar{y})$ satisfies (4.4).

If $z^+ \in L^q(\Omega)$ and $q > pn$, then $\bar{u} \in L^{\frac{q}{p-1}}_+(\Omega)$ and $\frac{q}{p-1} > \frac{p}{p-1}n = p'n$. Consequently, by Lemma 3.6(ii), $\bar{y} \in C^{1,\alpha}(\bar{\Omega})$ for some $\alpha \in (0,1)$.

In general, let $\bar{y}$ be an optimal control to Problem (C). Then it is easy to see that $\bar{y}$ must be the unique minimizer of the functional $y \mapsto \tilde{I}(y) + \int_\Omega |y - \bar{y}|^p dx$ over $\mathcal{H}^p_+(\Omega)$. Replacing $I^*(\cdot)$ by $I^*(\cdot) + \int_\Omega |T^*(\cdot) - \bar{y}|^p dx$, we can get the desired regularity of $\bar{y}$ by a discussion similar to that above.

The $C^{1,\alpha}$-regularity of $\bar{y}$ is the best possible result in general. To see this, let us first establish the following theorem.

THEOREM 4.1. *Let $1 < p < +\infty$, $z \in L^p(\Omega)$. Let $\bar{y}$ be an optimal control to Problem* (C) *and let Z minimize $\tilde{I}(\cdot)$ over $W^{1,p}_0(\Omega)$, i.e.,*

$$\begin{cases} -\mathrm{div}(|\nabla Z|^{p-2}\nabla Z) = |z - Z|^{p-2}(z - Z) \quad \text{in } \Omega, \\ Z|_{\partial\Omega} = 0. \end{cases} \tag{4.6}$$

Then $\bar{y} = Z$ if and only if $Z \le z$.

Proof. Suppose $\bar{y} = Z$. Then $Z = \bar{y} \le z$ since

$$|z - \bar{y}|^{p-2}(z - \bar{y}) = -\mathrm{div}(|\nabla \bar{y}|^{p-2}\nabla \bar{y}) \ge 0.$$

Now, suppose $Z \le z$. Then $Z \in \mathcal{H}^p_+(\Omega)$. Since Z minimizes $\tilde{I}(\cdot)$ over $W^{1,p}_0(\Omega)$, it must minimize $\tilde{I}(\cdot)$ over $\mathcal{H}^p_+(\Omega)$. Thus, $\tilde{I}(Z) = \tilde{I}(\bar{y})$. Therefore $\bar{y}$ also minimizes $\tilde{I}$ over $W^{1,p}_0(\Omega)$. Since $\tilde{I}$ is strictly convex in $W^{1,p}_0(\Omega)$, we get $\bar{y} = Z$. □

By the previous theorem, we can see that the $C^{1,\alpha}$-regularity of $\bar{y}$ is the best possible result in general. In fact, it is well known that a p-harmonic function may have no $C^{1,\alpha}$-regularity provided $\alpha > \alpha_p$, where $\alpha_p \to \frac{1}{3}$ as $p \to +\infty$ (see [22]). The case in our problem is quite similar. When $p = 2$, examples in [26] show that the $C^{1,1}$-regularity of $\bar{y}$ is the best possible result in general. The following is an example for the case $p > 2$.

Example 2. Let $2 < p < +\infty$, $\Omega = B_1$, $z \equiv 1$. Then by Theorem 4.1, we can prove that the minimizer $\bar{y}$ of $I(\cdot)$ corresponding to z satisfies

$$\begin{cases} -\mathrm{div}(|\nabla \bar{y}|^{p-2}\nabla \bar{y}) = |z - \bar{y}|^{p-2}(z - \bar{y}) \quad \text{in } \Omega, \\ \bar{y}|_{\partial\Omega} = 0. \end{cases}$$

It is easy to see that $\bar{y}$ is a radial function. Denote $\bar{y}(x) = h(|x|)$. We have

$$-h'(r) = \left\{ r^{1-n} \int_0^r \xi^{n-1}[1 - h(\xi)]^{p-1} d\xi \right\}^{\frac{1}{p-1}}. \tag{4.7}$$

Therefore

$$-h'(r) \le \left[r^{1-n} \int_0^r \xi^{n-1} d\xi \right]^{\frac{1}{p-1}} = \left(\frac{r}{n} \right)^{\frac{1}{p-1}} \tag{4.8}$$

and

$$h(r) \le h(0) = \int_0^1 -h'(r) dr \le \int_0^1 \left(\frac{r}{n} \right)^{\frac{1}{p-1}} dr = \left(\frac{1}{n} \right)^{\frac{1}{p-1}} \cdot \frac{p-1}{p} < 1.$$

Thus, by (4.7),

$$-h'(r) \ge [1 - h(0)] \left(\frac{r}{n} \right)^{\frac{1}{p-1}}. \tag{4.9}$$

Combining (4.8) with (4.9), we see that $h \notin C^{1,\alpha}[0,1)$ $\forall \alpha \in (\frac{1}{p-1}, 1]$. Consequently, $\bar{y} \notin C^{1,\alpha}(B_1)$ when $\alpha \in (\frac{1}{p-1}, 1]$.

5. Necessary condition. In section 3, we obtained the existence theorem and the necessary condition for a minimizer of the approximate problem. In section 4, we obtained the regularity of an optimal control to the original problem. However, it is difficult to characterize an optimality $\bar{y}$. The difficulty comes when we attempt to characterize the singular set $\{\nabla \bar{y} = 0\}$. But, in some special cases, we do get a characterization of $\bar{y}$.

LEMMA 5.1. *Let $-\infty < a < b < +\infty$, $1 < p < +\infty$. Suppose that $y \in W_0^{1,p}(a,b)$. Then, in the weak sense,*

$$-(|y'|^{p-2}y')' \geq 0 \quad \text{in } (a,b) \Longleftrightarrow -y'' \geq 0 \quad \text{in } (a,b). \tag{5.1}$$

Consequently, $\mathcal{H}_+^p(a,b)$ is convex.

Proof. We first suppose that in the weak sense,

$$-(|y'|^{p-2}y')' \geq 0 \quad \text{in } (a,b).$$

Denote

$$\mu = -(|y'|^{p-2}y')'.$$

Then $\mu \in W^{-1,p'}(a,b)$ and $\mu \geq 0$. As in the proof of Corollary 3.5, we have $u_\varepsilon \in C^\infty[a,b] \bigcap L_+^{p'}(\Omega)$ for any $\varepsilon > 0$ such that

$$\|u_\varepsilon - \mu\|_{W^{-1,p'}(a,b)} \leq \varepsilon.$$

Let y_ε satisfy (recalling that $y \in W^{1,p}(a,b) \hookrightarrow C[a,b]$)

$$\begin{cases} -[(\varepsilon^2 + |y_\varepsilon'|^2)^{\frac{p-2}{2}} y_\varepsilon']' = u_\varepsilon & \text{in } (a,b), \\ y_\varepsilon(a) = 0,\ y_\varepsilon(b) = 0. \end{cases} \tag{5.2}$$

By Lemma 3.4(ii), we have

$$y_\varepsilon \to y \quad \text{strongly in } W_0^{1,p}(a,b).$$

It is easy to see that $y_\varepsilon \in C^\infty(a,b)$. Therefore,

$$-(\varepsilon^2 + |y_\varepsilon'|^2)^{\frac{p-4}{2}}(\varepsilon^2 + (p-1)|y_\varepsilon'|^2)y_\varepsilon'' = u_\varepsilon \geq 0 \quad \text{in } (a,b).$$

Thus,

$$-y_\varepsilon'' \geq 0 \quad \text{in } (a,b).$$

Passing to the limit, we get

$$-y'' \geq 0 \quad \text{in } (a,b).$$

The remainder can be proved by a discussion similar to the above. □

THEOREM 5.2. *Let $-\infty < a < b < +\infty$, $\Omega = (a,b)$, $z \in L^p(a,b)$. Then Problem* (C) *admits a unique optimal control $\bar{y}$ and there exist a $\bar{\varphi} \in W_0^{1,p'}(a,b) \cap W^{2,p'}(a,b)$ such that*

$$\bar{\varphi} \leq 0 \quad \text{in } (a,b), \tag{5.3}$$

$$\begin{cases} -(|\bar{y}'|^{p-2}\bar{y}')' = |z - \bar{y}|^{p-2}(z - \bar{y})\chi_{\{\bar{\varphi}=0\}} & \text{in } (a,b), \\ \bar{y}(a) = \bar{y}(b) = 0, \end{cases} \tag{5.4}$$

and

$$\begin{cases} -\bar{\varphi}'' = |z-\bar{y}|^{p-2}(z-\bar{y})\chi_{\{\bar{\varphi}<0\}} \quad \text{in } (a,b), \\ \bar{\varphi}(a) = \bar{\varphi}(b) = 0. \end{cases} \tag{5.5}$$

Moreover, the pair $(\bar{y},\bar{\varphi}) \in \mathcal{H}_+^p(a,b)\times [W_0^{1,p'}(a,b)\cap W^{2,p'}(a,b)]$ *satisfying* (5.3)–(5.5) *is unique.*

On the other hand, if $z \in L^\infty(a,b)$, *then* $\bar{\varphi} \in C^{1,1}[a,b]$.

If $p \in (1,2]$ *and* $z^+ \in L^q(a,b)$ *for some* $p \le q \le +\infty$, *then* $\bar{y} \in W^{2,\frac{q}{p-1}}(a,b)$.

Proof. Obviously, $\mathcal{H}_+^p(a,b)$ is closed in $W_0^{1,p}(a,b)$ and $\tilde{I}(\cdot)$ is a strictly convex functional. By Lemma 5.1, $\mathcal{H}_+^p(a,b)$ is convex. Thus, we can prove that Problem (C) admits a unique optimal control $\bar{y}$ by the discussion in section 2. By Theorem 1.2, $\bar{u} \equiv -(|\bar{y}'|^{p-2}\bar{y}')' \in L_+^{p'}(a,b)$.

Let $y \in \mathcal{H}_+^p(a,b)$. Then

$$\bar{y} + \alpha(y-\bar{y}) \in \mathcal{H}_+^p(a,b) \quad \forall\ \alpha \in (0,1).$$

By the optimality of $\bar{y}$, we have

$$0 \le \frac{1}{p\alpha}\int_a^b \{|\bar{y}+\alpha(y-\bar{y})-z|^p + |\nabla\bar{y}+\alpha(\nabla y-\nabla\bar{y})|^p - |\bar{y}-z|^p - |\nabla\bar{y}|^p\}dx.$$

Let $\alpha \to 0^+$; we then get

$$\begin{aligned} 0 &\le \int_a^b \{|\bar{y}-z|^{p-2}(\bar{y}-z)(y-\bar{y}) + |\nabla\bar{y}|^{p-2}\nabla\bar{y}\cdot(\nabla y-\nabla\bar{y})\}dx \\ &= \int_a^b \{|\bar{y}-z|^{p-2}(\bar{y}-z)+\bar{u}\}(y-\bar{y})dx. \end{aligned} \tag{5.6}$$

Let $\bar{\varphi}$ be the solution of the following equation:

$$\begin{cases} -\bar{\varphi}'' = |z-\bar{y}|^{p-2}(z-\bar{y}) - \bar{u} \ \text{ in } (a,b), \\ \bar{\varphi}(a) = \bar{\varphi}(b) = 0. \end{cases} \tag{5.7}$$

Then (5.6) becomes

$$0 \le \int_a^b \bar{\varphi}''(y-\bar{y})dx = -\int_a^b \bar{\varphi}[(-y'')-(-\bar{y}'')]dx. \tag{5.8}$$

Consequently, by choosing $-y'' = -\bar{y}'' + v$, $v \in L_+^\infty(a,b)$, we get (5.3).

On the other hand, choosing $y = \frac{1}{2}\bar{y}$ in (5.8), we have

$$0 \le \frac{1}{2}\int_a^b \bar{\varphi}(-\bar{y}'')dx.$$

By (5.7), we have $\bar{\varphi} \in C^1[a,b]$. Thus $\{\bar{\varphi}<0\}$ is open and we have $\{\bar{\varphi}<0\} = \cup_k(a_k,b_k)$, where (a_k,b_k) are mutually disjoint. Combining (5.3) with $-\bar{y}'' \ge 0$, we get

$$\text{supp } \bar{y}'' \subseteq \{\bar{\varphi}=0\}.$$

Then, for any k, there exists a constant C_k such that

$$\bar{y}' = C_k, \quad \text{a.e. } (a_k,b_k).$$

That is,

$$|\bar{y}'|^{p-2}\bar{y}' = |C_k|^{p-2}C_k, \quad \text{a.e. } (a_k, b_k),$$

and we get

$$\bar{u} = 0, \quad \text{a.e. } (a_k, b_k).$$

On the other hand, by (5.7) and Lemma 3.2, we have

$$\bar{u} = |z - \bar{y}|^{p-2}(z - \bar{y}), \quad \text{a.e. } \{\bar{\varphi} = 0\}.$$

Thus, we get (5.4)–(5.5).

When $z \in L^\infty(a, b)$, we see that $\bar{\varphi} \in W^{2,\infty}(a, b) = C^{1,1}[a, b]$ by (5.5).

When $p \in (1, 2]$ and $z^+ \in L^q(a, b)$ for some $p \le q \le +\infty$, we have

$$-\bar{y}'' = \frac{1}{p-1}|\bar{y}'|^{2-p}|z - \bar{y}|^{p-2}(z - \bar{y})\chi_{\{\bar{\varphi}=0\}} \in L^{\frac{q}{p-1}}(a, b)$$

by (5.4). Consequently, $\bar{y} \in W^{2,\frac{q}{p-1}}(a, b)$.

Now, we prove the uniqueness. Suppose

$$(\tilde{y}, \tilde{\varphi}) \in \mathcal{H}_+^p(\Omega) \times W_0^{1,p'}(a, b)$$

satisfies (5.3)–(5.5) too. Then

$$\begin{aligned}
(5.9) \qquad & \int_a^b \{[|z - \tilde{y}|^{p-2}(z - \tilde{y}) - |z - \bar{y}|^{p-2}(z - \bar{y})][(z - \tilde{y}) - (z - \bar{y})] \\
& \qquad + (|\bar{y}'|^{p-2}\bar{y}' - |\tilde{y}'|^{p-2}\tilde{y}')(\bar{y}' - \tilde{y}')\}dx \\
& = \int_a^b \{-(|\bar{y}'|^{p-2}\bar{y}')' - |z - \bar{y}|^{p-2}(z - \bar{y}) \\
& \qquad + (|\tilde{y}'|^{p-2}\tilde{y}')' + |z - \tilde{y}|^{p-2}(z - \tilde{y})\}(\bar{y} - \tilde{y})dx \\
& = \int_a^b (\bar{\varphi}'' - \tilde{\varphi}'')(\bar{y} - \tilde{y})dx.
\end{aligned}$$

Since $\bar{\varphi} \in W_0^{1,p'}(a, b)$, $\bar{\varphi}$ is continuous. Therefore $\{\bar{\varphi} < 0\}$ is an open subset of (a, b). Thus $\{\bar{\varphi} < 0\} = \cup_k (a_k, b_k)$, where (a_k, b_k) are mutually disjoint. For each (a_k, b_k),

$$-(|\bar{y}'|^{p-2}\bar{y}')' = 0 \ \text{ in } (a_k, b_k).$$

Hence

$$\bar{y}' = C_k \ \text{ in } (a_k, b_k)$$

for some constant C_k. Noting that $\bar{\varphi}(a_k) = \bar{\varphi}(b_k) = 0$, $\bar{y} \in W_0^{1,p}(a, b)$, and $\bar{\varphi}' = 0$, a.e. in $\{\bar{\varphi} = 0\}$, by Lemma 3.2, we have

$$\begin{aligned}
& \int_a^b -\bar{\varphi}''\bar{y}dx = \int_a^b \bar{\varphi}'\bar{y}'dx = \int_{\{\bar{\varphi}<0\}} \bar{\varphi}'\bar{y}'dx = \sum_k \int_{a_k}^{b_k} \bar{\varphi}'\bar{y}'dx \\
& = \sum_k \int_{a_k}^{b_k} \bar{\varphi}'C_k dx = \sum_k C_k[\bar{\varphi}(a_k) - \bar{\varphi}(b_k)] = 0.
\end{aligned}$$

Similarly,

$$\int_a^b -\tilde{\varphi}''\tilde{y}dx = 0.$$

Therefore, by (5.3), (5.9), and Lemma 5.1,

$$\begin{aligned}&\int_a^b \{[|z-\tilde{y}|^{p-2}(z-\tilde{y}) - |z-\bar{y}|^{p-2}(z-\bar{y})][(z-\tilde{y}) - (z-\bar{y})] \\ &\qquad + (|\bar{y}'|^{p-2}\bar{y}' - |\tilde{y}'|^{p-2}\tilde{y}')(\bar{y}'-\tilde{y}')\}dx \\ &= \int_a^b \{-\bar{\varphi}''\tilde{y} - \tilde{\varphi}''\bar{y}\}dx = \int_a^b \{-\bar{\varphi}'\tilde{y}' - \tilde{\varphi}'\bar{y}'\}dx \le 0.\end{aligned}$$

Hence, by Lemma 3.3, $\bar{y} = \tilde{y}$. We complete the proof. □

When the domain is B_R and z is a radial function, we are interested in finding a minimizer of I over $W_0^{1,p}(B_R) \cap S_R$, where S_R denotes the set of all radial functions in B_R. We can also give results similar to those of Theorem 5.2.

THEOREM 5.3. *Suppose $1 < p < +\infty$ and $z \in L^p(B_R) \cap S_R$. Then $I(\cdot)$ admits a unique minimizer $\bar{y}$ over $W_0^{1,p}(B_R) \cap S_R$. We have $\bar{y} \in \mathcal{H}_+^p(B_R) \cap C^1(B_R \setminus \{0\})$. Moreover, if $z^+ \in L^q(B_R)$ for some $q > (p-1)n$, then $\bar{y} \in C^{1,\alpha}(\bar{B}_R)$ for some $\alpha \in (0,1)$.*

In the following theorems, $\bar{y}$ denotes the minimizer of $I(\cdot)$ over $W_0^{1,p}(B_R) \cap S_R$ corresponding to $z \in L^p(B_R) \cap S_R$.

THEOREM 5.4. *Let*

$$\bar{\psi}(x) = \bar{\psi}(|x|) = \int_{|x|}^R dr \int_0^r \xi^{n-1}|z(\xi)|^{p-2}z(\xi)r^{\frac{1-n}{p-1}}d\xi, \quad 0 < |x| \le R. \tag{5.10}$$

Then, $\bar{y} \equiv 0$ if and only if $\bar{\psi} \le 0$ in $\bar{B}_R \setminus \{0\}$.

THEOREM 5.5. *Suppose $1 < p < 2$, $\bar{\psi} \not\equiv 0$, where $\bar{\psi}$ is defined by (5.10) in $\bar{B}_R \setminus \{0\}$. Denote*

$$s = \sup\{r \in [0,R] | \nabla\bar{y} = 0, \text{ a.e. in } B_r\}. \tag{5.11}$$

Then $s \in [0,R)$, and

$$\nabla\bar{y} \ne 0 \quad \text{in } B_R \setminus \bar{B}_s. \tag{5.12}$$

(i) *If $s = 0$, then there exists a $\bar{\varphi} \in W_0^{1,\frac{p+1}{2}}(B_R) \cap C^1(\bar{B}_R \setminus \{0\})$ such that*

$$\bar{\varphi} \le 0 \quad \text{in } \bar{B}_R \setminus \{0\}, \tag{5.13}$$

$$\begin{cases} -\mathrm{div}(|\nabla\bar{y}|^{p-2}\nabla\bar{\varphi}) = |z-\bar{y}|^{p-2}(z-\bar{y})\chi_{\{\bar{\varphi}<0\}} & \text{in } B_R, \\ \bar{\varphi}|_{\partial B_R} = 0, \end{cases} \tag{5.14}$$

$$\begin{cases} -\mathrm{div}(|\nabla\bar{y}|^{p-2}\nabla\bar{y}) = |z-\bar{y}|^{p-2}(z-\bar{y})\chi_{\{\bar{\varphi}=0\}} & \text{in } B_R, \\ \bar{y}|_{\partial B_R} = 0. \end{cases} \tag{5.15}$$

Moreover, if $z \in L^q(B_R)$ for some $q > n(p-1)$, then $\bar{\varphi} \in C^1(\bar{B}_R)$. Consequently, $\bar{\varphi}(0) = 0$, $\nabla\bar{\varphi}(0) = 0$.

(ii) *If $s > 0$, then*

$$\int_{B_s} |z-\bar{y}|^{p-2}(z-\bar{y})dx = 0, \tag{5.16}$$

and there exists a $\bar{\varphi} \in C^1(\bar{B}_R \setminus B_s)$ such that

$$\bar{\varphi} \le 0 \quad \text{in } B_R \setminus \bar{B}_s, \tag{5.17}$$

$$\begin{cases} -\mathrm{div}(|\nabla\bar{y}|^{p-2}\nabla\bar{\varphi}) = |z-\bar{y}|^{p-2}(z-\bar{y})\chi_{\{\bar{\varphi}<0\}} & \text{in } B_R \setminus \bar{B}_s, \\ \bar{\varphi}|_{\partial B_R \cup \partial B_s} = 0, \nabla\bar{\varphi}|_{\partial B_s} = 0, \end{cases} \tag{5.18}$$

and

$$\begin{cases} -\mathrm{div}(|\nabla\bar{y}|^{p-2}\nabla\bar{y}) = |z-\bar{y}|^{p-2}(z-\bar{y})\chi_{\{\bar{\varphi}=0\}\cap(B_R\setminus\bar{B}_s)} & \text{in } B_R, \\ \bar{y}|_{\partial B_R} = 0. \end{cases} \tag{5.19}$$

THEOREM 5.6. *Suppose $2 < p < +\infty$. Then there exists a $\bar{\varphi} \in W_0^{1,p'}(B_R) \cap C^1(\bar{B}_R \setminus \{0\})$ such that*

$$\bar{\varphi} \le 0 \quad \text{in } \bar{B}_R, \tag{5.20}$$

$$\begin{cases} -\mathrm{div}(|x|^{-\gamma}\nabla\bar{\varphi}) = |z-\bar{y}|^{p-2}(z-\bar{y})\chi_{\{\bar{\varphi}<0\}} & \text{in } B_R, \\ \bar{\varphi}|_{\partial B_R} = 0, \end{cases} \tag{5.21}$$

$$\begin{cases} -\mathrm{div}(|\nabla\bar{y}|^{p-2}\nabla\bar{y}) = |z-\bar{y}|^{p-2}(z-\bar{y})\chi_{\{\bar{\varphi}=0\}} & \text{in } B_R, \\ \bar{y}|_{\partial B_R} = 0, \end{cases} \tag{5.22}$$

where $\gamma = \frac{(n-1)(p-2)}{p-1}$.

Though $\mathcal{H}_+^p(B_R)$ is not convex when $p \neq 2$ and $n > 1$, $\mathcal{H}_+^p(B_R) \cap S_R$ is convex by a straightforward computation. The proof of Theorem 5.3 is based on the fact that $\mathcal{H}_+^p(B_R) \cap S_R$ is a closed and convex subset of $W_0^{1,p}(B_R)$. The proofs of Theorems 5.4–5.6, which are somewhat similar to the proof of Theorem 5.2, need careful calculations. We omit the proofs.

REMARK 5.7. *If $\bar{\psi}$ in Theorem* 5.4 *has good regularity, then it is a solution of an equation like* (5.21) *in Theorem* 5.6.

REMARK 5.8. *Comparing Theorem* 5.5(ii) *with Theorems* 5.2 *and* 5.6, *we can see that the condition we get in case $1 < p < 2$ on the singular set $\{\nabla\bar{y} = 0\}$ is relatively weak. Roughly speaking,* (5.16) *is equivalent to $\bar{\varphi}(s) = 0$.*

REMARK 5.9. *In Theorem* 5.5, *if $\bar{u} \neq 0$, a.e. in $B_{s+\delta} \setminus B_s$ for some $\delta > 0$, then in* (5.14) *and* (5.18), $-\mathrm{div}(|\nabla\bar{y}|^{p-2}\nabla\bar{\varphi})$ *can be replaced by* $-\mathrm{div}(|x|^{-\gamma}\nabla\bar{\varphi})$ *as in* (5.21).

The results in Theorems 5.2, 5.5, and 5.6 are somewhat surprising. We guessed that on the singular set $\Omega_0 \equiv \{\nabla\bar{y} = 0\}$, $\bar{\varphi}$ in these theorems would satisfy $\bar{\varphi} \le 0$ and (if Ω_0 is a domain)

$$\begin{cases} -\triangle\bar{\varphi} = |z-\bar{y}|^{p-2}(z-\bar{y}) & \text{in } \Omega_0, \\ \bar{\varphi}|_{\partial\Omega_0} = 0. \end{cases}$$

Theorems 5.3 and 5.6 show that this is not the case. Though $\bar{y}_\varepsilon \to \bar{y}$ uniformly in $C^1(\bar{\Omega})$ under proper conditions, the set $\{\nabla\bar{y} = 0\}$ may be quite different from $\{\nabla\bar{y}_\varepsilon = 0\}$.

Acknowledgments. This paper was written under the guidance of Professor Jiongmin Yong. Many valuable suggestions were offered by Professor Xunjing Li. The author thanks both of them for their help.

REFERENCES

[1] D. R. Adams, S. M. Lenhart, and J. Yong, *Optimal control of the obstacle for an elliptic variational inequality*, Appl. Math. Optim., 38 (1998), pp. 121–140.
[2] R. A. Adams, *Sobolev Spaces*, Academic, New York, 1975.
[3] V. Barbu, *Optimal Control of Variational Inequalities*, Pitman, London, 1984.
[4] A. Bermudez and C. Saguze, *Pointwise control of variational inequality*, in Free Boundary Problems: Theory and Applications II, K. H. Hoffmann and J. Sprekels, eds., Pitman Res. Notes Math. Ser. 186, Longman Scientific and Technical, Harlow, UK, 1990, pp. 475–478.
[5] H. Brezis and D. Kinderlehrer, *The smoothness of solutions to nonlinear variational inequalities*, Indiana Univ. Math. J., 23 (1974), pp. 831–844.
[6] E. Casas and L. A. Fernández, *Optimal control of quasilinear elliptic equations with non-differentiable coefficients at the origin*, Rev. Mat. Univ. Complut. Madrid, 4 (1991), pp. 227–250.
[7] E. Casas and L. A. Fernández, *Distributed control of systems governed by a general class of quasilinear elliptic equations*, J. Differential Equations, 104 (1993), pp. 20–47.
[8] Q. Chen, *Indirect obstacle control problem for semilinear elliptic variational inequalities*, SIAM J. Control Optim., 38 (1999), pp. 138–158.
[9] Q. Chen, *A nonlinear parabolic system arising from the eddy currents problem*, Nonlinear Analysis, 42 (2000), pp. 759–770.
[10] Q. Chen, *Optimal control of semilinear elliptic variational bilateral problem*, Acta Math. Sin. (Engl. Ser.), 16 (2000), pp. 123–140.
[11] J. I. Diaz, *Nonlinear Partial Differential Equations and Free Boundaries, Vol.* I. *Elliptic Equations*, Res. Notes in Math. 106, Pitman, London, 1985.
[12] E. DiBenedetto, *$C^{1,\alpha}$ local regularity of weak solutions of degenerate elliptic equations*, Nonlinear Analysis, 7 (1983), pp. 827–850.
[13] A. Friedman, *Variational Principles and Free-boundary Problems*, Wiley, New York, 1982.
[14] A. Friedman, *Optimal control for variational inequalities*, SIAM J. Control Optim., 24 (1986), pp. 439–451.
[15] A. Friedman, *Optimal control for parabolic variational inequalities*, SIAM J. Control Optim., 25 (1987), pp. 482–497.
[16] A. Friedman and K.-H. Hoffmann, *Control of free boundary problems with hysteresis*, SIAM J. Control Optim., 26 (1988), pp. 42–55.
[17] A. Friedman, S. Huang, and J. Yong, *Bang-bang optimal control for the dam problem*, Appl. Math. Optim., 15 (1987), pp. 65–85.
[18] A. Friedman, S. H. Huang, and J. M. Yong, *Optimal periodic control for the two-phase Stefan problem*, SIAM J. Control Optim., 26 (1988), pp. 23–41.
[19] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, 1983.
[20] P. Jaillet, D. Lamberton, and B. Lapeyre, *Variational inequalities and the pricing of American options*, Acta Appl. Math., 21 (1990), pp. 263–289.
[21] I. Karatzas, *On the pricing of American options*, Appl. Math. Optim., 17 (1988), pp. 37–60.
[22] J. L. Lewis, *Smoothness of certain degenerate elliptic equations*, Proc. Amer. Math. Soc., 80 (1980), pp. 259–265.
[23] X. Li and J. Yong, *Optimal Control Theory for Infinite Dimensional Systems*, Birkhäuser, Boston, 1995.
[24] F. H. Lin and Y. Li, *Boundary $C^{1,\alpha}$-regularity for variational inequalities*, Comm. Pure Appl. Math., 44 (1991), pp. 715–732.
[25] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.
[26] H. Lou, *On the regularity of an obstacle control problem*, J. Math. Anal. Appl., 258 (2001), pp. 32–51.
[27] H. Lou, *On Singular Sets of Local Solutions to p–Laplace Equation*, Nonlinear Anal., submitted.
[28] F. Mignot and J.-P. Puel, *Optimal control in some variational inequalities*, SIAM J. Control Optim., 22 (1984), pp. 466–476.
[29] C. B. Morrey Jr., *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, Berlin, 1966.

[30] J. F. Rodrigues, *Obstacle Problems in Mathematical Physics*, North-Holland Math. Stud. 134, North-Holland, Amsterdam, The Netherlands, 1987.
[31] H. Royden, *Real Analysis*, 2nd ed., Collier-Macmillan, New York, 1968.
[32] T. W. Ting, *Elastic-plastic torsion of a square bar*, Trans. Amer. Math. Soc., 123 (1966), pp. 369–401.
[33] P. Tolksdorf, *Regularity for a more general case of quasilinear elliptic equations*, J. Differential Equations, 51 (1984), pp. 126–150.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 4, pp. 1254–1280

ASYMPTOTIC CONTROL OF PAIRS OF OSCILLATORS COUPLED BY A REPULSION, WITH NONISOLATED EQUILIBRIA I: THE REGULAR CASE*

ALEXANDRE CABOT† AND MARC-OLIVIER CZARNECKI†

Abstract. Let $\phi : H \to \mathbb{R}$ be a $\mathcal{C}^1$ function on a real Hilbert space H and let $\gamma > 0$ be a positive damping parameter. For any repulsive potential $V : H \to \mathbb{R}_+$ and any control function $\varepsilon : \mathbb{R}_+ \to \mathbb{R}_+$ which tends to zero as $t \to +\infty$, we study the asymptotic behavior of the trajectories of the coupled dissipative system of nonlinear oscillators

$$\text{(HBFC}^2\text{)} \qquad \begin{cases} \ddot{x} + \gamma\dot{x} + \nabla\phi(x) + \varepsilon(t)\nabla V(x-y) = 0, \\ \ddot{y} + \gamma\dot{y} + \nabla\phi(y) - \varepsilon(t)\nabla V(x-y) = 0. \end{cases}$$

We first provide general existence results and show that $\nabla\phi(x(t)) \to 0$ and $\nabla\phi(y(t)) \to 0$ when $t \to +\infty$, assuming either that the trajectory (x, y) is bounded, or that the potential V is bounded and that ϕ satisfies the following limit condition:

(LIM) For every sequence $(z_n) \subset H$ such that $\lim_{n\to+\infty} |z_n| = +\infty$, there exists a subsequence $(z_{\varphi(n)})$ such that

$$\lim_{n\to+\infty} \phi(z_{\varphi(n)}) = +\infty \qquad \text{or} \qquad \lim_{n\to+\infty} \nabla\phi(z_{\varphi(n)}) = 0.$$

If $\varepsilon(t)$ does not tend to zero too rapidly as $t \to +\infty$, then the term $\varepsilon(t)\nabla V(x-y)$ asymptotically repulses the trajectories one from the other. Precisely, when $H = \mathbb{R}$, and if ε is a "slow" control, i.e., $\int_0^{+\infty} \varepsilon(t)dt = +\infty$, then the trajectories x and y converge to extremal points of the set $S = \{\lambda \in \mathbb{R}, \nabla\phi(\lambda) = 0\}$ of the equilibria of ϕ (when $S \neq \emptyset$), or they have the same limit. In particular, when S is reduced to an interval—for example, if ϕ is convex—this allows us to obtain a global description of the set S. We provide numerical experiments which make our convergence results more precise.

Key words. nonlinear oscillator, coupled system, slow control, heavy ball with friction, global optimization

AMS subject classifications. Primary, 37N40, 34G20; Secondary, 34H05, 34D05, 34E10, 49K15, 70F99

PII. S0363012901385198

1. Introduction.

(a) Let H be a real Hilbert space, with scalar product and corresponding norm, respectively, denoted by $\langle ., . \rangle$ and $|\ .\ |$. Let $\phi : H \to \mathbb{R}$ be a given $\mathcal{C}^1$ real-valued function called the potential function. An important problem is the search for the equilibria of the function ϕ (i.e., the solutions of the equation $\nabla\phi(x) = 0$, where $\nabla\phi$ is the gradient of ϕ), among which the minima (global or local) play a particular role in the fields of optimization, physics, and economics, to name a few. To obtain equilibria or minima of the function ϕ, a powerful method is to follow the trajectories of an associated dissipative gradient-like dynamical system, possibly discretized for numerical applications. In many practical problems (for example, minimizing a convex function which is not strictly convex), the function ϕ has nonisolated equilibria, and one wants to choose a particular equilibrium or minimum. One may also desire a full and global description of the set of equilibria or minima of ϕ, and also estimations of its size. We provide some examples in section 3.1.

*Received by the editors February 17, 2001; accepted for publication (in revised form) April 1, 2002; published electronically October 29, 2002.
http://www.siam.org/journals/sicon/41-4/38519.html
†ACSIOM, CNRS-FRE 2311, Université Montpellier 2, place Eugène Bataillon, 34095 Montpellier cedex 5, France (cabot@math.univ-montp2.fr, marco@math.univ-montp2.fr).

It is not obvious that a dynamical system may help in these last cases, since a trajectory would likely lead to a single equilibrium at the limit when $t \to +\infty$. However, one could conceive, for example, that the cluster points of a dynamical system associated to the function ϕ could correspond exactly to the set of its minima. Even so, numerical applications might be tricky since the cluster points are likely to be more difficult to identify than simply limits. Another direction, which is our concern in this paper, is to consider coupled dynamical systems, which exchange information and thus are more able to globally explore the function ϕ than a single noncoupled system. Precisely, in order to obtain a global dynamical approach of the equilibria of the function ϕ, we study the asymptotic behavior of the following second order (in time) coupled gradient-like system:

$$\text{(HBFC}^2\text{)} \qquad \begin{cases} \ddot{x} + \gamma\dot{x} + \nabla\phi(x) + \varepsilon(t)\nabla V(x-y) = 0, \\ \ddot{y} + \gamma\dot{y} + \nabla\phi(y) - \varepsilon(t)\nabla V(x-y) = 0, \end{cases}$$

where $\gamma > 0$ is a positive damping parameter, $\varepsilon : \mathbb{R}_+ \to (0, +\infty)$ is a control function such that $\lim_{t\to+\infty} \varepsilon(t) = 0$, and $V : H \to \mathbb{R}_+$ is a (coupling) potential function.

(b) Let us briefly explain what led us to consider the (HBFC^2) system. To simply obtain equilibria or local minima of the function ϕ, the classical steepest descent method

$$\text{(SD)} \qquad \dot{x}(t) + \nabla\phi(x(t)) = 0$$

gives positive results (see Bruck's theorem [11], Lojasiewicz theorem [19, 20]; see also [9, 10], etc.) and is a descent method which corresponds to the motion of a drop of water on the graph of ϕ. Introducing acceleration in the motion, precisely considering a second order in time dynamical system, is likely to lead to more global exploration properties since it would not necessarily stop at the first equilibrium that it encounters. A particularly important system, the Heavy Ball with Friction system,

$$\text{(HBF)} \qquad \ddot{x}(t) + \gamma\dot{x}(t) + \nabla\phi(x(t)) = 0,$$

corresponds to the motion of a material point with positive mass, subjected to stay on the graph of ϕ. It is not a descent method but still a dissipative system ($\gamma > 0$ is the friction parameter) and enjoys many minimizing properties (see the recent papers of Alvarez [1], Attouch, Goudou, and Redont [6], Goudou [15], Haraux and Jendoubi [17], and Jendoubi [18]). However, only the first problem of reaching one—nonspecified—minimum of the function ϕ is answered.

In order to obtain more specifications on the properties of the equilibria which are selected as limits of the trajectories, Attouch and Czarnecki [4] consider the system

$$\text{(HBFC)} \qquad \ddot{x}(t) + \gamma\dot{x}(t) + \nabla\phi(x(t)) + \varepsilon(t)x(t) = 0$$

(Heavy Ball with Friction and Control) by introducing a Tikhonov-like[1] asymptotic regularization term $\varepsilon(t)x(t)$. When ϕ is convex and $\varepsilon : \mathbb{R}_+ \to \mathbb{R}_+$ is a $\mathcal{C}^1$ control function which tends to zero slowly, i.e., such that $\int_0^{+\infty} \varepsilon(t)dt = +\infty$, they proved that each trajectory of the (HBFC) system strongly converges to the point of minimal norm of the set $S = \operatorname{argmin}\phi$ [2] (which is assumed to be nonempty). The condition

[1] For the theory of Tikhonov regularization, we refer to [24].

[2] Precisely defined by $\operatorname{argmin}\phi = \{z \in H | \forall z' \in H, \phi(z) \leq \phi(z')\}$.

$\int_0^{+\infty} \varepsilon(t)dt = +\infty$ corresponds to the fact that $\varepsilon(t)$ does not tend to zero too rapidly, thus allowing the Tikhonov regularization term $\varepsilon(t)x(t)$ to be effective asymptotically.

This result shows in fact an asymptotic selection property: the effect of such a slow control ε forces all the trajectories to converge to the same equilibrium, namely, the equilibrium of minimal norm. This situation sharply contrasts with the noncontrolled situation (or fast control) where the limits of the trajectories are only weak limits, depend on the initial data, and may also be difficult to identify. The idea of coupling approximation methods with the dynamics of a gradient-like system had already been considered at the first order for the steepest descent by Attouch and Cominetti [3] (for related topics, see also Furuya, Miyashiba, and Kenmochi [14] and Baillon and Cominetti [8]).

(c) It is then a natural question to know if it is possible to adapt the selection properties of the Tikhonov regularization with a coupling potential in order to allow two different trajectories to "exchange" information so that they explore different parts of the set of equilibria $S = \{\lambda \in \mathbb{R}, \nabla\phi(\lambda) = 0\}$. Note that the (HBFC^2) system has a similar mechanical interpretation as the (HBF) system with an extra repulsion force—deriving from the potential V—between the two "balls" (for example, in the case where $V(z) = 1 - |z|^2$ ($|z| < 1$), it corresponds to a repulsive spring of varying stiffness $\varepsilon(t)$). One can easily conceive that if $\varepsilon(t)$ does not tend to zero too rapidly, the mechanical system will select two equilibria as far as possible from each other. Note also that the coupling term $\varepsilon(t)\nabla V(x-y)$ can be viewed as a time-varying feedback (see Coron [12] for a survey on stabilization of nonlinear systems by nonautonomous feedbacks).

The first main result of the paper (corresponding to Proposition 2.1 and part of Theorem 2.1 below) shows that $\nabla\phi(x(t)) \to 0$ and $\nabla\phi(y(t)) \to 0$ when $t \to +\infty$.

THEOREM A. *Assume that $\phi : H \to \mathbb{R}$ is a $\mathcal{C}^1$ function, bounded from below, such that $\nabla\phi$ is Lipschitz continuous on the bounded sets. Let $V : H \to \mathbb{R}_+$ be a C^1 function such that ∇V is locally Lipschitz continuous on H and bounded on the bounded subsets of H. Let $\varepsilon : [0, +\infty) \to \mathbb{R}_+$ be a C^1 function such that $\lim_{t\to+\infty} \varepsilon(t) = 0$ and $\dot\varepsilon(t) \le 0$ for every $t \in \mathbb{R}_+$. Then, for every $((x_0, y_0), (\dot x_0, \dot y_0)) \in H^2 \times H^2$, there is a unique solution $(x, y) : [0, +\infty) \to H^2$ of the* (HBFC^2) *Cauchy problem:*

$$\text{(HBFC}^2\text{)} \qquad \begin{cases} \ddot x + \gamma\dot x + \nabla\phi(x) + \varepsilon(t)\nabla V(x-y) = 0, \\ \ddot y + \gamma\dot y + \nabla\phi(y) - \varepsilon(t)\nabla V(x-y) = 0, \\ \Big(x(0), y(0), \dot x(0), \dot y(0)\Big) = \Big(x_0, y_0, \dot x_0, \dot y_0\Big). \end{cases}$$

Moreover,

$$\lim_{t\to+\infty} \nabla\phi(x(t)) = \lim_{t\to+\infty} \nabla\phi(y(t)) = 0$$

if one of the two following assumptions holds:

(a) *The trajectory (x, y) is bounded.*

(b) *The map $t \mapsto V(x(t) - y(t))$ is bounded and the map ϕ satisfies the limit condition*

(LIM) *for every sequence $(z_n) \subset H$ such that $\lim_{n\to+\infty} |z_n| = +\infty$, there exists a subsequence $(z_{\varphi(n)})$ such that*

$$\lim_{n\to+\infty} \phi(z_{\varphi(n)}) = +\infty \qquad \text{or} \qquad \lim_{n\to+\infty} \nabla\phi(z_{\varphi(n)}) = 0.$$

The second main result of the paper (corresponding to Theorem 2.2 and Corollary 2.4 below) is the convergence of the trajectories in the one-dimensional case. Let

$$\widehat{S} = \left\{ \overline{z} \in \overline{\mathbb{R}}, \lim_{z \to \overline{z}} \phi'(z) = 0 \right\}.$$

THEOREM B. *Under the assumptions of Theorem* A, *additionally assume that* $H = \mathbb{R}$ *and that* $zV'(z) \leq 0$ *for every* $z \in \mathbb{R}$ *(repulsion). Then the solution* (x, y) *satisfies the following asymptotical behavior:*

(i) *If* ϕ *satisfies* (LIM) *or if the trajectory* (x, y) *is bounded, there exists* $(x_\infty, y_\infty) \in \widehat{S} \times \widehat{S}$ *such that* $\lim_{t \to +\infty}(x(t)\,,\, y(t)) = (x_\infty\,,\, y_\infty)$.

(ii) *(Slow parametrization.) Additionally assume that* $\int_0^{+\infty} \varepsilon(t)\, dt = +\infty$, *that* $\widehat{S}$ *is an interval, and that for every* $z \neq 0$, $V'(z) \neq 0$. *Then one of the following cases holds:*

- $\lim_{t \to +\infty}(x(t)\,,\, y(t)) = (\sup \widehat{S}\,,\, \inf \widehat{S})$;
- $\lim_{t \to +\infty}(x(t)\,,\, y(t)) = (\inf \widehat{S}\,,\, \sup \widehat{S})$;
- $\lim_{t \to +\infty} x(t) = \lim_{t \to +\infty} y(t) \in \widehat{S}$.

In fact our convergence results are more precise and are not restricted to the case where $\widehat{S}$ is a nonempty interval. The first two cases give a global description of the set $\widehat{S}$ and give an answer to the initial problem of the global exploration of the equilibria of the potential ϕ. The last case does not provide global information and could be viewed as a drawback of our results. But numerical experiments (with different potentials ϕ and V) lead to the conjecture that this last case only happens on a negligible set of initial data, with a possibly fractal structure.

In order to precisely determine the convergence properties of the (HBFC^2) system in the infinite-dimensional setting, we show the weak convergence of the trajectories with a convex potential ϕ, and a "fast" control ε, i.e., such that $\int_0^{+\infty} \varepsilon(t)\, dt < +\infty$ (Proposition 2.2). But with a "fast" control, the (weak) limits depend on the initial data and are in general difficult to identify.

The one-dimensional case should be seen as a first step in which we are able to give precise results concerning the convergence, with a "slow control" ε. Though we provide remarks and counterexamples for such precision in higher dimensions, we believe that our paper could give a direction and justification for further extension of the global exploration of the equilibria of a (possibly nonconvex) function ϕ. Our numerical experiments seem to indicate that in "most" cases the coupled systems would lead to a full description of the set of the minima of ϕ, possibly involving a coupled finite number N of systems, with N greater than 2.

In numerical applications and simulations, the system (HBFC^2) is approximated by a discretized system, thus requiring one to compute only the potential on the sequence (x_n, y_n) of the discretized trajectory. For studies in this direction, see, for example, [2]. Note that the control term $\varepsilon(t)\nabla V(x-y)$ can be viewed as a perturbation term and thus allows us to take into account other perturbations (for example, of the potential ϕ). From the numerical point of view, it is essential to study the stability of the trajectory behavior of the (discretized) system under a perturbation of the potential, thus allowing for errors on the computation of ϕ and $\nabla\phi$ (which also can be interesting for the speed of computations). These studies are beyond the scope of our paper, but encouraging results are obtained in [21] by using incomplete sensitivities.

Besides its first optimization scope, note that applications are also to be found in the study of physical coupled systems, which of course requires results in higher dimensions.

(d) The paper is organized as follows. In section 2.1, we precisely state the global existence results and general asymptotic properties (Proposition 2.1 and Theorem 2.1), which are not reduced to the one-dimensional case. In section 2.2, we precisely state the asymptotic convergence results (Theorems 2.2 and 2.3 and Corollaries 2.3 and 2.4). In section 3, we provide remarks on and counterexamples of our results. The results are proved in section 4 (global existence) and section 5 (asymptotic convergence). Finally, in section 6, we give numerical results making our theoretical results more precise, and we indicate some possible directions for further research on the subject.

2. Main results. In this paper, we assume the following (rather standard) set of hypotheses.

HYPOTHESIS 1. *Let H be a real Hilbert space. Let us consider a map $\phi : H \to \mathbb{R}$ of class $\mathcal{C}^1$ which satisfies the following conditions:*

$$(\mathcal{H}_\phi) \quad \begin{cases} \text{(i)} & \text{the map } \phi \text{ is bounded from below on } H; \\ \text{(ii)} & \text{the map } \nabla\phi \text{ is Lipschitz continuous on the bounded subsets of } H. \end{cases}$$

Let $V : H \to \mathbb{R}_+$ be a map of class $\mathcal{C}^1$ such that

$$(\mathcal{H}_V) \quad \begin{cases} \text{(i)} & \text{the map } \nabla V \text{ is locally Lipschitz continuous on } H; \\ \text{(ii)} & \text{the map } \nabla V \text{ is bounded on the bounded subsets of } H. \end{cases}$$

Let $\varepsilon : [0, +\infty) \to \mathbb{R}_+$ be a function of class $\mathcal{C}^1$ such that

$$(\mathcal{H}_\varepsilon) \quad \begin{cases} \text{(i)} & \text{the function } \varepsilon \text{ is nonincreasing, i.e., } \forall t \in \mathbb{R}_+,\ \dot{\varepsilon}(t) \leq 0; \\ \text{(ii)} & \lim_{t \to +\infty} \varepsilon(t) = 0. \end{cases}$$

Let $\gamma > 0$, $((x_0, y_0), (\dot{x}_0, \dot{y}_0)) \in H^2 \times H^2$; the (HBFC^2) *system is defined as follows:*

$$\text{(HBFC}^2\text{)} \quad \begin{cases} \ddot{x} + \gamma\dot{x} + \nabla\phi(x) + \varepsilon(t)\nabla V(x - y) = 0, \\ \ddot{y} + \gamma\dot{y} + \nabla\phi(y) - \varepsilon(t)\nabla V(x - y) = 0, \\ \Big(x(0), y(0), \dot{x}(0), \dot{y}(0)\Big) = \Big(x_0, y_0, \dot{x}_0, \dot{y}_0\Big). \end{cases} \tag{2.1}$$

Remark. For the sake of readability, we take 0 as initial time. All the results of the paper clearly hold by taking any other initial time $t_0 \in \mathbb{R}$ and making the corresponding adaptations in the statements.

2.1. Global properties. The following proposition ensures the global existence of the solutions of the (HBFC^2) system.

PROPOSITION 2.1 (global existence). *Assume Hypothesis* 1. *Then*

(i) *there exists a unique maximal solution $(x, y) : [0, +\infty) \to H \times H$ of* (HBFC^2) *which is of class $\mathcal{C}^2$;*

(ii) $(\dot{x}, \dot{y}) \in L^\infty([0, +\infty); H \times H) \cap L^2([0, +\infty); H \times H)$ *and* $(\phi(x), \phi(y)) \in L^\infty([0, +\infty); \mathbb{R} \times \mathbb{R})$.

Proposition 2.1 is proved in section 4.1. The next result summarizes the global (convergence) properties of the solutions of the (HBFC^2) system.

THEOREM 2.1. *Under the assumptions of Proposition* 2.1 *(Hypothesis* 1*), additionally assume one of the two following assumptions:*

(a) *The trajectory (x, y) is bounded.*

(b) *The map $t \mapsto V(x(t) - y(t))$ is bounded and the map ϕ satisfies the limit condition*

(LIM) *For every sequence* $(z_n) \subset H$ *such that* $\lim_{n\to+\infty} |z_n| = +\infty$, *there exists a subsequence* $(z_{\varphi(n)})$ *such that*

$$\lim_{n\to+\infty} \phi(z_{\varphi(n)}) = +\infty \qquad \text{or} \qquad \lim_{n\to+\infty} \nabla\phi(z_{\varphi(n)}) = 0.$$

Then

(iii) $\lim_{t\to+\infty} \dot{x}(t) = \lim_{t\to+\infty} \dot{y}(t) = 0$;

(iv) $\lim_{t\to+\infty} \nabla\phi(x(t)) = \lim_{t\to+\infty} \nabla\phi(y(t)) = 0$.

Additionally assume that the map ϕ *is convex. Then, with* $S = \{z \in H, \nabla\phi(z) = 0\}$,

(v) *if* x *(resp.,* y*) is bounded, then* $\lim_{t\to+\infty} \phi(x(t)) = \inf \phi$ *(resp.,* $\lim_{t\to+\infty} \phi(y(t)) = \inf \phi$*)*;

(vi) *if* x_∞ *(resp.,* y_∞*) is a weak cluster point of* x *(resp.,* y*), then* x_∞ *(resp.,* y_∞*) belongs to* S.

Theorem 2.1 is proved in section 4.2. Note that Proposition 2.1 and Theorem 2.1 clearly imply Theorem A in the introduction.

Remark. Note that the condition (LIM) is equivalent to the following assertion:

$$\forall \alpha > 0,\ \forall A \in \mathbb{R}, \qquad \text{the set } \{z \in H, \phi(z) \le A \text{ and } |\nabla\phi(z)| \ge \alpha\} \text{ is bounded.}$$

The condition (LIM) is clearly satisfied in the two following simpler cases:

(c) The map ϕ is coercive, i.e., $\lim_{|z|\to+\infty} \phi(z) = +\infty$.

(d) $\lim_{|z|\to+\infty} \nabla\phi(z) = 0$.

In view of Proposition 2.1(ii), the map ϕ is bounded and the assumption (c) implies that the trajectory (x, y) is bounded. In this case, the assumption that the map $t \mapsto V(x(t) - y(t))$ is bounded is automatically satisfied.

We provide additional remarks and counterexamples in section 3.2.

2.2. Convergence in the one-dimensional case. Once the (global) existence is acquired, the main point in the study of a dissipative system is to investigate the convergence properties of the solution map. In this section, we explore the convergence and stabilizing properties of the solutions of the (HBFC^2) system under additional assumptions. We assume that the space H is one-dimensional (i.e., $H = \mathbb{R}$). Note that since the map $V' : \mathbb{R} \to \mathbb{R}$ is continuous, it is bounded on the bounded sets and assumption $(\mathcal{H}_V)$(ii) is automatically satisfied. Since our point is a global exploration, and in order to push the trajectories away one from another, we consider the case where the potential V is a repulsion, i.e., satisfies

$$(\mathcal{H}_V)\text{(iii)} \qquad \forall z \in \mathbb{R}, \quad zV'(z) \le 0.$$

2.2.1. Convergence of the trajectory. The next result gives the general convergence and stabilizing properties of the solutions of the (HBFC^2) system without further assumptions on the control ε. In the following, S denotes the set of the equilibria of ϕ: $S = \{z \in \mathbb{R}, \phi'(z) = 0\}$. In order to give a unified presentation of our results, we define the set $\widehat{S}$ as the union of S, of $\{+\infty\}$ if $\lim_{z\to+\infty} \phi'(z) = 0$, and of $\{-\infty\}$ if $\lim_{z\to-\infty} \phi'(z) = 0$. Equivalently,

$$\widehat{S} = \left\{ \overline{z} \in \overline{\mathbb{R}}, \lim_{z\to\overline{z}} \phi'(z) = 0 \right\}.$$

Finally, in order to include the case where the trajectories x and y may not be bounded, we consider the limit condition (LIM). Since $H = \mathbb{R}$, the limit condition (LIM) is satisfied if the map ϕ is convex and bounded from below.

THEOREM 2.2 (convergence of the solutions). *Assume Hypothesis* 1, *with* $H = \mathbb{R}$, *together with* $(\mathcal{H}_V)$(iii), $\forall z \in \mathbb{R},\ zV'(z) \leq 0$.

Assume that the map ϕ *satisfies* (LIM) *(for example, if* ϕ *is convex) or that the trajectory* (x, y) *is bounded.*[3] *Let (x,y) be the solution of the* (HBFC^2) *system. Then there exists* $(x_\infty, y_\infty) \in \widehat{S} \times \widehat{S}$ *such that*

$$\lim_{t \to +\infty} (x(t)\,,\, y(t)) = (x_\infty\,,\, y_\infty).$$

Theorem 2.2 is proved in section 5.1 and commented on in section 3.4. When ϕ is convex, the trajectory minimizes ϕ. The following corollary states this precisely.

COROLLARY 2.1. *Under the assumptions of Theorem* 2.2, *additionally assume that the map* ϕ *is convex. Then*

$$\lim_{t \to +\infty} (\phi(x(t))\,,\, \phi(y(t))) = (\inf \phi\,,\, \inf \phi).$$

Proof of Corollary 2.1. Consider the different cases (i) $x_\infty \in \mathbb{R}$ (hence x is bounded), (ii) $x_\infty = -\infty$ (hence $\phi' \geq 0$), (iii) $x_\infty = +\infty$ (hence $\phi' \leq 0$). ∎

Corollary 2.1 is commented on in section 3.5. As a consequence of Theorem 2.2, we deduce the convergence result of Haraux [16, Example 2.2.6] for the (not controlled) (HBF) system in dimension one.

COROLLARY 2.2 (Haraux [16]). *Let* $f \in \mathcal{C}^1(\mathbb{R})$ *and* $\gamma > 0$*; then every bounded solution of*

$$\ddot{x}(t) + \gamma \dot{x}(t) = f(x(t))$$

converges to some $x_\infty \in \mathbb{R}$ *such that* $f(x_\infty) = 0$.

Proof of Corollary 2.2. In Theorem 2.2, consider the (not controlled) case where $\varepsilon(t) = 0$ for every t, with a bounded $\mathcal{C}^2$ potential ϕ such that $\phi(z) = -\int_0^z f(u)du$ for $|z| \leq \|x\|_\infty$. ∎

2.2.2. Slow control. The results in this section specify the convergence of the solution map (x, y) toward specific points of $\widehat{S}$ with a "slow" control ε. For every $\lambda \in \overline{\mathbb{R}}$, we denote by $P^+(\lambda)$ (resp., $P^-(\lambda)$) the following proposition:

$$\text{For every neighborhood } V(\lambda) \text{ of } \lambda \quad \exists \mu \in V(\lambda) \cap \mathbb{R}, \quad \phi'(\mu) > 0 \quad (\text{resp., } \phi'(\mu) < 0).$$

THEOREM 2.3 (slow parametrization). *Under the assumptions of Theorem* 2.2, *additionally assume that*

$$(\mathcal{H}_\varepsilon)\text{(iii)} \qquad \int_0^{+\infty} \varepsilon(t)\, dt = +\infty.$$

$$(\mathcal{H}_V)\text{(iv)} \qquad \forall z \in \mathbb{R} \setminus \{0\}, \quad V'(z) \neq 0.$$

Let $m_\infty = \min\{x_\infty, y_\infty\}$ *and* $M_\infty = \max\{x_\infty, y_\infty\}$, *where* $(x_\infty, y_\infty) \in \widehat{S} \times \widehat{S}$ *are the limits of the trajectories. Then the solution* (x, y) *of the* (HBFC^2) *system satisfies the following properties:*

- *If* $-\infty < m_\infty < M_\infty < +\infty$, *then we have* $P^-(m_\infty)$ *and* $P^+(M_\infty)$.
- *If* $m_\infty = -\infty$ *and* $M_\infty < +\infty$, *then we have* $P^+(m_\infty)$ *or* $P^+(M_\infty)$.
- *If* $m_\infty > -\infty$ *and* $M_\infty = +\infty$, *then we have* $P^-(m_\infty)$ *or* $P^-(M_\infty)$.

[3] In fact this last case could be deduced from the previous one—the (LIM) assumption—by changing the map ϕ.

Theorem 2.3 is proved in section 5.2. In fact Theorem 2.3 specifies the points of $\widehat{S}$ where the trajectories x and y converge. In order to ensure readability, we consider only the special case below (see Corollary 3.1 for a more extensive description).

COROLLARY 2.3 (slow parametrization). *Under the assumptions of Theorem* 2.3, *let* $I(\lambda)$ *be the connected component of* λ *in* $\widehat{S}$. *If* $y_\infty < x_\infty$ *and if one of the conditions*

(a) $-\infty < y_\infty$;

(b) $y_\infty = -\infty$ *and* $\exists \lambda \in \mathbb{R}$, $\phi'((-\infty, \lambda]) \leq 0$

is satisfied, then the trajectory x *converges to an extremal point of* $I(x_\infty)$; *precisely,*

$$x_\infty \in \{\inf I(x_\infty), \sup I(x_\infty)\}.$$

Corollary 2.3 is proved in section 5.3 and commented on in section 3.6. When the set of equilibria is an interval, the behavior of the trajectories is more precise.

COROLLARY 2.4 (slow parametrization with a connected set of equilibria). *Under the assumptions of Theorem* 2.3, *additionally assume that the set* $\widehat{S}$ *is an interval (in* $\overline{\mathbb{R}}$*) (for example, if* ϕ *is convex). Then the solution (x,y) of the* (HBFC^2) *system satisfies one of the following cases:*

(i) $\lim_{t\to+\infty}(x(t)\,,\, y(t)) = (\sup \widehat{S}\,,\, \inf \widehat{S})$.

(ii) $\lim_{t\to+\infty}(x(t)\,,\, y(t)) = (\inf \widehat{S}\,,\, \sup \widehat{S})$.

(iii) *There exists* $x_\infty \in \widehat{S}$ *such that* $\lim_{t\to+\infty} x(t) = \lim_{t\to+\infty} y(t) = x_\infty$.

Note that Theorem 2.2 and Corollary 2.4 clearly imply Theorem B in the introduction.

Remark. In Corollary 2.4, the three cases reduce to only one when $\widehat{S}$ is reduced to a singleton. But then other simple methods of optimization (like the steepest descent method in the convex case) also apply. In the general case where $\widehat{S}$ is not reduced to a singleton, the three cases are disjoint.

Remark. When case (i) or (ii) of Corollary 2.4 holds, one obtains a global description of the set $\widehat{S}$ since the trajectories converge toward the extremal points of $\widehat{S}$. After some numerical experiments, we conjecture that case (iii) only happens for initial data in a negligible set (see section 6.2).

To illustrate Corollary 2.4 (hence Theorem 2.3 and Corollaries 2.3 and 3.1), we represent some trajectories $(x(t), y(t))$ in the plane $\mathbb{R}^2$. We take $\phi(z) = (z+1)^2$ if $z \leq -1$, $\phi(z) = 0$ if $z \in [-1, 1]$, $\phi(z) = (z-1)^2$ if $z \geq 1$, $V(z) = \frac{1}{2}e^{\frac{-z^2}{10}}$, $\varepsilon(t) = \frac{1}{\log(t+2)}$, and $\gamma = 0.4$. On Figure 2.1, we draw the trajectories $(x(t), y(t))$ for the three different initial conditions $(x(0), y(0), \dot{x}(0), \dot{y}(0)) = (0, 1.9, 0, 0)$, respectively, $(1.5, -2.5, 0, 0)$, respectively, $(2.5, 1.525718, 0, 0)$ corresponding to case (i), respectively, (ii), respectively, (iii). The gray square is the set $S \times S = [-1, 1] \times [-1, 1]$.

As a (clear) consequence of Corollary 2.4, the limit of the difference of the trajectories either maximizes the diameter of the set S or is equal to 0.

COROLLARY 2.5. *Under the assumptions of Theorem* 2.3, *additionally assume that* S *is a nonempty interval and that* $\widehat{S} = \mathrm{cl}_{\overline{\mathbb{R}}}(S)$. *Then the solution* (x, y) *of the* (HBFC^2) *system satisfies*

$$\lim_{t\to+\infty} |x(t) - y(t)| = \mathrm{diam}\,(S) \qquad \textit{or} \qquad \lim_{t\to+\infty} |x(t) - y(t)| = 0.$$

The proof of Corollary 2.5 is immediate.

Remark. Corollary 2.5 may not hold if we do not assume $\widehat{S} = \mathrm{cl}_{\overline{\mathbb{R}}}(S)$, even if S is a nonempty interval. Consider the counterexample in section 3.5.

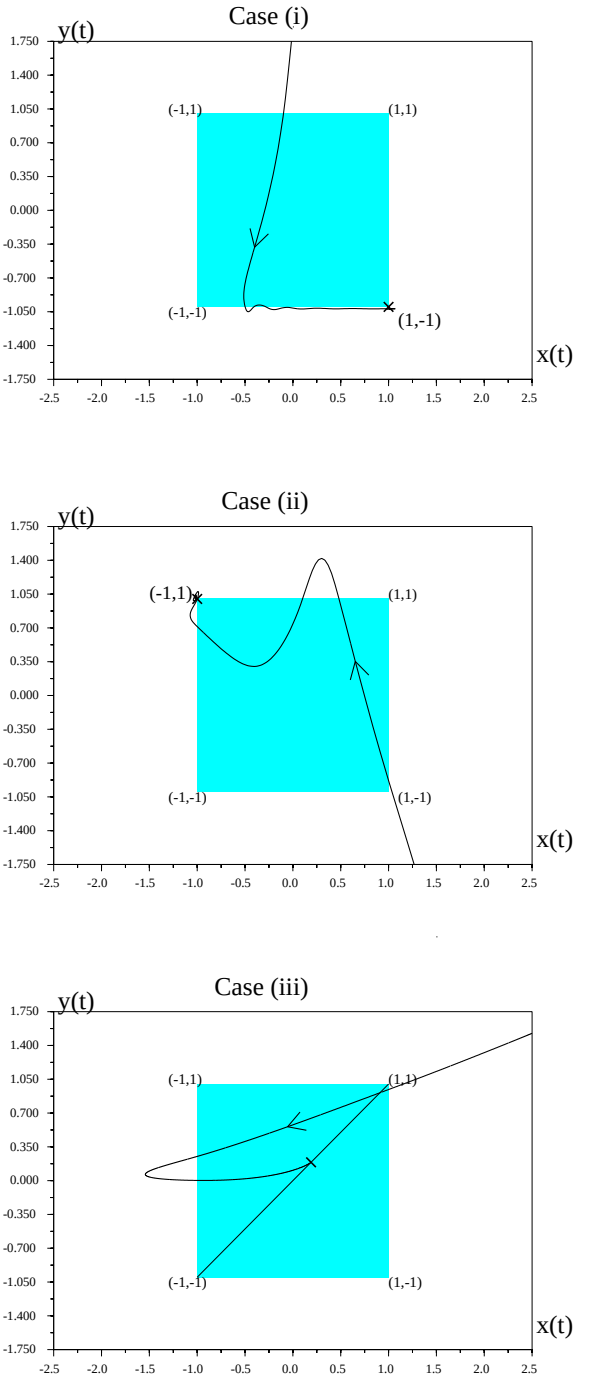


FIG. 2.1. *Illustration of the three cases* (i), (ii), *and* (iii).

2.3. Fast convergence in the convex and infinite-dimensional case. In order to precisely determine the convergence properties of the (HBFC^2) system, we show that the solutions of the (HBFC^2) system weakly converge, with a convex potential and a "fast control." Note that in this section we do not necessarily assume the potential V to be repulsive.

PROPOSITION 2.2 (fast parametrization in the convex case). *Assume Hypothesis* 1, *that* ϕ *is convex with* $S = \operatorname{argmin}\phi \neq \emptyset$, *and that*

$$(\mathcal{H}_\varepsilon)(\text{iv}) \qquad \int_0^{+\infty} \varepsilon(t)\,dt < +\infty.$$

Additionally assume that the trajectories x *and* y *are bounded. Then there exists some* $(x_\infty, y_\infty) \in S \times S$ *such that* (x, y) *weakly converges to* (x_∞, y_∞)*; precisely,* $w - \lim_{t\to+\infty}(x(t), y(t)) = (x_\infty, y_\infty)$ *and* $\lim_{t\to+\infty}\phi(x(t)) = \lim_{t\to+\infty}\phi(y(t)) = \min\phi$.

Proposition 2.2 is proved in section 5.5.

3. Remarks and counterexamples.

3.1. Some applications. In this section, we give some illustrations of practical cost or potential functions with nonunique and nonisolated equilibria. These are of course academic examples, while examples closer to real-world applications would require considering higher dimensions, PDEs, and also constraints (for example, if one needs to describe the full optimal face in mathematical programming). As pointed out in [5], this last study raises nontrivial difficulties, since the existence of constraints implies the possibility of shocks, with $\dot{x}$, $\dot{y}$ being discontinuous and $\ddot{x}$, $\ddot{y}$ being measures in the (HBFC^2) system.

Example. Let us consider the minimization problem

$$(\text{P}) \qquad \min_{x\in\mathbb{R}^n}\{\alpha\,||Ax - b||^2 + \beta\,d_C^2(x)\},$$

where A is an $(m\times n)$ matrix, $b \in \mathbb{R}^m$ is a vector, α and β are nonnegative coefficients, and d_C stands for the distance function of a convex set C, which should be seen as a penalization. When $\beta = 0$, (P) reduces to the classical least-squares resolution of the equation $Ax = b$. The optimal set S is then an affine subspace $F \subset \mathbb{R}^n$. On the other hand, when $\alpha = 0$, the optimal set clearly coincides with the set C. Now suppose we mix both problems by taking $\alpha > 0$ and $\beta > 0$. Assuming moreover that $C \cap F \neq \emptyset$, the optimal set S is equal to $C \cap F$: our problem then amounts to solving the linear system $Ax = b$ by a least-squares technique under the constraint $[x \in C]$. One question which naturally arises here concerns the size of the set $C \cap F$. We can then use the (HBFC)^2 system with $\phi(x) = \alpha\,||Ax - b||^2 + \beta\,d_C^2(x)$ to obtain a global exploration of the set $C \cap F$, and possibly its diameter. A second important question concerns the full description of the affine space F. This can be achieved by using N coupled system (for N larger than $\dim F$) and taking $C = \overline{B}(0, M)$ (with M large enough). If they do not collapse, the limit points of the trajectories will generate (by affine combination) the space F.

Example. An investor who wants to set up a firm will face and minimize an aggregate cost function, which is likely to have nonisolated equilibria. For example, the cost of linking to various utilities (water, electricity, etc.) is usually constant in a city, while outside the city the cost depends on the distance to the existing network. The same applies for employment costs, due to the existence of public transportation, but also due to possible differences of wage policies in different states.

3.2. On the global existence results.

Remark. Proposition 2.1 holds without assuming that the map ∇V is bounded on the bounded subsets of H ($(\mathcal{H}_V)$(ii)) and that $\lim_{t\to+\infty} \varepsilon(t) = 0$ ($(\mathcal{H}_\varepsilon)$(ii)).

Remark. If one does not assume that the trajectory (x, y) is bounded, then assertion (iv) of Theorem 2.1 may not hold, even in dimension one. Consider indeed a C^2 potential $\phi : \mathbb{R} \to \mathbb{R}$ such that $\phi' > 0$, $\lim_{z\to-\infty} \phi(z) = -1$, $\lim_{z\to+\infty} \phi(z) = 1$, $\limsup_{z\to-\infty} \phi'(z) \geq 1$. Take as initial conditions $x_0 = y_0$ and $\dot{x}_0 = \dot{y}_0 = 0$. Then by the Cauchy–Lipschitz theorem, $x \equiv y$. Then x satisfies $\ddot{x}(t) + \gamma\dot{x}(t) + \phi'(x(t)) = 0$. Since $\phi' > 0$, then $\ddot{x} + \gamma\dot{x} < 0$. We deduce that the map $\dot{x} + \gamma x$ converges (in $\overline{\mathbb{R}}$). Since $\dot{x}$ is bounded, $\ddot{x}$ is bounded from above. Since $\dot{x} \in L^2([0, +\infty), \mathbb{R})$, $\lim_{t\to+\infty} \dot{x}(t) = 0$, and the trajectory x converges (in $\overline{\mathbb{R}}$). Then $\liminf_{t\to+\infty} \phi'(x(t)) \leq 0$ (otherwise we would have $\lim_{t\to+\infty} \dot{x}(t) = -\infty$), and we obtain $\lim_{t\to+\infty} x(t) = -\infty$; hence $\limsup_{t\to+\infty} \phi'(x(t)) \geq 1$, a contradiction with assertion (iv).

Remark. Proposition 2.1 and Theorem 2.1 can easily be generalized to the case of N coupled oscillators with different potentials $\phi_1, \ldots, \phi_N$, with $N \geq 2$. For example, taking $\phi_2 = -\phi_1$ should lead to the exploration of saddle points, but this study is beyond the scope of this paper.

3.3. On condition (LIM).

Remark. When $H = \mathbb{R}$, note that the limit condition (LIM) is satisfied if

$$(\mathrm{LIM}_{strong}) \qquad \begin{cases} \lim_{z\to-\infty} \phi(z) = +\infty \text{ or } \lim_{z\to-\infty} \phi'(z) = 0, \\ \lim_{z\to+\infty} \phi(z) = +\infty \text{ or } \lim_{z\to+\infty} \phi'(z) = 0. \end{cases}$$

The converse is not true. Consider, for example, a $\mathcal{C}^2$ convolution-type approximation ϕ of the function $f + g$, where f is the simplest continuous piecewise affine function such that $f(2^{2n}) = n$ and $f(2^{2n+1}) = 0$, and g is the simplest continuous piecewise affine function such that $g(2^{2n}) = 1$, $g([2^{2n} + 1/2^n, 2^{2(n+1)} - 1/2^{n+1}]) = 0$. Note also that the natural generalization of (LIM_{strong}) in higher dimensions would be

$$\forall z \in H \setminus \{0\}, \qquad \lim_{\lambda\to+\infty} \phi(\lambda z) = +\infty \qquad \text{or} \qquad \lim_{\lambda\to+\infty} \nabla\phi(\lambda z) = 0.$$

But contrary to the one-dimensional case, the above condition does not imply (LIM). For example, consider $H = \mathbb{R}^2$ and $\phi(x, y) = x^2$.

Remark. Condition (LIM) can be generalized as follows:

$$(\mathrm{LIM}_{weak}) \qquad \begin{array}{l} \forall \alpha > 0, \forall A \in \mathbb{R}, \text{ for every connected component} \\ \Gamma_A \text{ of the set } \{z \in H, \phi(z) \leq A\}, \text{ the set} \\ \{z \in \Gamma_A, |\nabla\phi(z)| \geq \alpha\} \text{ is bounded.} \end{array}$$

Then all the results in this paper hold by replacing (LIM) by (LIM_{weak}). When $H = \mathbb{R}$, (LIM_{weak}) is equivalent to

$$\begin{cases} (\exists (z_n) \subset \mathbb{R}, \lim_{n\to+\infty} z_n = -\infty \text{ and } \lim_{n\to+\infty} \phi(z_n) = +\infty) \text{ or } \lim_{z\to-\infty} \phi'(z) = 0, \\ (\exists (z_n) \subset \mathbb{R}, \lim_{n\to+\infty} z_n = +\infty \text{ and } \lim_{n\to+\infty} \phi(z_n) = +\infty) \text{ or } \lim_{z\to+\infty} \phi'(z) = 0. \end{cases}$$

3.4. On Theorem 2.2.

Remark. It is possible to obtain the convergence of the solutions (but not necessarily in $\widehat{S}$; see below) under other sets of assumptions (for example, $\phi(z)$ converges when $z \to -\infty$, and when $z \to +\infty$). But our concern being the exploration of the equilibria of ϕ by the (HBFC^2) system, the extensive study of the converging properties of (HBFC^2) is beyond the scope of this paper.

3.5. On Corollary 2.1.

Remark. Corollary 2.1 may not hold if the map ϕ is not assumed to be convex, even if it is quasi-convex. Consider indeed the map $\phi : \mathbb{R} \to \mathbb{R}$ defined by $\phi(z) = 1 - e^{-z^2}$, a C^1 repulsion V such that $V(z) = 1/z$ if $z \geq 1$. Then ϕ is quasi-convex and $(x(t)\,,\, y(t)) = (\log t, -\log t)$ (for t large enough) is a solution of the corresponding (HBFC^2) system with $\varepsilon(t) = (\log t)^2(\frac{1}{t} - \frac{1}{t^2} + 2e^{-(\log t)^2}\log t)$ and $\gamma = 1$. One easily finds that when $t \to +\infty$, $\varepsilon(t) \sim \frac{(\log t)^2}{t}$ and $\dot{\varepsilon}(t) \sim \frac{\log t}{t^2}(2 - \log t)$, and hence that $\varepsilon(t) > 0$ and $\dot{\varepsilon}(t) < 0$ for t large enough. Moreover, note that $\int_0^{+\infty} \varepsilon(t)\, dt = +\infty$. The solution map $(x(t)\,,\, y(t)) = (\log t, -\log t)$ clearly does not satisfy the conclusion of Corollary 2.1.

3.6. On Corollary 2.3.

Remark. The conclusion of Corollary 2.3 may not be more precise in general. Precisely, one cannot replace the conclusion $x_\infty \in \{\inf I(x_\infty), \sup I(x_\infty)\}$ by $x_\infty = \sup I(x_\infty)$. Indeed, consider the even $\mathcal{C}^1$ map $\phi : \mathbb{R} \to \mathbb{R}$ defined by $\phi(z) = z^2 - \frac{1}{2}$ for $0 \leq z \leq \frac{1}{2}$, $\phi(z) = -(z-1)^2$ for $\frac{1}{2} \leq z \leq 1$, $\phi(z) = 0$ for $z \geq 1$. Let $x(t) = 1 - \frac{1}{t}$ for $t \geq 2$ and $y(t) = -x(t)$. Then (x, y) is a solution of the corresponding (HBFC^2) system, with $V(z) = \frac{1}{z}$ for $z \geq 1$ and $\varepsilon(t) = (2 - \frac{2}{t})^2(\frac{2}{t} + \frac{1}{t^2} - \frac{2}{t^3})$. Note that when $t \to +\infty$, $\varepsilon(t) \sim \frac{8}{t}$ and $\dot{\varepsilon}(t) \sim -\frac{8}{t^2}$, and hence $\int_0^{+\infty} \varepsilon(t)\, dt = +\infty$ and $\dot{\varepsilon}(t) < 0$ for t large enough.

Remark. The conclusion of Corollary 2.3 does not hold if $y_\infty = -\infty$ and there exists no $\lambda \in \mathbb{R}$ such that $\phi'((-\infty, \lambda]) \leq 0$. Indeed, consider a $\mathcal{C}^1$ increasing map $\phi : \mathbb{R} \to \mathbb{R}$ such that $\phi(z) = e^z$ for $z \leq 0$, $\phi(z) = 2$ for $z \geq 1$, and such that $S = [1, +\infty)$. Let $\gamma = 1$, $V(z) = e^{-z}$ for $z \geq 1$. Let $y(t) = -(1 + \frac{2}{t})\log(t)$ for $t \geq 1$. Then there is a map $x : [1, +\infty) \to \mathbb{R}$ such that $x(t) = 2 - \frac{1}{t} + o(\frac{1}{t})$ and such that (x, y) is a solution of the corresponding (HBFC^2) system for t large enough, with a control ε which satisfies $\varepsilon(t) \sim \frac{1}{C_2 t}$ and $\dot{\varepsilon}(t) \sim \frac{-1}{C_2 t^2}$ for some $C_2 > 0$. Precisely, $x(t) = x(1) + \int_1^t e^{-s} \int_1^s -e^u(\ddot{y}(u) + \dot{y}(u) + e^{y(u)})du ds$ and $\varepsilon(t) = \frac{\ddot{y}(t) + \dot{y}(t) + e^{y(t)}}{V'(x(t) - y(t))}$. Then $\widehat{S} = \{-\infty\} \cup [1, +\infty]$ and $\lim_{t \to +\infty} x(t) = 2$, which is clearly not an extremal point of its connected component $[1, +\infty]$.

Remark. From Corollary 2.3, one can deduce a full description of the different cases, as seen more precisely in the following corollary.

COROLLARY 3.1 (slow parametrization). *Under the assumptions of Theorem* 2.3*, the solution (x,y) of the* (HBFC^2) *system satisfies one of the following cases:*

(i) $y_\infty < x_\infty$
- (a) *if* $-\infty < y_\infty$, *or* ($y_\infty = -\infty$ *and* $\exists \lambda \in \mathbb{R}$, $\phi'((-\infty, \lambda]) \leq 0$), *then* $x_\infty \in \{\inf I(x_\infty), \sup I(x_\infty)\}$;
- (b) *if* $x_\infty < +\infty$, *or* ($x_\infty = +\infty$ *and* $\exists \lambda \in \mathbb{R}$, $\phi'([\lambda, +\infty)) \geq 0$), *then* $y_\infty \in \{\inf I(y_\infty), \sup I(y_\infty)\}$.

(ii) $x_\infty < y_\infty$
- (a) *if* $-\infty < x_\infty$, *or* ($x_\infty = -\infty$ *and* $\exists \lambda \in \mathbb{R}$, $\phi'((-\infty, \lambda]) \leq 0$), *then* $y_\infty \in \{\inf I(y_\infty), \sup I(y_\infty)\}$;
- (b) *if* $y_\infty < +\infty$, *or* ($y_\infty = +\infty$ *and* $\exists \lambda \in \mathbb{R}$, $\phi'([\lambda, +\infty)) \geq 0$), *then* $x_\infty \in \{\inf I(x_\infty), \sup I(x_\infty)\}$.

(iii) $x_\infty = y_\infty$.

Proof of Corollary 3.1. Let $\tilde{\phi}(z) = \phi(-z)$ and $\tilde{V}(z) = V(-z)$. If (x, y) is a solution of the (HBFC^2) system associated to ϕ and V, $(-y, -x)$ is a solution of the (HBFC^2) system associated to $\tilde{\phi}$ and V, which proves (ii)(b), and (y, x) is a solution

of the (HBFC^2) system associated to ϕ and $\tilde{V}$, which proves (ii)(a). Assertion (i)(b) follows. ∎

3.7. On Corollary 2.5.

Remark. The formulation of Theorem 2.3 and Corollaries 2.3, 2.4, and 3.1 is specific to dimension one. This is not the case for Corollary 2.5, even if this last one may not remain true in higher dimensions. Indeed, in $\mathbb{R}^2$, consider the set $C = [-1,1] \times [-1,1]$, the function $\phi = d_C^2$, and $V(z) = e^{-|z|^2}$. Take any $\gamma > 0$ and slow control function ε (satisfying the assumptions $(\mathcal{H}_\varepsilon)$(i)(ii)(iii)). Let (x,y) be the solution of the corresponding (HBFC^2) system for initial data $x(0) \in \{0\} \times \mathbb{R}$, $y(0) \in \{0\} \times \mathbb{R}$, $\dot{x}(0) \in \{0\} \times \mathbb{R}$, $\dot{y}(0) \in \{0\} \times \mathbb{R}$. It is then clear that the trajectories of x and y are one-dimensional, and from Corollary 2.4 one easily shows that either $\lim_{t\to+\infty} x(t) = \lim_{t\to+\infty} y(t)$ or x and y converge toward the extremities of the segment $[(0,-1),(0,1)]$. In this last case, $\lim_{t\to+\infty} |x(t)-y(t)| = 2 \neq 2\sqrt{2} = \operatorname{diam}(S)$.

3.8. On Proposition 2.2.

Remark. In general, it does not seem possible to obtain a better result (in the sense of the specification of the limit point, or by obtaining strong convergence rather than weak convergence) without further assumptions. Indeed, if $x = y$, a case which happens if and only if $(x_0, \dot{x}_0) = (y_0, \dot{y}_0)$, then the (HBFC^2) system reduces to a (HBF) system. It is then known (see Baillon [7] and [6]) that the trajectory (x,y) may not strongly converge and that the weak limits depend on the initial data.

Remark. If the map ϕ is not assumed to be convex, the trajectories (x,y) may not weakly converge. Considering again the case where $x = y$, we refer to the counterexample of Redont [23] (see also [6]).

4. Proof of the global existence and of the main properties.

4.1. Proof of Proposition 2.1. Let

$$X = (x,y) \in H^2, \qquad \Phi(X) = \phi(x) + \phi(y), \qquad U(X) = V(x-y),$$
$$X_0 = (x_0, y_0), \qquad \dot{X}_0 = (\dot{x}_0, \dot{y}_0).$$

The system (HBFC^2) then reduces to:

$$\text{(HBFC)} \qquad \begin{cases} \ddot{X} + \gamma \dot{X} + \nabla\Phi(X) + \varepsilon(t)\nabla U(X) = 0, \\ X(0) = X_0, \quad \dot{X}(0) = \dot{X}_0. \end{cases}$$

4.1.1. Proof of (i). The second order system (HBFC) can be written as a first order system in $H^2 \times H^2$:

$$\dot{Y} = F(t,Y)$$

with

$$Y(t) = \begin{pmatrix} X(t) \\ \dot{X}(t) \end{pmatrix} \qquad \text{and} \qquad F(t,u,v) = \begin{pmatrix} v \\ -\gamma v - \nabla\Phi(u) - \varepsilon(t)\nabla U(u) \end{pmatrix}. \tag{4.1}$$

For $Y_0 = \binom{X_0}{\dot{X}_0}$ given in $H^2 \times H^2$, the Cauchy–Lipschitz theorem and Hypothesis 1 ensure the existence of a unique local solution to the problem

$$\begin{cases} \dot{Y} = F(t,Y), \\ Y(0) = Y_0. \end{cases} \tag{4.2}$$

Let X denote the maximal solution defined on the interval $[0, T_{max})$ with $0 < T_{max} \leq +\infty$. In order to prove that $T_{max} = +\infty$, let us show that the map $\dot{X}$ is bounded. We first observe that (HBFC) and the regularity assumptions on ϕ, V, and ε automatically imply that the map X is $\mathcal{C}^2$ on $[0, T_{max})$.

Along every trajectory of (HBFC), we define the energy by

$$E(t) = \frac{1}{2}|\dot{X}(t)|^2 + \Phi(X(t)) + \varepsilon(t)U(X(t)).$$

By differentiation of $E(t)$, and in view of (HBFC), we obtain for every $t \in [0, T_{max})$

$$\begin{aligned} \dot{E}(t) &= \langle \dot{X}(t), \ddot{X}(t) + \nabla\Phi(X(t)) + \varepsilon(t)\nabla U(X(t))\rangle + \dot{\varepsilon}(t)U(X(t)) \\ &= -\gamma|\dot{X}(t)|^2 + \dot{\varepsilon}(t)U(X(t)). \end{aligned} \tag{4.3}$$

Since $\dot{\varepsilon}(t) \leq 0$ (assumption $(\mathcal{H}_\varepsilon)$(i)) and $U \geq 0$, we have $\dot{E}(t) \leq 0$. Hence the function E is nonincreasing and for every $t \in [0, T_{max})$, $E(t) \leq E(0)$. Equivalently,

$$\frac{1}{2}|\dot{X}(t)|^2 + \Phi(X(t)) + \varepsilon(t)U(X(t)) \leq E(0). \tag{4.4}$$

Since Φ is bounded from below, $U \geq 0$, and $\varepsilon \geq 0$, we obtain

$$\sup_{t\in[0,T_{max})} |\dot{X}(t)| < +\infty.$$

By a standard argument, we derive that $T_{max} = +\infty$. Indeed, assume that $T_{max} < +\infty$. Since

$$|X(t) - X(t')| \leq ||\dot{X}||_\infty |t - t'|$$

and since $T_{max} < +\infty$, $\lim_{t\to T_{max}} X(t) := X_\infty$ exists. Hence, the maps X and $\dot{X}$ are bounded on $[0, T_{max})$. Since $\lim_{t\to T_{max}} X(t) = X_\infty$, the map

$$\nabla\Phi(X) + \varepsilon\nabla U(X) \quad \text{ is bounded on } [0, T_{max}).$$

From (HBFC), we deduce that $\ddot{X}$ is bounded on this interval. Hence $\lim_{t\to T_{max}} \dot{X}(t) = \dot{X}_\infty$ exists. Applying again the local existence theorem with initial data $(X_\infty, \dot{X}_\infty) \in H^2 \times H^2$, we can extend the maximal solution to a strictly larger interval, a contradiction. Hence $T_{max} = +\infty$, which completes the proof of Proposition 2.1(i). □

4.1.2. Proof of (ii). We already proved that the function E is nonincreasing. For every $t \geq 0$, $E(0) \geq E(t) \geq \Phi(X(t))$, which implies that the map Φ is bounded from above, and hence bounded. Moreover, the function E is also bounded from below. Hence, there exists $E_\infty \in \mathbb{R}$ such that $\lim_{t\to+\infty} E(t) = E_\infty$. From (4.4), and since Φ is bounded from below and $U \geq 0$, we obtain, for every $t \geq 0$,

$$\frac{1}{2}|\dot{X}(t)|^2 \leq E(0) - \inf\Phi.$$

Hence $\dot{X} \in L^\infty([0, +\infty); H^2)$. From (4.3) and the fact that $\dot{\varepsilon} \leq 0$ and $U \geq 0$, we derive, for every $t \geq 0$,

$$\int_0^t |\dot{X}(s)|^2 ds \leq \frac{1}{\gamma}(E(0) - E(t)).$$

Since $E(t)$ decreases to E_∞ as t increases to $+\infty$, we obtain

$$\int_0^{+\infty} |\dot{X}(s)|^2 ds \leq \frac{1}{\gamma}(E(0) - E_\infty),$$

and $\dot{X} \in L^2([0,+\infty); H^2)$. The inequality (4.4) implies, for every $t \geq 0$,

$$\Phi(X(t)) \leq E(0),$$

i.e., $\Phi(X)$ is bounded. □

4.2. Proof of Theorem 2.1.

4.2.1. Proof of (iii). We first claim that the maps $\nabla\Phi(X)$ and $U(X)$ are bounded. To prove it, we distinguish the two cases (a) and (b).

Case (a). Since the map $\nabla\Phi$ is Lipschitz continuous on the bounded sets, it is bounded on the bounded sets. Since the map ∇U is bounded on the bounded sets, the map U is Lipschitz continuous on the bounded sets, and hence bounded on the bounded sets. Since the trajectory $X = (x, y)$ is bounded, the maps $\nabla\Phi(X)$ and $U(X)$ are bounded.

Case (b). In this case, the map $U(X)$ is assumed to be bounded. From Proposition 2.1(ii), the map $\phi(x)$ is bounded. The condition (LIM) implies that the set

$$C = \{z \in H, \phi(z) \leq \|\phi(x)\|_\infty \text{ and } |\nabla\phi(z)| \geq 1\}$$

is bounded. Since the map $\nabla\phi$ is bounded on the bounded sets, it is bounded on C. If $x(t) \notin C$, since $\phi(x(t)) \leq \|\phi(x)\|_\infty$, we deduce $|\nabla\phi(x(t))| < 1$. Hence the map $\nabla\phi(x)$ is bounded, and so is the map $\nabla\phi(y)$. We now conclude the proof thanks to the following claim. □

CLAIM 4.1. *If the maps $\nabla\Phi(X)$ and $U(X)$ are bounded, then $\lim_{t\to+\infty} \dot{X}(t) = 0$.*

Proof of Claim 4.1. Since the map $\nabla\Phi(X)$ is bounded, and since the map $\dot{X}$ is bounded (Proposition 2.1(ii)), the map $t \mapsto \Phi(X(t))$ is Lipschitz continuous. Let us recall that the energy function

$$E(t) = \frac{1}{2}|\dot{X}(t)|^2 + \Phi(X(t)) + \varepsilon(t)U(X(t))$$

converges to some $E_\infty \in \mathbb{R}$. Since $t \mapsto U(X(t))$ is bounded, $\lim_{t\to+\infty} \varepsilon(t)U(X(t)) = 0$. Hence, the map $|\dot{X}|^2$ is the sum of a convergent map and of a Lipschitz continuous map. From Proposition 2.1(ii), $|\dot{X}|^2 \in L^1([0,+\infty), H)$. By a classical argument, we deduce that $\lim_{t\to+\infty} |\dot{X}(t)|^2 = 0$, and hence deduce also that the map $\Phi(X(t))$ converges when $t \to +\infty$. □

4.2.2. Proof of (iv). We distinguish the two cases (a) and (b).

Proof in Case (a). The trajectory X is bounded, and $\lim_{t\to+\infty} \nabla\Phi(X(t)) = 0$—which proves (iv)—is an immediate consequence of the following lemma. □

LEMMA 4.1. *Assume Hypothesis* 1. *Let $(t_n) \subset \mathbb{R}_+$ be a sequence such that $\lim_{n\to+\infty} t_n = +\infty$ and $X(t_n)$ is bounded. Then $\lim_{n\to+\infty} \nabla\Phi(X(t_n)) = 0$.*

Proof of Lemma 4.1. Assume that it is not true. Then there exist $\alpha > 0$ and a subsequence of (t_n), still denoted by (t_n), such that $\lim_{n\to+\infty} t_n = +\infty$ and $|\nabla\Phi(X(t_n))| \geq \alpha$ for every n. By assumption, the map $\nabla\Phi$ is K-Lipschitz continuous on the bounded set

$$C = \{Z \in H \times H, \quad \exists n \in \mathbb{N}, \quad |Z - X(t_n)| \leq 1\}$$

for some $K > 0$. Let $\tau = \frac{\alpha}{K\|\dot{X}\|_\infty}$ (assuming without any loss of generality that $\|\dot{X}\|_\infty \neq 0$) and let $t \in [t_n, t_n + \tau]$. Then

$$|X(t) - X(t_n)| \leq \|\dot{X}\|_\infty \tau = \frac{\alpha}{K}. \tag{4.5}$$

Assuming without any loss of generality that $\frac{\alpha}{K} \leq 1$, then from (4.5), $X(t) \in C$ and

$$|\nabla\Phi(X(t)) - \nabla\Phi(X(t_n))| \leq K|X(t) - X(t_n)| \leq K\frac{\alpha}{K} = \alpha. \tag{4.6}$$

Let us now integrate (HBFC) on the interval $[t_n, t_n + \tau]$:

$$\begin{aligned}\dot{X}(t_n + \tau) - \dot{X}(t_n) + \gamma \int_{t_n}^{t_n+\tau} \dot{X}(t)\, dt + \int_{t_n}^{t_n+\tau} \nabla\Phi(X(t))\, dt \\ + \int_{t_n}^{t_n+\tau} \varepsilon(t)\nabla U(X(t))\, dt = 0.\end{aligned} \tag{4.7}$$

From (iii), $\lim_{n\to+\infty} \sup_{t\in[t_n,t_n+\tau]} \dot{X}(t) = 0$ and therefore

$$\lim_{n\to+\infty} \dot{X}(t_n + \tau) - \dot{X}(t_n) + \gamma \int_{t_n}^{t_n+\tau} \dot{X}(t)\, dt = 0.$$

Since the map ∇U is bounded on the bounded set C, we obtain

$$\lim_{n\to+\infty} \int_{t_n}^{t_n+\tau} \varepsilon(t)\nabla U(X(t))\, dt = 0.$$

Noticing that

$$\begin{aligned}\left| \int_{t_n}^{t_n+\tau} \nabla\Phi(X(t))\, dt - \int_{t_n}^{t_n+\tau} \nabla\Phi(X(t_n))\, dt \right| &\leq K\|\dot{X}\|_\infty \int_{t_n}^{t_n+\tau} |t - t_n|\, dt \\ &\leq \frac{K\|\dot{X}\|_\infty \tau^2}{2} = \frac{\alpha\tau}{2}\end{aligned}$$

and that $\left| \int_{t_n}^{t_n+\tau} \nabla\Phi(X(t_n))\, dt \right| \geq \alpha\tau$, we obtain

$$\left| \int_{t_n}^{t_n+\tau} \nabla\Phi(X(t))\, dt \right| \geq \frac{\alpha\tau}{2}.$$

Taking the limit in equation (4.7) when $n \to +\infty$, we obtain a contradiction. Hence $\lim_{t\to+\infty} |\nabla\Phi(X(t))| = 0$. □

Proof in Case (b). Let us argue by contradiction and assume that it is not true; i.e., there exists $\alpha > 0$ and a sequence $(t_n) \subset \mathbb{R}_+$ such that $\lim_{n\to+\infty} t_n = +\infty$ and $|\nabla\Phi(X(t_n))| \geq \alpha$. Without any loss of generality, we may assume that $|\nabla\phi(x(t_n))| \geq \alpha$. Since the map $\phi(x)$ is bounded (Proposition 2.1(ii)) and since the map ϕ satisfies the condition (LIM), the sequence $(x(t_n))$ is bounded. The following lemma shows that the sequence $(y(t_n))$ is also bounded. From Lemma 4.1, $\lim_{n\to+\infty} \nabla\Phi(X(t_n)) = 0$, a contradiction. Hence $\lim_{t\to+\infty} \nabla\Phi(X(t)) = 0$. □

LEMMA 4.2. *Assume Hypothesis* 1, *and that the map ϕ satisfies the limit condition* (LIM). *Let $\alpha > 0$ and $(t_n) \subset \mathbb{R}_+$ be a sequence such that $\lim_{n\to+\infty} t_n = +\infty$, $(x(t_n))$ is bounded and $|\nabla\phi(x(t_n))| \geq \alpha$. Then the sequence $(y(t_n))$ is also bounded.*

Proof of Lemma 4.2. By assumption, the map $\nabla\phi$ is K-Lipschitz continuous on the bounded set $C = \{z \in H, \quad \exists n \in \mathbb{N}, \quad |z - x(t_n)| \leq 1\}$ for some $K > 0$. Let $\tau = \frac{\alpha}{2K\|\dot{x}\|_\infty}$ (assuming without any loss of generality that $\|\dot{x}\|_\infty \neq 0$) and let $t \in [t_n, t_n + \tau]$. Then

$$|x(t) - x(t_n)| \leq \|\dot{x}\|_\infty \tau = \frac{\alpha}{2K}. \tag{4.8}$$

Assuming without any loss of generality that $\frac{\alpha}{2K} \leq 1$, $x(t) \in C$, and

$$|\nabla\phi(x(t)) - \nabla\phi(x(t_n))| \leq K|x(t) - x(t_n)| \leq K\frac{\alpha}{2K} = \frac{\alpha}{2}. \tag{4.9}$$

By adding the two equations of (HBFC^2), we obtain

$$\ddot{x}(t) + \ddot{y}(t) + \gamma(\dot{x}(t) + \dot{y}(t)) + \nabla\phi(x(t)) + \nabla\phi(y(t)) = 0. \tag{4.10}$$

Integrating (4.10) between t_n and $t_n + \tau$, we obtain

$$\begin{aligned}\dot{x}(t_n + \tau) + \dot{y}(t_n + \tau) - (\dot{x}(t_n) + \dot{y}(t_n)) + \gamma \int_{t_n}^{t_n+\tau} \dot{x}(t) + \dot{y}(t)\, dt \\ + \int_{t_n}^{t_n+\tau} \nabla\phi(x(t)) + \nabla\phi(y(t))\, dt = 0.\end{aligned} \tag{4.11}$$

From assertion (iii), $\lim_{t\to+\infty} \dot{x}(t) + \dot{y}(t) = 0$; hence $\lim_{n\to+\infty} \sup_{t\in[t_n,t_n+\tau]} \dot{x}(t) + \dot{y}(t) = 0$. Therefore,

$$\lim_{n\to+\infty} \dot{x}(t_n + \tau) + \dot{y}(t_n + \tau) - (\dot{x}(t_n) + \dot{y}(t_n)) + \gamma \int_{t_n}^{t_n+\tau} \dot{x}(t) + \dot{y}(t)\, dt = 0$$

and

$$\lim_{n\to+\infty} \int_{t_n}^{t_n+\tau} \nabla\phi(x(t)) + \nabla\phi(y(t))\, dt = 0. \tag{4.12}$$

Writing

$$\nabla\phi(x(t_n)) = \nabla\phi(x(t_n)) - \nabla\phi(x(t)) \; + \; \nabla\phi(x(t)) + \nabla\phi(y(t)) \; - \; \nabla\phi(y(t)),$$

we obtain

$$\begin{aligned}\left|\int_{t_n}^{t_n+\tau} \nabla\phi(x(t_n))\, dt\right| \leq \left|\int_{t_n}^{t_n+\tau} \nabla\phi(x(t_n)) - \nabla\phi(x(t))\, dt\right| \\ + \left|\int_{t_n}^{t_n+\tau} \nabla\phi(x(t)) + \nabla\phi(y(t))\, dt\right| + \left|\int_{t_n}^{t_n+\tau} \nabla\phi(y(t))\, dt\right|.\end{aligned}$$

Now using $|\nabla\phi(x(t_n))| \geq \alpha$, (4.9), and (4.12), we deduce

$$\liminf_{n\to+\infty} \left|\int_{t_n}^{t_n+\tau} \nabla\phi(y(t))\, dt\right| \geq \frac{\alpha}{2}\tau,$$

and therefore, for n large enough,

$$\int_{t_n}^{t_n+\tau} |\nabla\phi(y(t))|\, dt \geq \frac{\alpha}{3}\tau. \tag{4.13}$$

From the above inequality, we deduce that there exists $\theta_n \in [t_n, t_n+\tau]$ such that, for n large enough, $|\nabla\phi(y(\theta_n))| \geq \frac{\alpha}{3}$. Indeed, if it were not true, up to a subsequence, then for every n and for every $t \in [t_n, t_n+\tau]$, $|\nabla\phi(y(t))| < \frac{\alpha}{3}$, which contradicts (4.13). Since the map $\phi(y)$ is bounded (Proposition 2.1(ii)) and since the map ϕ satisfies the condition (LIM), the sequence $(y(\theta_n))$ is bounded. Since $|y(t_n) - y(\theta_n)| \leq ||\dot{y}||_\infty \tau$, we deduce that the sequence $(y(t_n))$ is also bounded. □

4.2.3. Proof of (v) and (vi). Let us write the classical convexity inequality

$$\forall \xi \in H, \qquad \phi(\xi) \geq \phi(x(t)) + \langle \nabla\phi(x(t)), \xi - x(t)\rangle.$$

By noticing that in the duality bracket $\langle \nabla\phi(x(t)), \xi - x(t)\rangle$ the two terms are, respectively, norms converging to zero and bounded, we can pass to the upper limit to obtain

$$\forall \xi \in H, \qquad \phi(\xi) \geq \limsup_{t\to+\infty} \phi(x(t)) \geq \liminf_{t\to+\infty} \phi(x(t)) \geq \inf \phi.$$

This being true for any $\xi \in H$, we deduce that $\lim_{t\to+\infty} \phi(x(t)) = \inf \phi$, which proves assertion (v).

From assertion (iv), $\nabla\phi(x(t)) \to 0$ as $t \to +\infty$. If $x(t_n) \rightharpoonup x_\infty$ weakly, by using the graph closedness property of the maximal monotone operator $\nabla\phi$ in $w-H \times s-H$, we conclude that $\nabla\phi(x_\infty) = 0$, i.e., $x_\infty \in S$. □

5. Proof of the convergence results. In this section, we consider the one-dimensional case, i.e., $H = \mathbb{R}$.

5.1. Proof of Theorem 2.2. The proof of the convergence of the trajectories goes in three steps. We first establish the inequality fulfilled by the map $\max\{x, y\}$. Then we deduce the convergence of the maps $\max\{x, y\}$ and $\min\{x, y\}$. Finally we check the convergence of the trajectories.

5.1.1. Properties of the map $\max\{x, y\}$.

LEMMA 5.1. *Under the assumptions of Theorem* 2.2, *let* (x, y) *be the unique solution of the* (HBFC^2) *system. Then*

(i) *there exists a discrete and closed set* $\mathcal{D} \subset \mathbb{R}_+$ *such that the map* $w = \max\{x, y\}$ *is of class* $\mathcal{C}^2$ *on* $\mathbb{R}_+ \setminus \mathcal{D}$, *and for every* $t \in \mathbb{R}_+ \setminus \mathcal{D}$,

$$\ddot{w}(t) + \gamma\dot{w}(t) + \phi'(w(t)) \geq 0; \tag{5.1}$$

(ii) *for every* $(t, t') \in \mathbb{R}_+ \times \mathbb{R}_+$ *such that* $t' \leq t$, *the following inequality holds:*

$$w(t) \geq w(t') - \frac{1}{\gamma}(\dot{w}_-(t) - \dot{w}_+(t')) - \frac{1}{\gamma}\int_{t'}^{t} \phi'(w(s))\, ds.$$

Proof of Lemma 5.1.

Proof of (i). Let us remark at once that if $\forall t \in \mathbb{R}_+$, $x(t) = y(t)$, then we obviously have

$$\forall t \in \mathbb{R}_+, \qquad \ddot{w}(t) + \gamma\dot{w}(t) + \phi'(w(t)) = 0.$$

We can then take $\mathcal{D} = \emptyset$. Assume now that $x \not\equiv y$. Let us prove that the set $\mathcal{D} = \{t \in \mathbb{R}_+,\ x(t) = y(t)\}$ satisfies the conditions of the statement. Since x and y are continuous, the set $\mathcal{D}$ is closed. Let $t_0 \in \mathcal{D}$, i.e., such that $x(t_0) = y(t_0)$.

Necessarily $\dot{x}(t_0) \neq \dot{y}(t_0)$; in fact, in the opposite case, one would have $x \equiv y$ by the Cauchy–Lipschitz theorem. Indeed, take the solution z of $\ddot{z}(t) + \gamma\dot{z}(t) + \phi'(z(t)) = 0$, $z(t_0) = x(t_0)$ $\dot{z}(t_0) = \dot{x}(t_0)$ and note that (z, z) is a solution of (HBFC^2), and hence that $(x, y) = (z, z)$. Then there exists $\varepsilon > 0$ such that $\forall t \in [t_0-\varepsilon, t_0[\cup]t_0, t_0+\varepsilon]$, $x(t) \neq y(t)$, i.e., $[t_0 - \varepsilon, t_0 + \varepsilon] \cap \mathcal{D} = \{t_0\}$, which proves that the point t_0 is isolated.

Let us now verify (5.1); let $t_0 \in \mathbb{R}_+ \setminus \mathcal{D}$. There exists a neighborhood $\mathcal{V}$ of t_0 such that w coincides with x (resp., y) on $\mathcal{V}$. We then have $x(t_0) \geq y(t_0)$ (resp., $y(t_0) \geq x(t_0)$) and therefore, using assumption $(\mathcal{H}_V)$(iii), $V'(x - y)(t_0) \leq 0$ (resp., $V'(x-y)(t_0) \geq 0$). We then deduce from the first (resp., second) equation of (HBFC^2) that

$$\ddot{w}(t_0) + \gamma\dot{w}(t_0) + \phi'(w(t_0)) \geq 0.$$

Proof of (ii). Let us remark at once that the set $\mathcal{D} \cap [t', t]$ is discrete and compact, and hence finite. Relation (5.1) is then true on $[t', t]$ outside a finite number of points. Let us denote by $a_1, \dots, a_n$ the elements of $\mathcal{D} \cap]t', t[$ and set $a_0 = t'$ and $a_{n+1} = t$. We now integrate the inequality (5.1) on each interval $]a_k, a_{k+1}[$, $0 \leq k \leq n$. By adding the obtained inequalities, one finds

$$\sum_{k=0}^{n} \int_{a_k}^{a_{k+1}} \ddot{w}(s)\, ds \, + \gamma \sum_{k=0}^{n} \int_{a_k}^{a_{k+1}} \dot{w}(s)\, ds \, + \sum_{k=0}^{n} \int_{a_k}^{a_{k+1}} \phi'(w(s))\, ds \, \geq 0. \tag{5.2}$$

Since the functions $w = \max\{x, y\}$ and $\phi'(w)$ are continuous on $[t', t]$, we have

$$\sum_{k=0}^{n} \int_{a_k}^{a_{k+1}} \dot{w}(s)\, ds \, = \sum_{k=0}^{n} \Big(w(a_{k+1}) - w(a_k) \Big) = w(t) - w(t') \tag{5.3}$$

and

$$\sum_{k=0}^{n} \int_{a_k}^{a_{k+1}} \phi'(w(s))\, ds \, = \, \int_{t'}^{t} \phi'(w(s))\, ds. \tag{5.4}$$

On the other hand, the map $w = \max\{x, y\}$ verifies $\dot{w}_+ \geq \dot{w}_-$, and hence

$$\begin{aligned} \sum_{k=0}^{n} \int_{a_k}^{a_{k+1}} \ddot{w}(s)\, ds \, &= \sum_{k=0}^{n} \dot{w}_-(a_{k+1}) - \dot{w}_+(a_k) \\ &= \dot{w}_-(t) - \dot{w}_+(t') - \sum_{k=1}^{n} \Big(\dot{w}_+(a_k) - \dot{w}_-(a_k) \Big) \leq \dot{w}_-(t) - \dot{w}_+(t'). \end{aligned} \tag{5.5}$$

By combining (5.2), (5.3), (5.4), and (5.5), we obtain the expected formula. ∎

5.1.2. Convergence of $\max\{x, y\}$ and $\min\{x, y\}$.

LEMMA 5.2. *Under the assumptions of Theorem* 2.2, *the maps* $\max\{x, y\}$ *and* $\min\{x, y\}$ *converge in* $\overline{\mathbb{R}}$.

Proof of Lemma 5.2. Without any loss of generality, we prove the convergence of the map $w = \max\{x, y\}$ only. Assume that it is not true, i.e., $\liminf_{t\to+\infty} w(t) < \limsup_{t\to+\infty} w(t)$. Let us first prove that $]\liminf_{t\to+\infty} w(t), \limsup_{t\to+\infty} w(t)[\subset S$. Let $\lambda \in]\liminf_{t\to+\infty} w(t), \limsup_{t\to+\infty} w(t)[$. There is a sequence (t_n) in $\mathbb{R}$ such that $\lim_{n\to+\infty} t_n = +\infty$ and $\lim_{n\to+\infty} w(t_n) = \lambda$. Without any loss of generality, up to a subsequence, we may assume that $w(t_n) = x(t_n)$ for every n; hence from Theorem 2.1(iv), $\phi'(\lambda) = \lim_{n\to+\infty} \phi'(w(t_n)) = \lim_{n\to+\infty} \phi'(x(t_n)) = 0$.

Now consider λ and μ in $\mathbb{R}$ such that

$$\liminf_{t\to+\infty} w(t) < \lambda < \mu < \limsup_{t\to+\infty} w(t).$$

Since $\mu < \limsup_{t\to+\infty} w(t)$, there exists a sequence (t_n) such that $t_n \to +\infty$ and $w(t_n) > \mu$. Let

$$T_n = \sup\left\{u \geq t_n,\ w([t_n,u]) \geq \lambda\right\}.$$

Since $\liminf_{t\to+\infty} w(t) < \lambda$, then $T_n < +\infty$. By the continuity of w, we have $w(T_n) = \lambda$. Let

$$\tau_n = \inf\left\{u \in [t_n,T_n],\ w([u,T_n]) \leq \limsup_{t\to+\infty} w(t)\right\}.$$

If $\tau_n = t_n$ then $w(\tau_n) = w(t_n) > \mu$. If $\tau_n > t_n$, then $w(\tau_n) = \limsup_{t\to+\infty} w(t) > \mu$. Hence, in both cases, $w(\tau_n) > \mu$. For every $u \in [\tau_n, T_n]$,

$$\liminf_{t\to+\infty} w(t) < \lambda \leq w(u) \leq \limsup_{t\to+\infty} w(t);$$

hence $\phi'(w(u)) = 0$. We then deduce from Lemma 5.1 that $w(T_n) \geq w(\tau_n) - \frac{1}{\gamma}(\dot{w}_-(T_n) - \dot{w}_+(\tau_n))$, and since $w(T_n) = \lambda$ and $w(\tau_n) > \mu$,

$$\lambda > \mu - \frac{1}{\gamma}(\dot{w}_-(T_n) - \dot{w}_+(\tau_n)). \tag{5.6}$$

Since $\lim_{t\to+\infty} \dot{x}(t) = \lim_{t\to+\infty} \dot{y}(t) = 0$ (Theorem 2.1(iii)), then $\lim_{t\to+\infty} \dot{w}_-(t) = \lim_{t\to+\infty} \dot{w}_+(t) = 0$. Passing to the limit when $n \to +\infty$, (5.6) then yields $\lambda \geq \mu$, a contradiction. □

5.1.3. Convergence of the trajectories. We now come back to the proof of Theorem 2.2. In view of Lemma 5.2, the maps $\max\{x,y\}$ and $\min\{x,y\}$, respectively, converge to some m_∞ and M_∞ in $\overline{\mathbb{R}}$. If $m_\infty < M_\infty$, we directly deduce the convergence of the trajectories x and y. If $m_\infty = M_\infty = +\infty$ (resp., $-\infty$), then clearly $\lim_{t\to+\infty} x(t) = \lim_{t\to+\infty} y(t) = +\infty$ (resp., $-\infty$). Let us now examine the case $m_\infty = M_\infty \in \mathbb{R}$. The functions $x + y = \max\{x,y\} + \min\{x,y\}$ and $|x - y| = \max\{x,y\} - \min\{x,y\}$, respectively, converge to $2\,m_\infty (= 2\,M_\infty)$ and 0. Hence, $\lim_{t\to+\infty} x(t) = \lim_{t\to+\infty} y(t) = m_\infty = M_\infty$. Consequently, in all cases, there exists $(x_\infty, y_\infty) \in \overline{\mathbb{R}} \times \overline{\mathbb{R}}$ such that $\lim_{t\to+\infty} x(t) = x_\infty$ and $\lim_{t\to+\infty} y(t) = y_\infty$. Since $\lim_{t\to+\infty}(\phi'(x(t)), \phi'(y(t))) = (0,0)$ (Theorem 2.1(iv)), then $(x_\infty, y_\infty) \in \widehat{S} \times \widehat{S}$, which ends the proof of Theorem 2.2. □

5.2. Proof of Theorem 2.3. Let us first assume that $-\infty < m_\infty < M_\infty < +\infty$. Without any loss of generality, we prove only the assertion $P^+(M_\infty)$. Let us argue by contradiction and assume that there exists a neighborhood $V(M_\infty)$ of M_∞ such that $\phi'_{|V(M_\infty)} \leq 0$. Setting $w = \max\{x,y\}$, there exists $t_0 \geq 0$ such that, $\forall t \geq t_0$, $w(t) \in V(M_\infty)$, and hence

$$\forall t \geq t_0, \qquad \phi'(w(t)) \leq 0. \tag{5.7}$$

Since, by assumption, $-\infty < m_\infty < M_\infty < +\infty$, there exist $\alpha > 0$, $M > 0$, and $t_1 \geq t_0$ such that, $\forall t \geq t_1$, $\alpha \leq |x(t) - y(t)| \leq M$. From $(\mathcal{H}_V)$(iv), there exists $\eta > 0$ such that

$$\forall t \geq t_1, \qquad |V'(x(t) - y(t))| \geq \inf_{\alpha \leq |z| \leq M} |V'(z)| = \eta. \tag{5.8}$$

On the other hand, since $x(t) \neq y(t)$ for every $t \geq t_1$, then $w(t) = x(t)$ for every $t \geq t_1$ or $w(t) = y(t)$ for every $t \geq t_1$. Hence the map w is of class $\mathcal{C}^2$ on $[t_1, +\infty)$. In view of (HBFC^2) and $(\mathcal{H}_V)$(iii), the map w verifies the following differential equation:

$$\ddot{w} + \gamma \dot{w} + \phi'(w) - \varepsilon(t)|V'(x - y)| = 0. \tag{5.9}$$

In view of (5.7), (5.8), and (5.9), we find

$$\forall t \geq t_1, \qquad \ddot{w}(t) + \gamma \dot{w}(t) \geq \eta \varepsilon(t).$$

In view of Claim 5.1 below, we obtain $\lim_{t \to +\infty} w(t) = +\infty$, a contradiction.

Claim 5.1. *Let $u : [0, +\infty[\to \mathbb{R}$ be a function of class $\mathcal{C}^2$ and let $\varepsilon : [0, +\infty) \to \mathbb{R}_+$ be a continuous function such that*

$$\forall t \geq 0, \qquad \ddot{u}(t) + \gamma \dot{u}(t) \geq \eta \varepsilon(t),$$

where $\eta > 0$, $\gamma > 0$, and $\int_0^{+\infty} \varepsilon(t)\, dt = +\infty$. Then $\lim_{t \to +\infty} u(t) = +\infty$.

Proof of Claim 5.1. Let us multiply the differential inequality by $e^{\gamma s}$ and integrate twice; we find

$$u(t) \geq u(0) + \frac{\dot{u}(0)}{\gamma}(1 - e^{-\gamma t}) + \eta \int_0^t \int_0^u e^{-\gamma(u-s)} \varepsilon(s)\, ds\, du.$$

On the other hand,

$$\lim_{t \to +\infty} \int_0^t \int_0^u e^{-\gamma(u-s)} \varepsilon(s)\, ds\, du = \frac{1}{\gamma} \int_0^{+\infty} \varepsilon(s)\, ds = +\infty.$$

Hence, $\lim_{t \to +\infty} u(t) = +\infty$. ∎

Let us now assume that $m_\infty = -\infty$ and $M_\infty < +\infty$. Let us argue by contradiction and assume that we have neither $P^+(m_\infty)$ nor $P^+(M_\infty)$. Then there exists a neighborhood $V(M_\infty)$ of M_∞, respectively, $V(m_\infty)$ of m_∞, such that ${\phi'}_{|V(M_\infty)} \leq 0$, respectively, ${\phi'}_{|V(m_\infty)} \leq 0$. Consequently there exists $t_0 \geq 0$ such that, for every $t \geq t_0$,

$$\phi'(x(t)) \leq 0 \quad \text{and} \quad \phi'(y(t)) \leq 0.$$

By adding the two equations of (HBFC^2), we deduce that

$$\forall t \geq t_0, \qquad \ddot{x}(t) + \ddot{y}(t) + \gamma(\dot{x}(t) + \dot{y}(t)) \geq 0.$$

A direct computation shows that $x + y$ is bounded from below, a contradiction with $m_\infty = -\infty$. The proof of the last case $m_\infty > -\infty$ and $M_\infty = +\infty$ goes along the same lines. ∎

5.3. Proof of Corollary 2.3. Let $I(x_\infty)$ be the connected component of x_∞ in $\widehat{S}$. Let us argue by contradiction and assume that $x_\infty \notin \{\inf I(x_\infty), \sup I(x_\infty)\}$. In particular, we have $x_\infty < +\infty$. If condition (a) is satisfied, we obtain by Theorem 2.3 the assertions $P^-(y_\infty)$ and $P^+(x_\infty)$. If condition (b) is satisfied, the assertion $P^+(y_\infty)$ is false and then, by Theorem 2.3, we obtain the assertion $P^+(x_\infty)$. In both cases, we have $P^+(x_\infty)$, which is inconsistent with $x_\infty \in \operatorname{int} I(x_\infty)$, a contradiction. ∎

5.4. Proof of Corollary 2.4. We assume that the third conclusion of Corollary 2.4 does not hold, i.e., $x_\infty \neq y_\infty$. Without any loss of generality, we may assume that $x_\infty > y_\infty$. Since $\widehat{S}$ is connected, $I(x_\infty) = \widehat{S}$. From Corollary 2.3, $x_\infty \in \{\inf \widehat{S}, \sup \widehat{S}\}$. But $x_\infty > y_\infty$ with $y_\infty \in \widehat{S}$ implies that $x_\infty = \sup \widehat{S}$. ∎

5.5. Proof of Proposition 2.2. Not surprisingly, our proof of Proposition 2.2 is greatly inspired by the proof of the Alvarez theorem given in [6] and its extension to the controlled case in [4]. The Alvarez theorem is itself an extension of Bruck's theorem [11] (first order steepest descent method) to the second order dissipative (HBF) system. For the sake of completeness, we recall the outline of the proof. We refer to [4] for further details.

Let $z \in S$; we define the function $h_z : [0, +\infty) \to \mathbb{R}_+$ by

$$h_z(t) = \frac{1}{2}|x(t) - z|^2.$$

Since $\dot{h}_z(t) = \langle x(t) - z, \dot{x}(t)\rangle$ and $\ddot{h}_z(t) = |\dot{x}(t)|^2 + \langle x(t) - z, \ddot{x}(t)\rangle$, and since the map x is the solution of the (HBFC^2) system, we have

$$\ddot{h}_z(t) + \gamma \dot{h}_z(t) = |\dot{x}(t)|^2 - \langle x(t) - z, \nabla\phi(x(t))\rangle - \varepsilon(t)\langle x(t) - z, \nabla V(x(t) - y(t))\rangle.$$

Since $z \in S$, we have $\nabla\phi(z) = 0$. From the monotonicity of $\nabla\phi$, we have, for every t, $\langle x(t) - z, \nabla\phi(x(t))\rangle = \langle x(t) - z, \nabla\phi(x(t)) - \nabla\phi(z)\rangle \geq 0$. Since the trajectories x and y are bounded and since ∇V is bounded on the bounded sets, there is $C > 0$ such that $|\langle x(t) - z, \nabla V(x(t) - y(t))\rangle| \leq C$ for every t. Hence

$$\ddot{h}_z(t) + \gamma \dot{h}_z(t) \leq |\dot{x}(t)|^2 + C\varepsilon(t). \tag{5.10}$$

Since $|\dot{x}|^2 \in L^1([t_0, +\infty), \mathbb{R}_+)$ and $\varepsilon \in L^1([t_0, +\infty), \mathbb{R}_+)$, in view of [6, Lemma 4.2], the above equation implies that h_z converges. Since the function x is bounded, from Theorem 2.1(vi) it follows that for every sequence $(t_n) \subset [0, +\infty)$ such that $t_n \to +\infty$ and $x(t_n) \rightharpoonup \bar{x}$ weakly in H, we have $\bar{x} \in S$ and $\lim_{n\to+\infty} \phi(x(t_n)) = \min \phi$. Since, from above, $\lim_{t\to+\infty} |x(t) - z|$ exists for every $z \in S$, we deduce from Opial's lemma (given below) that the map x weakly converges to some element x_∞ of S.

Lemma 5.1 (Opial [22]). *Let H be a Hilbert space and let $x : [0, +\infty) \to H$ be a function such that there exists a nonempty set $S \subset H$ which verifies*

(i) *$\forall t_n \to +\infty$ with $x(t_n) \rightharpoonup \bar{x}$ weakly in H, we have $\bar{x} \in S$;*

(ii) *$\forall z \in S$, $\lim_{t\to+\infty} |x(t) - z|$ exists.*

Then $x(t)$ weakly converges as $t \to +\infty$ to some element x_∞ of S.

6. Further remarks: Precisions and generalizations. In this paper the convergence results are restricted to the one-dimensional case. They clearly call for precisions and generalizations. With numerical applications in mind, an important issue is the study and the control of the rates of convergence of the solutions, of the exponential decay of the energy, etc.

6.1. Toward higher dimensions. To obtain further results, it is natural to directly study the map $x-y$ in order to obtain more general conclusions on the maps x and y. The function $h(t) := \frac{1}{2}|x(t)-y(t)|^2$ satisfies the following equation:

$$\begin{aligned}\ddot{h}(t)+\gamma\dot{h}(t) &= |\dot{x}(t)-\dot{y}(t)|^2 - \langle \nabla\phi(x(t)) - \nabla\phi(y(t)), x(t)-y(t)\rangle \\ &\quad - 2\varepsilon(t)\langle \nabla V(x(t)-y(t)), x(t)-y(t)\rangle. \end{aligned} \tag{6.1}$$

First recall that from Proposition 2.1, $|\dot{x}-\dot{y}|^2 \in L^1([0,+\infty),\mathbb{R}_+)$. We then obtain positive results when $\phi = 0$. In the general case, it seems difficult to deduce global information from (6.1). If we assume the map ϕ to be convex, $\langle \nabla\phi(x(t)) - \nabla\phi(y(t)), x(t)-y(t)\rangle \geq 0$. But the inequality that follows from (6.1) may at most lead to upper bounds on the map h. On the other hand, since the map $\nabla\phi$ is Lipschitz continuous on the bounded sets, $\langle \nabla\phi(x(t)) - \nabla\phi(y(t)), x(t)-y(t)\rangle \leq Kh(t)$ if the maps x and y are bounded. However, even in the slow case the term $\varepsilon(t)$ is then negligible in front of K and the effect of the "slow" control does not appear in the equation deduced from (6.1).

But, as we have seen in the proof of the global existence results (section 4.1), it is possible to rewrite the (HBFC^2) system as a (HBFC) system in $H \times H$. It then seems to be natural to adapt the proofs of the convergence results of [4] for the (HBFC) system in order to obtain convergence results for the (HBFC^2) system. This direction gives indeed positive results in the case where $H = \mathbb{R}$ but does not lead to all the conclusions of Theorems 2.2 and 2.3 because of the problems arising from the direct study of the map $x - y$. But taking into account the remark in section 3.7, this method could likely help to generalize some of our results to a higher, possibly infinite dimension, of course obtaining more possible cases than in Theorem 2.3.

6.2. Numerical experiments: The three cases in Corollary 2.4. When $\widehat{S}$ is a nonempty interval, for example, when ϕ is convex, the conclusion of Corollary 2.4 shows that either the solutions x and y of the (HBFC^2) system globally explore the set $\widehat{S}$ or converge to the same limit in $\widehat{S}$. In this last case, the (HBFC^2) system does not provide more information than a simple (noncoupled) (HBF) system. A first investigation would be to find, numerically and theoretically, when this "bad" case happens—hopefully on a negligible set. A natural way is to distinguish the sets where each case happens:

$$\begin{aligned} C_i &= \{(x_0,y_0,\dot{x}_0,\dot{y}_0)\in\mathbb{R}^4, \lim_{t\to+\infty}(x(t),y(t)) = (\sup\widehat{S},\inf\widehat{S})\};\\ C_{ii} &= \{(x_0,y_0,\dot{x}_0,\dot{y}_0)\in\mathbb{R}^4, \lim_{t\to+\infty}(x(t),y(t)) = (\inf\widehat{S},\sup\widehat{S})\};\\ C_{iii} &= \{(x_0,y_0,\dot{x}_0,\dot{y}_0)\in\mathbb{R}^4, \lim_{t\to+\infty}x(t) = \lim_{t\to+\infty}y(t)\}, \end{aligned}$$

where (x,y) denotes the solution of the (HBFC^2) system with initial data $(x_0,y_0,\dot{x}_0,\dot{y}_0)$. Then from Corollary 2.4, $C_i \cup C_{ii} \cup C_{iii} = \mathbb{R}^4$ and the sets C_i, C_{ii}, and C_{iii} form a partition of $\mathbb{R}^4$ when the set $\widehat{S}$ is not reduced to a singleton. Since each case is characterized by the value $\lim_{t\to+\infty} x(t)-y(t)$, we obtain equivalently

$$\begin{aligned} C_i &= \{(x_0,y_0,\dot{x}_0,\dot{y}_0)\in\mathbb{R}^4, \lim_{t\to+\infty}x(t)-y(t) = \sup\widehat{S}-\inf\widehat{S}\};\\ C_{ii} &= \{(x_0,y_0,\dot{x}_0,\dot{y}_0)\in\mathbb{R}^4, \lim_{t\to+\infty}x(t)-y(t) = \inf\widehat{S}-\sup\widehat{S}\};\\ C_{iii} &= \{(x_0,y_0,\dot{x}_0,\dot{y}_0)\in\mathbb{R}^4, \lim_{t\to+\infty}x(t)-y(t) = 0\}. \end{aligned}$$

Numerically, we approximate the function $\Delta : (x_0,y_0,\dot{x}_0,\dot{y}_0) \mapsto \lim_{t\to+\infty} x(t)-y(t)$ by the functions $\Delta_t : (x_0,y_0,\dot{x}_0,\dot{y}_0) \mapsto x(t)-y(t)$, which converge to Δ as $t\to+\infty$. We obtain positive (numerical) results on the example considered for the illustration

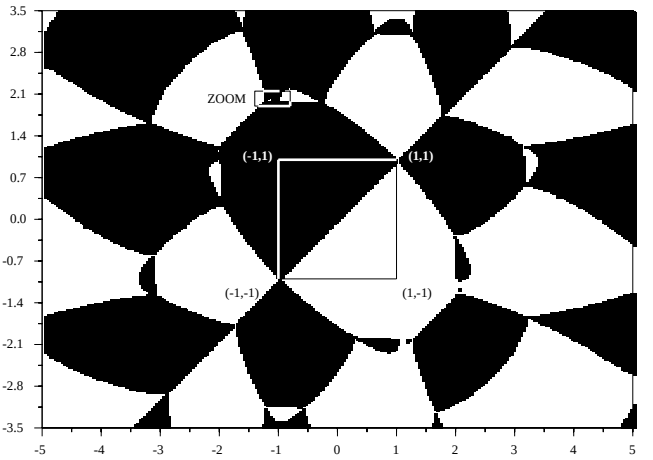


FIG. 6.1. *Case* (i) *in white and case* (ii) *in black.*

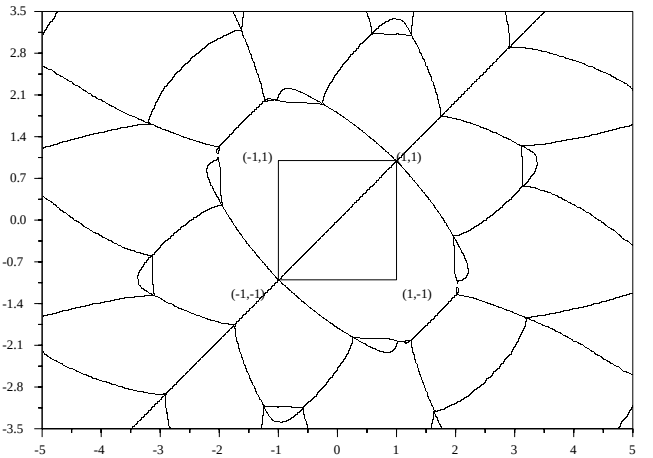


FIG. 6.2. *Cases* (i) *and* (ii) *in white and case* (iii) *in black.*

of Corollary 2.4 ($\phi(z) = (z+1)^2$ if $z \leq -1$, $\phi(z) = 0$ if $z \in [-1,1]$, $\phi(z) = (z-1)^2$ if $z \geq 1$, $V(z) = \frac{1}{2}e^{\frac{-z^2}{10}}$, $\varepsilon(t) = \frac{1}{\log(t+2)}$, $\gamma = 0.4$).

6.2.1. About the structure of the sets C_i, C_{ii}, and C_{iii}. To give an idea of the possible structure of the sets C_i, C_{ii}, and C_{iii}, we compute the function $\Delta_t : (x_0, y_0, \dot{x}_0, \dot{y}_0) \mapsto x(t) - y(t)$ for initial data $(x_0, y_0, \dot{x}_0, \dot{y}_0)$ in the set $[-5,5] \times [-3.5, 3.5] \times \{0\} \times \{0\}$ and for different times t. Figures 6.1 and 6.2 correspond to $t = 60$, respectively with a grid of 62500 points (250×250) and a grid of 10^6 points (1000×1000). On Figure 6.1, the white parts correspond to the set C_i (case (i)), and the black parts to the set C_{ii} (case (ii)). The set C_{iii} (case (iii)) is easier to visualize in Figure 6.2, where it appears in black. Note that the symmetries in the figures are only due to the symmetries in ϕ and V in our example (if $(x(t), y(t))$ is a solution, then $(y(t), x(t))$ and $(-x(t), -y(t))$ are also solutions).

Enlarging a part of Figure 6.1 (precisely, with a computation on a new grid

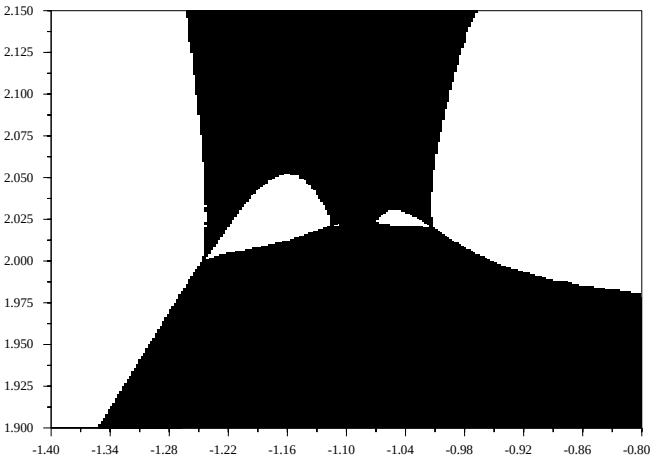


FIG. 6.3. *Zoom on the set* $[-1.4, -0.8] \times [1.9, 2.15]$. *Case* (i) *in white and case* (ii) *in black.*

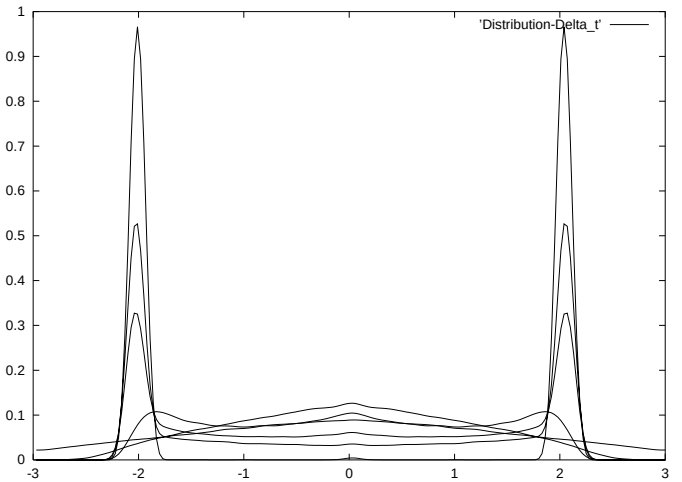


FIG. 6.4. *Distribution of the function* $x(t) - y(t)$ *at times* 0, 5, 10, 15, 20, 25, *and* 30.

of 62500 points corresponding to the zoomed part) suggests in Figure 6.3 a fractal structure of the sets C_i, C_{ii}, and C_{iii}. This also happens with other potentials ϕ and V (for example, $\phi = d^{\alpha}_{[a,b]}$ with $\alpha \geq 2$, etc.).

6.2.2. Relative weight of the sets C_i, C_{ii}, and C_{iii}. The second numerical experiment illustrates on Figure 6.4 the relative weight of the sets C_i, C_{ii}, and C_{iii} and also gives an idea of the rates of convergence. It precisely consists of evaluating the distribution of the function $x(t) - y(t)$. For initial data $(x_0, y_0, \dot{x}_0, \dot{y}_0)$ in the set $[-5, 5] \times [-5, 5] \times [-5, 5] \times [-5, 5]$ and for different times t, we compute the function Δ_t and, for a given number $p > 0$, the proportion of points which belong to an interval $[\frac{k}{p}, \frac{k+1}{p})$ for $k \in \mathbf{Z}$. In our example, the values 2 and -2, respectively, correspond to cases (i) and (ii), and the value 0 corresponds to case (iii). The experiments are computed on a grid of 10^4 points, and we limited the representation at $t = 30$ for a

matter of readability. Again, the symmetries in the figure are due to the symmetries in ϕ and V. The function corresponding at $t = 0$ is simply the density of probability $p(z = x - y)$ for $(x, y) \in [-5, 5] \times [-5, 5]$.

6.3. The singular case to avoid case (iii). An idea to avoid case (iii) consists of preventing the solutions x and y from collapsing by taking a "singular" potential V, precisely defined on $\mathbb{R} \setminus \{0\}$ and such that $\lim_{z \to 0} V(z) = +\infty$. Then the trajectories x and y never cross, and this effect could hopefully influence the asymptotic behavior of the solutions and thus avoid the case $\lim_{t \to +\infty} x(t) - y(t) = 0$ (see [13]). Note, however, that the regular case studied in this paper is still relevant from a numerical point of view since a singular potential would be numerically approximated by a regular potential.

Acknowledgment. We wish to thank Hedy Attouch for suggesting the global approach of equilibria by coupled systems and for useful comments.

REFERENCES

[1] F. Alvarez, *On the minimizing property of a second order dissipative system in Hilbert spaces*, SIAM J. Control Optim., 38 (2000), pp. 1102–1119.

[2] F. Alvarez and H. Attouch, *An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping*, Set-Valued Anal., 9 (2001), pp. 3–11.

[3] H. Attouch and R. Cominetti, *A dynamical approach to convex minimization coupling approximation with the steepest descent method*, J. Differential Equations, 128 (1996), pp. 519–540.

[4] H. Attouch and M.-O. Czarnecki, *Asymptotic control and stabilization of nonlinear oscillators with non isolated equilibria*, J. Differential Equations, 179 (2002), pp. 278–310.

[5] H. Attouch, A. Cabot, and P. Redont, *Shock solutions via epigraphical regularization of a second order in time gradient-like differential inclusion*, Adv. Math. Sci. Appl., 12 (2002), pp. 273–306.

[6] H. Attouch, X. Goudou, and P. Redont, *The heavy ball with friction method.* I. *The continuous dynamical system*, Commun. Contemp. Math., 2 (2000), pp. 1–34.

[7] J.-B. Baillon, *Un exemple concernant le comportement asymptotique de la solution du problème* $du/dt + \partial\phi(u) = 0$, J. Funct. Anal., 28 (1978), pp. 369–376.

[8] J.-B. Baillon and R. Cominetti, *A convergence result for nonautonomous subgradient evolution equations and its application to the steepest descent exponential penalty trajectory in linear programming*, J. Funct. Anal., 187 (2001), pp. 263–273.

[9] H. Brézis, *Monotonicity methods in Hilbert spaces and some applications to nonlinear partial differential equations*, in Contributions to Nonlinear Analysis, E. H. Zarantonello, ed., Academic Press, New York, 1971, pp. 101–156.

[10] H. Brézis, *Asymptotic behaviour of some evolution systems*, in Nonlinear Evolution Equations, Academic Press, New York, 1978, pp. 141–154.

[11] R. E. Bruck, *Asymptotic convergence of nonlinear contraction semigroups in Hilbert space*, J. Funct. Anal., 18 (1975), pp. 15–26.

[12] J.-M. Coron, *On the stabilization of some nonlinear control systems: Results, tools, and applications*, in Nonlinear Analysis, Differential Equations and Control (Montreal, QC, 1998), NATO Sci. Ser. C Math. Phys. Sci. 528, Kluwer Academic, Dordrecht, The Netherlands, 1999, pp. 307–367,

[13] M.-O. Czarnecki, *Asymptotic Control of Pairs of Oscillators Coupled by a Repulsion, with Non Isolated Equilibria* II: *The Singular Case*, manuscript.

[14] H. Furuya, K. Miyashiba, and N. Kenmochi, *Asymptotic behaviour of solutions to a class of nonlinear evolution equations*, J. Differential Equations, 62 (1986), pp. 73–94.

[15] X. Goudou, *Genericity of the Convergence towards a Local Minimum of the Heavy Ball Method*, working paper.

[16] A. Haraux, *Systèmes dynamiques dissipatifs et applications*, Rech. Math. Appl. 17, Masson, Paris, 1991.

[17] A. Haraux and M. A. Jendoubi, *Convergence of bounded weak solutions of the wave equation with dissipation and analytic nonlinearity*, Calc. Var. Partial Differential Equations, 9 (1999), pp. 95–124.

[18] M. A. Jendoubi, *Convergence of global and bounded solutions of the wave equation with linear dissipation and analytic nonlinearity*, J. Differential Equations, 144 (1998), pp. 302–312.
[19] S. Lojasiewicz, *Une propriété topologique des sous-ensembles analytiques réels*, Colloques internationaux du CNRS 117, Les équations aux dérivées partielles, 1963.
[20] S. Lojasiewicz, *Ensembles semi-analytiques*, notes, Bures-sur Yvette, Institut des Hautes Etudes Scientifiques, 1965.
[21] B. Mohammadi and O. Pironneau, *Applied Shape Optimization for Fluids*, Oxford University Press, London, 2001.
[22] Z. Opial, *Weak convergence of the sequence of successive approximations for nonexpansive mappings*, Bull. Amer. Math. Soc., 73 (1967), pp. 591–597.
[23] P. Redont, *Equation de la boule pesante avec frottement: Exemple de solution non convergente*, Prépublication 99, Département de Mathématiques, Université de Montpellier II; available online from http://www.math.univ-montp2.fr.
[24] A. N. Tikhonov and V. Ya. Arsenine, *Méthodes de résolution de problèmes mal posés*, MIR, Moscow, 1976.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 4, pp. 1281–1294

FEEDBACK SPREADING CONTROL UNDER SPEED CONSTRAINTS*

K. KASSARA†

Abstract. This paper is concerned with the control of spread in semilinear parabolic systems. It first introduces a formula in order to measure the speed of a spread. Then feedback spreading control laws under speed constraints are studied by a set-valued approach. The optimality of such laws is examined in the case in which the system is affine dependent upon the control.

Key words. semilinear parabolic systems, spreading control, lower semicontinuity of maps, constrained optimization

AMS subject classifications. Primary, 93C20, 49J20; Secondary, 54C60

PII. S0363012900369915

1. Introduction.

Spreadable distributed parameter systems provide a mathematical context for modeling expansion phenomena which may arise in spatially distributed processes; cf. [6, 7] and the references therein.

In handling the control aspects of that concept, a first attempt has been made in [8], where it is shown that spreading control can be determined by minimizing a rather unusual criterion, which is partly quadratic but contains a nonquadratic term. Conditions for a solution are given, the optimality system is derived, and algorithms for the resolution are determined. It is even of interest to cite [9], which relates spreading control to actuators for a class of linear distributed parameter systems.

Nevertheless, all of the approaches cited above have the disadvantage of being restricted to linear systems and concern only a few situations. In a recent study [13], it has been pointed out that feedback spreading controls for semilinear partial differential equations may be investigated in the framework of *monotone* solutions with respect to a *preorder*; cf. [1, 17]. Then the application of some results on monotonicity by [4] has allowed us to characterize these controls as selections of a certain set-valued map, which is defined by a set of tangential conditions.

The present study continues the investigation of the field as expounded in [13] by essentially concentrating on the speed of a spread. For this, we are motivated by the technical need to design spreads, taking into consideration both the speed and the time of spreading; cf. [8]. First, we propose a convenient setting in which the measure of the spread speed can be rigorously made. Then, due to some set-valued analysis facts, we examine the existence of feedback spreading control laws, which generate a spread either slower or quicker than a desired given speed.

In this paper, the following definitions and notation are used. Let Y be a Hilbert space; then a set-valued map $Q : \mathcal{S} \to 2^Y \setminus \{\emptyset\}$ is said to be lower semicontinuous (lsc) whenever the following property holds: For each $z_0 \in \mathcal{S}$ and any sequence of elements $z_n \in \mathcal{S}$ converging to z_0, for every $y_0 \in Q(z_0)$, there exists a sequence of elements $y_n \in Q(z_n)$ which converges to y_0.

*Received by the editors March 27, 2000; accepted for publication (in revised form) May 7, 2002; published electronically December 3, 2002.
http://www.siam.org/journals/sicon/41-4/36991.html
†Department of Mathematics, University of Casablanca 1, P.O. Box 5366, Casablanca, Morocco (kassara@facsc-achok.ac.ma).

The graph of Q is denoted by

$$\operatorname{graph}(Q) \doteq \{(z,y) \in \mathcal{S} \times Z \mid y \in Q(z)\}.$$

The inverse of Q is the map $Q^{-1} : Y \to 2^{\mathcal{S}}$ defined by

$$Q^{-1}(y) \doteq \{z \in \mathcal{S} \mid y \in Q(z)\} \quad \text{for each } y \in Y.$$

A selection of the map Q is a mapping $\nu : \mathcal{S} \to Y$, which satisfies

$$\nu(z) \in Q(z) \quad \text{for each } z \in \mathcal{S}.$$

We quote Michael's selection theorem, which states that any lsc set-valued map with closed convex values has a continuous selection; cf. [5].

A mapping from Z to Y is said to be demicontinuous if it maps strongly convergent sequences in Z into weakly convergent sequences in Y; cf. [15].

When the scalar product in Y is clear from the context, it is denoted by $\langle\ ;\ \rangle$.

The projector of best approximation on a closed convex subset $\mathcal{K}$ of Y will be denoted by $\pi_{\mathcal{K}}(\cdot)$.

The directional derivative of a functional $\ell : \mathcal{S} \to \mathbb{R}$ in the direction of $y \in Y$, if it exists at a point $z \in \mathcal{S}$, is denoted by

$$d\ell(z)(y) \doteq \liminf_{h \downarrow O} \frac{\ell(z+hy) - \ell(z)}{h}. \tag{1.1}$$

Note that, if ℓ is Gâteaux differentiable at z, then we get

$$d\ell(z)(y) = \langle \nabla \ell(z), y \rangle \quad \text{for each } y \in Y,$$

where $\nabla \ell$ denotes the Gâteaux derivative of ℓ; cf. [11].

The paper is organized as follows: In section 2, we set the spreading control problems in their open loop form. Then section 3 gives the basic results on feedback spreading control laws. In section 4, we state the speed functional and show some results which justify its definition. In section 5, we deal with feedback spreading control laws under speed constraints. Finally, section 6 is devoted to the optimality of these control laws.

2. Statement of the problem. Let $\Omega \subset \mathbb{R}^n$ be an open and bounded domain with sufficiently smooth boundary $\partial\Omega$, and set $Q = \Omega \times (0, \infty[$. Let A be a second order elliptic operator on Ω given in the form

$$A \doteq -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial}{\partial x_j}\right) + \sum_{i=1}^{n} a_i(x)\frac{\partial}{\partial x_i} + a_0(x), \tag{2.1}$$

with the smooth functions a_{ij}, a_i, and a_0. For convenient boundary data, it can be assumed that the operator $-A$ stands for an unbounded densely defined linear operator which generates a $\mathcal{C}_0$ analytic semigroup $(S(t))_{t \geq 0}$ on $Z = L^2(\Omega)$; cf. [2, 3].

We consider the semilinear parabolic control system

$$\frac{\partial z}{\partial t} + Az = \varphi(z,v) \qquad \text{in } Q \tag{2.2a}$$

with initial data

$$z(x,0) = z_0(x) \qquad \text{in } \Omega, \tag{2.2b}$$

where $z_0 \in \mathrm{dom}(A)$ (i.e., the domain of A) and φ denotes a nonlinear operator which maps $\mathcal{S} \times V$ into Z, with V another Hilbert space and $\mathcal{S}$ a closed subset of Z. Let ω be a map defined as follows:

$$\omega : \mathcal{S} \subset Z \to 2^{\Omega}. \tag{2.3}$$

DEFINITION 2.1 (cf. [6]). *A measurable function $\bar{v} : [0, t_1[\to V$ is called a spreading control with respect to ω if there exists a solution $\bar{z}$ which satisfies*

$$\bar{z}(t) \in \mathcal{S} \quad \textit{for all } t \in [0, t_1[\tag{2.4a}$$

and

$$(\omega(\bar{z}(t)))_{0 \le t < t_1} \textit{ is nondecreasing.} \tag{2.4b}$$

As an instance of the map ω, we consider the pollution process; cf. [8, 12]. It takes place when the system which describes the *concentration of pollutant* is spreadable with respect to ω, with

$$\omega(z) \doteq \{x \in \Omega \mid z(x) > z_{\max}\},$$

where $z_{\max}$ is a tolerance coefficient. Let $t_1 > 0$, and set

$$\mathcal{V}^s_{t_1} \doteq \{v \in L^2(0, t_1, V) \mid v \text{ is a spreading control with respect to } \omega\}.$$

For each control $v \in \mathcal{V}^s_{t_1}$, denote by $z(\cdot, v)$ the solution of (2.2) on the interval $[0, t_1[$; then a natural way to define the speed of the generated spread $(\omega(z(t, v)))_t$ may be

$$\mathrm{speed}(t, v) \doteq \liminf_{h \downarrow O} \frac{\lambda(\omega(z(t+h, v)) \setminus \omega(z(t, v)))}{h} \ge 0 \quad \text{for each } t \in [0, t_1[\,, \tag{2.5}$$

where λ stands for the Lebesgue measure on Ω. Then, roughly, the control problems we shall consider in this paper are stated as follows:

$$P^+_m \qquad \begin{array}{l} \text{Find a control } v^+_m \in \mathcal{V}^s_{t_1} \text{ such that} \\ \mathrm{speed}(t, v^+_m) \ge m(t) \text{ for all } t \in [0, t_1[\end{array}$$

and

$$P^-_m \qquad \begin{array}{l} \text{Find a control } v^-_m \in \mathcal{V}^s_{t_1} \text{ such that} \\ \mathrm{speed}(t, v^-_m) \le m(t) \text{ for all } t \in [0, t_1[\,, \end{array}$$

where $m : [0, t_1[\to \mathbb{R}^+$ stands for a measurable function. Also, we are concerned with investigating the optimal control problems

$$P^+_{\theta,m} \qquad \begin{array}{l} \min \|v_s\|^2_{L^2(0,t_1,V)} \text{ subject to} \\ v_s \text{ is a solution of } P^+_m \end{array}$$

and

$$P^-_{\theta,m} \qquad \begin{array}{l} \min \|v_s\|^2_{L^2(0,t_1,V)} \text{ subject to} \\ v_s \text{ is a solution of } P^-_m. \end{array}$$

Note that there are two technical notes which might be taken into account:

(i) According to [13], it should be of interest to seek feedback spreading control laws in the form

$$v_s = \psi(z) \quad \text{for each } z \in \mathcal{S}. \tag{2.6}$$

(ii) In general, the Lebesgue measure λ does not provide a well-defined function $\mathrm{speed}(\cdot, \cdot)$. That, a priori, depends upon the differentiability of $\lambda \circ \omega$ in the sense of Dini; cf. [11].

3. Preliminaries on feedback spreading control laws. In this section, we present a summary of the main definitions and results related to the concept of feedback spreading control. Let ω be as in (2.3).

DEFINITION 3.1 (cf. [13]). *The mapping $\varsigma : \mathcal{S} \to V$ is said to be a feedback spreading control (fsc) law with respect to ω if, for all initial data z_0 in $\mathcal{S}$, there exists a solution $\bar{z}$ which satisfies*

$$\bar{z}(t) \in \mathcal{S} \quad \textit{for all } t \in [0, t_1[\tag{3.1a}$$

and

$$\bar{v} = \varsigma(\bar{z}) \textit{ is a spreading control.} \tag{3.1b}$$

Next, for each couple $(y, z) \in Z \times \mathcal{S}$, consider the following tangential condition:

$$\begin{array}{l} \text{for all } \delta > 0, \exists\ 0 < h < \delta, \text{ and } \|p\| \le \delta \text{ such that} \\ S(h)z + h(y+p) \in \mathcal{S} \text{ and} \\ \omega(S(h)z + h(y+p)) \supset \omega(z). \end{array} \tag{3.2}$$

Then define the set-valued maps

$$\mathcal{T}_\omega(z) \doteq \{y \in Z \mid \text{ (3.2) holds with } (y, z)\} \quad \text{for each } z \in \mathcal{S} \tag{3.3}$$

and

$$\mathcal{F}_\omega(z) \doteq \{v \in V \mid \varphi(z, v) \in \mathcal{T}_\omega(z)\} \quad \text{for each } z \in \mathcal{S}. \tag{3.4}$$

Also, we need to let

$$\Sigma_\omega \doteq \{(y, z) \in \mathcal{S}^2 \mid \omega(y) \supset \omega(z)\} \tag{3.5}$$

and make the following assumption.

Assumption 3.2. The semigroup $S(\cdot)$ is compact.

We are ready to present the following basic result which characterizes fsc laws.

THEOREM 3.3. *Let Assumption* 3.2 *hold, and let $\varsigma : \mathcal{S} \to V$ be a measurable function. Furthermore, assume that*

(i) *Σ_ω is closed,*

(ii) *$\varphi(\cdot, \varsigma(\cdot))$ is demicontinuous on $\mathcal{S}$.*

Then ς is an fsc law with respect to ω iff ς is a selection of $\mathcal{F}_\omega$.

Proof. See [13, Theorem 3.1]. □

It should be convenient to emphasize that, in Theorem 3.3, only $\varphi(\cdot, \varsigma(\cdot))$ is required to be demicontinuous, and there are no continuity assumptions on φ or ς. Also, note that Assumption 3.2 is generic for parabolic systems; cf. [3, 16].

Remark 3.4. It is useful to notice that the subset $\mathcal{T}_\omega(z)$ may be expressed in terms of contingent subsets [17], which are given by

$$T_D^A(z) = \left\{ y \in Z \mid \liminf_{h \downarrow 0} \frac{d(S(h)z + hy, D)}{h} = 0 \right\}.$$

We have, by considering (3.2),

$$\mathcal{T}_\omega(z) = T^A_{P_\omega(z)}(z) \quad \text{for each } z \in \mathcal{S},$$

where

$$P_\omega(z) = \{y \in Z \mid \omega(y) \supset \omega(z)\} \quad \text{for each } z \in \mathcal{S}. \tag{3.6}$$

In the preliminary result below, we use the following assumption.

Assumption 3.5. Σ_ω is closed, and the map ω^{-1} has convex values.

LEMMA 3.6.

(i) *The map $\mathcal{T}_\omega$ has closed values.*

(ii) *Under Assumption* 3.5*, the map $\mathcal{T}_\omega$ has convex values.*

Proof. Let z belong to $\mathcal{S}$; then the tangential condition (3.2) yields

$$\mathcal{T}_\omega(z) = \bigcap_{\delta>0} cl \bigcup_{h\in(0,\delta)} \frac{1}{h}[P_\omega(z) - S(h)z].$$

Then it is obvious that $\mathcal{T}_\omega(z)$ is closed.

Regarding (ii), let $y, \bar{y} \in \mathcal{T}_\omega(z)$, and $\alpha, \beta \geq 0$ such that $\alpha + \beta = 1$. It follows that

$$S(h)z + h(\alpha y + \beta\bar{y}) = \alpha(S(h)z + hy) + \beta(S(h)z + h\bar{y}).$$

Now it is not hard to show that Assumption 3.5 implies that the map P_ω has closed convex values. It follows that the function

$$y \in Z \to d(y, P_\omega(z))$$

is convex, and therefore we get

$$d(S(h)z + h(\alpha y + \beta\bar{y}), P_\omega(z)) \leq \alpha d(S(h)z + hy, P_\omega(z)) + \beta d(S(h)z + h\bar{y}, P_\omega(z)).$$

Consequently, Remark 3.4 yields the result. ∎

4. The speed functional. Let μ be a measure on Ω. By the *speed functional*, we mean the functional defined on graph$(\mathcal{T}_\omega)$ by

$$\theta(y,z) \doteq \liminf_{h\downarrow 0, \|p\|\to 0} \frac{\mu(\omega(S(h)z + h(y+p)) \setminus \omega(z))}{h} \quad \text{for each } z \in \mathcal{S} \text{ and } y \in \mathcal{T}_\omega(z). \tag{4.1}$$

Let

$$\tau_\omega \doteq \mu \circ \omega : \begin{array}{ll} \mathcal{S} & \to \mathbb{R}^+, \\ z & \to \mu(\omega(z)). \end{array}$$

Next, we prove some immediate properties which are verified by the speed functional.

PROPOSITION 4.1.

(i) *θ is well defined and has values ranging in $[0, \infty]$.*

(ii) *Assume τ_ω to be locally Lipschitz on $\mathcal{S}$. Let $\bar{v}$ and $\bar{z}$ be as in Definition* 2.1*, with μ instead of λ; then we have*

$$\text{speed}(t, \bar{v}) = \theta(\varphi(\bar{z}(t), \bar{v}(t)), \bar{z}(t)) \quad \textit{for each } t \in [0, t_1[\,. \tag{4.2}$$

(iii) *Suppose that $\mathcal{S}$ and τ_ω are convex; then we have*

$$\theta(y,z) = d\tau_\omega(z)(y - Az) \quad \textit{for each } y \in \mathcal{T}_\omega(z) \textit{ and } z \in \mathcal{S} \cap \text{dom}(A). \tag{4.3}$$

Proof. First, note that (i) is simply a consequence of (3.2) and (3.3). To show (ii), let $\bar{v}$ and $\bar{z}$ be as in Definition 2.1, and denote $\bar{\varphi}(\cdot) \doteq \varphi(\bar{z}(\cdot), \bar{v}(\cdot))$. For $t \in [0, t_1[$, we get

$$\bar{z}(t+h) = \bar{z}(t) + h(\bar{\varphi}(t) + p_h) \quad \text{with } p_h \to 0 \text{ when } h \to 0.$$

Then, applying formula (4.1) yields

$$\begin{aligned} \theta(\varphi(\bar{z}(t), \bar{v}(t)), \bar{z}(t)) &= \liminf_{h \downarrow 0, \|p\| \to 0} \frac{\tau_\omega(S(h)\bar{z}(t) + h(\bar{\varphi}(t) + p)) - \tau_\omega(\bar{z}(t))}{h} \\ &= \liminf_{h \downarrow 0, \|p\| \to 0} \frac{\tau_\omega(\bar{z}(t+h) + h(p - p_h)) - \tau_\omega(\bar{z}(t))}{h}. \end{aligned}$$

Now, we observe that

$$\begin{aligned} \tau_\omega(\bar{z}(t+h) + h(p-p_h)) - \tau_\omega(\bar{z}(t)) &= \tau_\omega(\bar{z}(t+h)) - \tau_\omega(\bar{z}(t)) \\ &\quad + \tau_\omega(\bar{z}(t+h) + h(p-p_h)) - \tau_\omega(\bar{z}(t+h)). \end{aligned}$$

We use the fact that τ_ω is locally Lipschitz to obtain

$$\lim_{h \downarrow 0, \|p\| \to 0} \frac{\tau_\omega(\bar{z}(t+h) + h(p-p_h)) - \tau_\omega(\bar{z}(t))}{h} = 0.$$

Therefore, (ii) is proved if we refer to (2.5).

Regarding statement (iii), we first remark that, due to its convexity, the mapping τ_ω has a directional derivative on $\mathcal{S}$, and $d\tau_\omega(z)(\cdot)$ is continuous for each z; cf. [11]. On the other hand, we have

$$S(h)z = z - hAz + hp_h \quad \text{for each } h \geq 0, \quad \text{with } p_h \to 0 \text{ when } h \to 0.$$

It follows that

$$\theta(y, z) = \liminf_{h \downarrow 0, \|p\| \to 0} \frac{\tau_\omega(z + h(y - Az + p)) - \tau_\omega(z)}{h}.$$

Therefore, we have

$$\theta(y, z) = \liminf_{\|p\| \to 0} d\tau_\omega(z)(y - Az + p),$$

and consequently we obtain (4.3) thanks to the continuity of the directional derivative. □

As an important consequence, we stress that the speed functional provides a proper tool in order to measure the speed of the spread generated by a spreading control, especially when τ_ω has a directional derivative, in which case formula (4.3) can easily be used.

Remark 4.2. Note that in (iii) the assumption "τ_ω is convex" may be replaced by "τ_ω is Gâteaux differentiable." In this case, we obtain the formula

$$\theta(y, z) = \langle \nabla \tau_\omega(z); y - Az \rangle \quad \text{for each } y \in \mathcal{T}_\omega(z) \text{ and } z \in \mathcal{S} \cap \operatorname{dom}(A). \tag{4.4}$$

Next, we show a technical result to be used in the subsequent sections. To this end, let us consider the following assumption.

Assumption 4.3. For each sequence $(z_n)_n \subset \mathcal{S}$ and $(y_n)_n \subset Z$ such that $y_n \in \mathcal{T}_\omega(z_n)$ for every n, we have

$$\begin{array}{l} z_n \to z \text{ (strong)} \\ y_n \to y \text{ (weak)} \end{array} \implies y \in \mathcal{T}_\omega(z), \text{ and } \theta(y_n, z_n) \to \theta(y, z).$$

Then we can prove the following result, which studies the convexity of the mapping $\theta(\cdot, z)$ on $\mathcal{T}_\omega(z)$.

LEMMA 4.4. *Let Assumptions* 3.5 *and* 4.3 *be satisfied; then we have the following statements:*

(i) *If τ_ω is convex, then $\theta(\cdot, z)$ is convex on $\mathcal{T}_\omega(z)$ for each $z \in \mathcal{S}$.*

(ii) *If τ_ω is Gâteaux differentiable, then, for each $\alpha, \beta > 0$ with $\alpha + \beta = 1$, we have*

$$\theta(\alpha y + \beta \bar{y}, z) = \alpha\theta(y, z) + \beta\theta(\bar{y}, z) \quad \textit{for each } z \in \mathcal{S} \textit{ and } y, \bar{y} \in \mathcal{T}_\omega(z).$$

Proof. For $z \in \mathcal{S} \cap \operatorname{dom}(A)$, by considering Proposition 4.1(iii), we can easily see that $\theta(\cdot, z)$ is convex on $\mathcal{T}_\omega(z)$ because $d\tau_\omega(z)$ is such. Now let $z \in \mathcal{S}$; then $z = \lim_{n\to\infty} z_n$ for a sequence $(z_n)_n \subset \mathcal{S} \cap \operatorname{dom}(A)$. Let $\alpha, \beta \geq 0$ such that $\alpha + \beta = 1$ and $y, \bar{y} \in \mathcal{T}_\omega(z)$; then using Assumption 4.3 yields

$$\begin{aligned} \theta(\alpha y + \beta \bar{y}, z) &= \lim_{n\to\infty} \theta(\alpha y + \beta \bar{y}, z_n) \\ &\leq \alpha \lim_{n\to\infty} \theta(y, z_n) + \beta \lim_{n\to\infty} \theta(\bar{y}, z_n) \\ &\leq \alpha\theta(y, z) + \beta\theta(\bar{y}, z), \end{aligned}$$

and hence (i) is shown. Similarly, statement (ii) easily follows from Remark 4.2 and Assumption 4.3. ∎

5. Feedback spreading controls with constraints on the speed. In this section, based on the results provided by Theorem 3.3 and Proposition 4.1, we suitably restate the spreading control problems P_m^+ and P_m^- of section 2 in their feedback version. Let ν be a nonnegative measurable function on $\mathcal{S}$; then problem P_m^+ may read as follows:

$$\mathbb{P}_\nu^+ \qquad \begin{array}{l} \text{Find an fsc law } v = \varsigma_\nu^+(z) \text{ such that} \\ \rho(z, v) \geq \nu(z) \text{ for each } z \in \mathcal{S}. \end{array}$$

Also, problem P_m^- can be reformulated as

$$\mathbb{P}_\nu^- \qquad \begin{array}{l} \text{Find an fsc law } v = \varsigma_\nu^-(z) \text{ such that} \\ \rho(z, v) \leq \nu(z) \text{ for each } z \in \mathcal{S}, \end{array}$$

where the functional ρ is defined according to (4.2) by

$$\rho(z, v) \doteq \theta(\varphi(z, v), z) \quad \text{for each } z \in \mathcal{S} \text{ and } v \in \mathcal{F}_\omega(z), \tag{5.1}$$

and $\mathcal{F}_\omega(\cdot)$ is as in (3.4). Now define the following maps for each $z \in \mathcal{S}$:

$$\mathcal{T}_\omega^{\nu^+}(z) \doteq \{y \in \mathcal{T}_\omega(z) \mid \theta(y, z) \geq \nu(z)\} \tag{5.2a}$$

and

$$\mathcal{F}_\omega^{\nu^+}(z) \doteq \{v \in V \mid \varphi(z, v) \in \mathcal{T}_\omega^{\nu^+}(z)\}. \tag{5.2b}$$

We also need to set

$$\mathcal{T}_\omega^{\nu^-}(z) \doteq \{y \in \mathcal{T}_\omega(z) \mid \theta(y,z) \leq \nu(z)\} \tag{5.3a}$$

and

$$\mathcal{F}_\omega^{\nu^-}(z) \doteq \{v \in V \mid \varphi(z,v) \in \mathcal{T}_\omega^{\nu^-}(z)\}. \tag{5.3b}$$

Consequently, providing that the assumptions of Theorem 3.3 are satisfied by ς_ν^ϵ (with ϵ denoting $+$ or $-$), the following statement holds:

$$\varsigma_\nu^\epsilon \text{ is a solution of problem } \mathbb{P}_\nu^\epsilon \iff \varsigma_\nu^\epsilon \text{ is a selection of } \mathcal{F}_\omega^{\nu^\epsilon}. \tag{5.4}$$

For both problems $\mathbb{P}_\nu^+$ and $\mathbb{P}_\nu^-$, we respectively define the subsets of admissible speeds ν as follows:

$$\mathcal{A}_\omega^+ \doteq \{\nu : \mathcal{S} \to \mathbb{R}^+ \mid \text{for all } z \in \mathcal{S}, \exists y \in \mathcal{T}_\omega(z) \text{ such that } \theta(y,z) > \nu(z)\} \tag{5.5a}$$

and

$$\mathcal{A}_\omega^- \doteq \{\nu : \mathcal{S} \to \mathbb{R}^+ \mid \text{for all } z \in \mathcal{S}, \exists y \in \mathcal{T}_\omega(z) \text{ such that } \theta(y,z) < \nu(z)\}. \tag{5.5b}$$

In order to state an existence result for problem $\mathbb{P}_\nu^+$ for appropriate speeds ν, we first begin by proving the following lemma, which studies the lower semicontinuity of the map $\mathbb{P}_\nu^+$.

LEMMA 5.1. *Let Assumptions* 3.5 *and* 4.3 *be satisfied. Furthermore, suppose that*

(i) $\mathcal{T}_\omega$ *is lsc,*
(ii) τ_ω *is Gâteaux differentiable,*
(iii) $\nu \in \mathcal{A}_\omega^+$ *and is upper semicontinuous.*

Then the map $\mathcal{T}_\omega^{\nu^+}$ *is lsc.*

Proof. Since $\nu \in \mathcal{A}_\omega^+$, it can easily be seen that $\mathcal{T}_\omega^{\nu^+}(z) \neq \emptyset$ for each $z \in \mathcal{S}$. Now, to see that this map is lsc, it suffices to show that the functional

$$\jmath : z \in \mathcal{S} \to d(y_0, \mathcal{T}_\omega^{\nu^+}(z))^2$$

is upper semicontinuous for each $y_0 \in Y$; cf. [5, Lemma 4.2]. Indeed, given $y_0 \in Y$ and $z \in Z$, we have

$$\jmath(z) = \min_{\substack{y \in \mathcal{T}_\omega(z) \\ \nu(z) - \theta(y,z) \leq 0}} \|y_0 - y\|^2. \tag{5.6}$$

By virtue of Lemma 3.6, $\mathcal{T}_\omega(z)$ is closed and convex. On the other hand, the function $\nu(z) - \theta(\cdot, z)$ is continuous (by Assumption 4.3) and convex (due to Lemma 4.4). Then there is a unique $y_+(z) \in \mathcal{T}_\omega(z)$ which solves the optimization problem (5.6). For such a problem, the fact that $\nu \in \mathcal{A}_\omega^+$ obviously provides the Slater condition as in [11, Theorem 6.7] yields the formula

$$\jmath(z) = \sup_{\lambda \geq 0} \inf_{y \in \mathcal{T}_\omega(z)} \{\|y_0 - y\|^2 + \lambda(\nu(z) - \theta(y,z))\} \quad \text{for each } z \in \mathcal{S}. \tag{5.7}$$

Now let $(z_n)_n$ be a sequence in $\mathcal{S}$ which converges to z. By condition (i) and the fact that $y_+(z) \in \mathcal{T}_\omega(z)$, there exists a sequence $y_n \in \mathcal{T}_\omega(z)$ which converges to $y_+(z)$. It follows that

$$\inf_{y \in \mathcal{T}_\omega(z_n)} \{\|y_0 - y\|^2 + \lambda(\nu(z_n) - \theta(y,z_n))\} \leq \|y_0 - y_n\|^2 + \lambda(\nu(z_n) - \theta(y_n,z_n)) \tag{5.8}$$

for each $\lambda \geq 0$ and $n \in \mathbb{N}$. However, since ν is upper semicontinuous and

$$\theta(y_n, z_n) \to \theta(y_+(z), z),$$

we get

$$\limsup_{n\to\infty}(\nu(z_n) - \theta(y_n, z_n)) \leq \nu(z) - \theta(y_+(z), z) \leq 0.$$

Consequently, by passing to the lim sup in (5.8), we obtain

$$\limsup_{n\to\infty} \inf_{y\in\mathcal{T}_\omega(z_n)} \{\|y_0 - y\|^2 + \lambda(\nu(z_n) - \theta(y, z_n))\} \leq \|y_0 - y_+(z)\|^2 \tag{5.9}$$

for each $\lambda \geq 0$. Next, by writing $\jmath(z_n)$ by (5.7) and noting that

$$\limsup_{n\to\infty} \sup_{\lambda\geq 0}[\cdot] \leq \sup_{\lambda\geq 0} \limsup_{n\to\infty}[\cdot],$$

we get the desired inequality

$$\limsup_{n\to\infty} \jmath(z_n) \leq \jmath(z),$$

ending the proof of the lemma. □

Now we turn our attention to examine the lower semicontinuity of the map $\mathcal{T}_\omega^{\nu^-}$ as defined by (5.3a). Arguing as in the proof of Lemma 5.1, we can easily show the following result.

LEMMA 5.2. *Let Assumptions* 3.5 *and* 4.3 *be satisfied. Furthermore, suppose that*

(i) $\mathcal{T}_\omega$ *is lsc,*
(ii) τ_ω *is convex or Gâteaux differentiable,*
(iii) $\nu \in \mathcal{A}_\omega^-$ *and is lsc.*

Then the map $\mathcal{T}_\omega^{\nu^-}$ *is lsc.*

Consequently, we are in a position to provide existence results for problems $\mathbb{P}_\nu^+$ and $\mathbb{P}_\nu^-$. For that purpose, we need to take into consideration the following hypothesis.

Assumption 5.3. For each $z \in \mathcal{S}$ and $y \in \mathcal{T}_\omega(z)$, there exists $v \in V$ such that $\varphi(z, v) = y$.

PROPOSITION 5.4. *Let Assumptions* 3.2 *and* 5.3 *hold, and assume that all of the conditions of Lemma* 5.1 *(resp., Lemma* 5.2*) are satisfied. Then there exists an fsc law which solves problem* $\mathbb{P}_\nu^+$ *(resp., problem* $\mathbb{P}_\nu^-$*).*

Proof. By Lemma 5.1 (resp., Lemma 5.2), the map $\mathcal{T}_\omega^{\nu^+}$ (resp., $\mathcal{T}_\omega^{\nu^-}$) is lsc. Moreover, due to Assumption 4.3 and Lemma 4.4, it has closed convex values. Then, thanks to Michael's selection theorem which is stated in section 1, the map $\mathcal{T}_\omega^{\nu^+}$ (resp., $\mathcal{T}_\omega^{\nu^-}$) admits a continuous selection $y_+(\cdot)$ (resp., $y_-(\cdot)$). Next, we can use Assumption 5.3 to construct a mapping $\varsigma_\nu^+ : \mathcal{S} \to V$ (resp., ς_ν^-) in such a manner that

$$\varphi(z, \varsigma_\nu^\epsilon(z)) = y_\epsilon(z) \quad \text{for each } z \in \mathcal{S} \text{ with } \epsilon \in \{+, -\}.$$

Consequently, by using Theorem 3.3, it follows that ς_ν^+ (resp., ς_ν^-) stands for the desired fsc law. □

6. Optimality of fsc laws with speed constraints. In the same spirit of section 5, we investigate in this section the feedback versions of optimal control problems $P^+_{\theta,m}$ and $P^-_{\theta,m}$, which are stated in section 2. Let ν be a measurable function defined from $\mathcal{S}$ with nonnegative values; then problem $P^+_{\theta,m}$ may read as follows:

$$\mathbb{P}^+_{\theta,\nu} \quad \begin{array}{l} \text{Find an fsc law } v = \hat{\varsigma}_+(z) \text{ which solves} \\ \min \|v\|^2 \text{ subject to } \varphi(z,v) \in \mathcal{T}_\omega^{\nu^+}(z) \text{ for each } z \in \mathcal{S}. \end{array}$$

Also, problem $P^-_{\theta,m}$ can be restated as

$$\mathbb{P}^-_{\theta,\nu} \quad \begin{array}{l} \text{Find an fsc law } v = \hat{\varsigma}_-(z) \text{ which solves} \\ \min \|v\|^2 \text{ subject to } \varphi(z,v) \in \mathcal{T}_\omega^{\nu^-}(z) \text{ for each } z \in \mathcal{S}. \end{array}$$

By virtue of Theorem 3.3, the above problems can be treated through satisfying what follows.

(a) The involved parameterized minimization problems

$$\min \|v\|^2 \text{ subject to } \varphi(z,v) \in \mathcal{T}_\omega^{\nu^\epsilon}(z) \quad \text{for each } z \in \mathcal{S} \tag{6.1}$$

uniquely have solutions $\hat{\varsigma}_\epsilon(\cdot)$ for each $z \in \mathcal{S}$ and $\epsilon \in \{+,-\}$.

(b) Let the mappings $\varphi(\cdot, \hat{\varsigma}_\epsilon(\cdot))$ be demicontinuous.

Next, we need to assume that (2.2a) is affine in the controls; i.e.,

$$\varphi(z,v) = B(z)v + f(z) \quad \text{for each } z \in \mathcal{S} \text{ and } v \in V, \tag{6.2}$$

where f and B act in $\mathcal{S}$ and have images, respectively, in Z and $\mathcal{L}(V,Z)$.

First, we show a result on the optimization technique to be used in order to solve the problems (6.1).

LEMMA 6.1. *Let $f \in Z$ and $\mathcal{T}$ be a closed convex subset of Z. Let $B \in \mathcal{L}(V,Z)$ be a linear operator satisfying the following condition:*

$$\|B^\star \mu\|^2 \geq m\|\mu\|^2 \quad \textit{for each } \mu \in Z \textit{ with } m > 0. \tag{6.3}$$

Then the minimization problem

$$\min_{Bv+f \in \mathcal{T}} \|v\|^2 \tag{6.4}$$

has a unique solution $v_0 = -B^\star \mu_0$, where μ_0 is uniquely given by the optimality system

$$\begin{array}{l} \|B^\star \mu_0\|^2 \leq \langle \mu_0, f - y \rangle \quad \textit{for each } y \in \mathcal{T}, \\ f - BB^\star \mu_0 \in \mathcal{T}. \end{array} \tag{6.5}$$

Proof. See the appendix. □

Now we consider the following assumption.

Assumption 6.2.

(i) The mapping $f : \mathcal{S} \to Z$ is continuous.

(ii) For each sequence $(z_n)_n \subset \mathcal{S}$, $(v_n)_n \subset V$, and $(\mu_n)_n \subset Z$,

$$\begin{array}{l} z_n \to z \text{ (strong) and } v_n \to v \text{ (weak)} \implies B(z_n)v_n \to B(z)v \text{ (weak)}, \\ z_n \to z \text{ (strong) and } \mu_n \to \mu \text{ (weak)} \implies B^\star(z_n)\mu_n \to B^\star(z)\mu \text{ (weak)}. \end{array}$$

(iii) For each $z \in \mathcal{S}$, the operator $B^\star(z)$ satisfies the coercivity condition which is required in (6.3),

$$\|B^\star(z)\mu\|^2 \geq m_z\|\mu\|^2 \quad \text{for each } \mu \in Z, \tag{6.6}$$

where the coefficient $m_z > 0$ is such that, for each $\alpha > 0$, there exists $M > 0$ such that $m_z > M$ for each $z \in \mathcal{S}, \|z\| < \alpha$.

Then we are ready to examine problem $\mathbb{P}^+_{\theta,\nu}$.

THEOREM 6.3. *Let Assumptions* 3.2, 3.5, *and* 6.2 *be satisfied. In addition, suppose that*

(i) *the map* $\mathcal{T}_\omega$ *is lsc,*

(ii) τ_ω *is Gâteaux differentiable,*

(iii) $\nu \in \mathcal{A}^+_\omega$ *and is upper semicontinuous.*

Then the mapping $\hat{\varsigma}_+(\cdot)$ *of* (6.1) *stands for the unique solution of problem* $\mathbb{P}^+_{\theta,\nu}$.

Proof. By Lemma 4.4, the map $\mathcal{T}_\omega$ has closed convex values. Then the map $\mathcal{T}_\omega^{\nu^+}$ of (5.2a) also has closed convex values. This results, respectively, from Assumption 4.3 and Lemma 4.4 (ii). Therefore, all conditions of Lemma 6.1 are satisfied for each $z \in \mathcal{S}$ with $B \doteq B(z)$, $f \doteq f(z)$, $\mathcal{T} \doteq \mathcal{T}_\omega^{\nu^+}(z)$, and the coercivity condition (6.6). Hence the minimization problem (6.1) has $\hat{\varsigma}_+(z)$ as a unique solution for each $z \in \mathcal{S}$. In addition, by using Lemma 6.1, we have

$$\hat{\varsigma}_+(z) = -B^\star(z)\mu_0(z) \quad \text{for each } z \in \mathcal{S}, \tag{6.7}$$

where $\mu_0(z) \doteq \mu_0$ is uniquely determined by

$$\begin{aligned} &\|B^\star(z)\mu_0\|^2 \leq \langle \mu_0, f(z) - y\rangle \quad \text{for each } y \in \mathcal{T}_\omega^{\nu^+}(z) \text{ and } z \in \mathcal{S}, \\ &f(z) - B(z)B^\star(z)\mu_0 \in \mathcal{T}_\omega^{\nu^+}(z). \end{aligned} \tag{6.8}$$

Now it remains to show that $\hat{\varsigma}_+(\cdot)$ stands for an fsc law. According to (b) above, this holds if the mapping

$$\phi_s \doteq f + B(\cdot)\hat{\varsigma}_+ = f - B(\cdot)B^\star(\cdot)\mu_0(\cdot)$$

is demicontinuous.

Indeed, let $(z_n)_n$ be a sequence with (strong) limit $z \in \mathcal{S}$ and $y \in \mathcal{T}_\omega^{\nu^+}(z)$. Due to Lemma 5.1, the map $\mathcal{T}_\omega^{\nu^+}$ is lsc; then there exists a sequence $(y_n)_n$ which converges to y and satisfies

$$y_n \in \mathcal{T}_\omega^{\nu^+}(z_n) \quad \text{for each } n.$$

Therefore, by (6.8), we have

$$\|B^\star(z_n)\mu_0(z_n)\|^2 \leq \langle \mu_0(z_n), f(z_n) - y_n\rangle \quad \text{for each } n.$$

Consequently, since the sequence $(f(z_n))_n$ is bounded (due to Assumption 6.2 (i)), it follows that the sequence $(\mu_0(z_n))_n$ is bounded too. It therefore has a subsequence $(\mu_0(z_k))_k$ which is weakly convergent to $\bar{\mu}_0 \in Z$.

Now, since $\langle \mu_0(z_k); f(z_k) - y_k\rangle \to \langle \bar{\mu}_0; f(z) - y\rangle$ (because $f(z_k) - y_k \to f(z) - y$ strongly and $\mu_0(z_k) \to \bar{\mu}_0$ weakly), we get by passing to the lim inf in the last inequality

$$\liminf \|B^\star(z_k)\mu_0(z_k)\|^2 \leq \langle \bar{\mu}_0; f(z) - y\rangle.$$

Therefore, due to Assumption 6.2 (ii), it follows that

$$\|B^\star(z)\bar{\mu}_0\|^2 \leq \liminf \|B^\star(z_k)\mu_0(z_k)\|^2 \leq \langle \bar{\mu}_0; f(z) - y\rangle \tag{6.9}$$

for every $y \in \mathcal{T}_\omega^{\nu^+}(z)$. By using Assumption 6.2 (ii), we get

$$\phi_s(z_k) = f(z_k) - B(z_k)B^\star(z_k)\mu_0(z_k) \overset{we}{\to} f(z) - B(z)B^\star(z)\bar{\mu}_0.$$

This implies, due to Assumption 3.5 and the upper semicontinuity of ν, that

$$f(z) - B(z)B^\star(z)\bar{\mu}_0 \in \mathcal{T}_\omega^{\nu^+}(z).$$

Therefore, by (6.9), $\bar{\mu}_0$ satisfies the optimality system (6.8), and then we obtain, by uniqueness,

$$\bar{\mu}_0 = \mu_0(z) \text{ and } f(z) - B(z)B^\star(z)\bar{\mu}_0 = \phi_s(z).$$

Consequently, the sequences $(\mu_0(z_n))_n$ and $(\phi_s(z_n))_n$ are, respectively, weakly convergent to $\mu_0(z)$ and $\phi_s(z)$. Thus, as desired, the mapping ϕ_s is demicontinuous on the subset $\mathcal{S}$. □

Remark 6.4. From the proof of Theorem 6.3, it follows by Assumption 6.2 that the minimal fsc law

$$\hat{\varsigma}_\cdot = -B^\star(\cdot)\mu_0(\cdot)$$

is also demicontinuous.

Remark 6.5. The proof of Lemma 6.1 in the appendix is informative on the technique to use in order to compute $\hat{\varsigma}_\cdot$. In fact, we can use the optimality system (A.3), from which a sequence of suboptimal fsc laws can be derived by successive approximation.

Similarly, we can follow the same approach to examine problem $\mathbb{P}^-_{\theta,\nu}$.

THEOREM 6.6. *Let Assumptions* 3.2, 3.5, 5.3, *and* 6.2 *be satisfied. In addition, suppose that*

(i) *the map* $\mathcal{T}_\omega$ *is lsc,*
(ii) τ_ω *is convex or Gâteaux differentiable,*
(iii) $\nu \in \mathcal{A}_\omega^-$ *and is lsc.*

Then the mapping $\hat{\varsigma}_-(\cdot)$ *of* (6.1) *stands for the unique solution of problem* $\mathbb{P}^-_{\theta,\nu}$.

Also, note that Remarks 6.4 and 6.5 remain valid regarding problem $\mathbb{P}^-_{\theta,\nu}$.

Appendix. Proof of Lemma 6.1. Since $\mathcal{C} = \{v \in V \mid Bv + f \in \mathcal{T}\}$ is a nonempty closed convex subset in V, (6.4) has a unique solution which is $v_0 = \pi_{\mathcal{C}}(0)$. Now we can use a saddle point method to compute v_0. Define the Lagrangian functional (cf. [10, 14])

$$L(v, y, \mu) = \frac{1}{2}\|v\|^2 + \langle Bv + f - y; \mu\rangle \quad \text{for each } v \in V, y \in \mathcal{T}, \mu \in Z.$$

In fact, it can be easily shown that, if (u_0, y_0, μ_0) is a saddle point for L, i.e.,

$$\max_{\mu \in Z} L(u_0, y_0, \mu) = L(u_0, y_0, \mu_0) = \min_{v \in V, y \in \mathcal{T}} L(v, y, \mu_0),$$

then u_0 is a solution of (6.4), and, by uniqueness, $u_0 = v_0$. Now, since both L and $\mathcal{T}$ are convex, the saddle point (v_0, y_0, μ_0) is characterized by

$$\frac{\partial L}{\partial v}(v_0, y_0, \mu_0) = 0,$$

$$\left\langle \frac{\partial L}{\partial y}(v_0, y_0, \mu_0); y - y_0 \right\rangle \geq 0 \quad \text{for each } y \in \mathcal{T},$$

$$\frac{\partial L}{\partial \mu}(v_0, y_0, \mu_0) = 0$$

so that we have

$$\begin{aligned} &v_0 + B^\star \mu_0 = 0, \\ &\langle \mu_0; y - y_0 \rangle \leq 0 \quad \text{for each } y \in \mathcal{T}, \\ &Bv_0 + f = y_0 \in \mathcal{T}. \end{aligned}$$

Therefore, in an equivalent way, we get $v_0 = -B^\star \mu_0$, where μ_0 is uniquely given by the system

$$\begin{aligned} &-BB^\star \mu_0 + f = y_0, \\ &\langle \mu_0, y - y_0 \rangle \leq 0 \quad \text{for each } y \in \mathcal{T}, \\ &y_0 \in \mathcal{T}, \end{aligned} \tag{A.1}$$

which is equivalent to

$$\begin{aligned} &\|B^\star \mu_0\|^2 \leq \langle \mu_0, f - y \rangle \quad \text{for each } y \in \mathcal{T}, \\ &f - BB^\star \mu_0 \in \mathcal{T}. \end{aligned} \tag{A.2}$$

Now it remains to show that such a μ_0 exists. In fact, by multiplying by $\rho > 0$ in (A.1) and using the operator of best approximation $\pi_{\mathcal{T}}$, we obtain the equivalent system

$$\begin{aligned} &v_0 = B^\star R^{-1}(y_0 - f), \\ &y_0 = \pi_{\mathcal{T}}[(1 - \rho R^{-1})y_0 + \rho R^{-1} f] \end{aligned} \tag{A.3}$$

for some $\rho > 0$, where the operator $R = BB^\star$. Then we are led to seek a fixed point of the mapping

$$\begin{aligned} \Theta_\rho : \quad \mathcal{T} &\to \mathcal{T} \\ y &\to \pi_{\mathcal{T}}[(1 - \rho R^{-1})y + \rho R^{-1} f]. \end{aligned}$$

Indeed, we have

$$\begin{aligned} \|\Theta_\rho(y) - \Theta_\rho(\bar{y})\|^2 &\leq \|(1 - \rho R^{-1})e\|^2 \\ &= \|e\|^2 - 2\rho \langle R^{-1} e; e \rangle + \rho^2 \|R^{-1} e\|^2, \end{aligned}$$

where $y, \bar{y} \in \mathcal{T}$, and $e = y - \bar{y}$.

Since the operator R^{-1} is coercive, we have, for some $m' > 0$,

$$\langle R^{-1} y; y \rangle \geq m' \|y\|^2 \quad \text{for each } y \in Z.$$

It follows that

$$\|\Theta_\rho(y) - \Theta_\rho(\bar{y})\|^2 \leq (1 - 2\rho m' + \rho^2 \|R^{-1}\|^2) \|y - \bar{y}\|^2.$$

Therefore, Θ_ρ is a contraction for $\rho < 2m'/\|R^{-1}\|^2$, and thereby it has a unique fixed point y_0, which belongs to $\mathcal{T}$. This ends the proof of Lemma 6.1.

Acknowledgments. The author is grateful to the unknown reviewers for their careful reading of the original manuscript, constructive criticism, and helpful suggestions.

REFERENCES

[1] J. P. Aubin and A. Cellina, *Differential Inclusions*, Birkhäuser Boston, Boston, 1991.
[2] A. V. Balakrishnan, *Applied Functional Analysis*, Springer-Verlag, New York, 1981.
[3] R. Curtain and A. J. Pritchard, *Functional Analysis in Modern Applied Mathematics*, Academic Press, London, 1977.
[4] I. Chiş-Şter, *Existence of monotone solutions for a parabolic problem*, An. Ştiinţ. Univ. Al. I. Cuza Iaşi. Mat. (N.S.), 43 (1997), pp. 403–414.
[5] K. Deimling, *Multivalued Differential Equations*, Walter de Gruyter, Berlin, 1992.
[6] A. El Jai and K. Kassara, *Spreadable distributed systems*, Math. Comput. Modelling, 20 (1994) pp. 47–64.
[7] A. El Jai and K. Kassara, *Spreadability of transport systems*, Internat. J. Systems Sci., 27 (1996), pp. 681–688.
[8] A. El Jai, K. Kassara, and O. Cabrera, *Spray control*, Internat. J. Control, 68 (1997), pp. 709–730.
[9] S. El Yacoubi, A. El Jai, and J. Karrakchou, *Spreadability and spray actuators*, Appl. Math. Comput. Sci., 8 (1998), pp. 367–379.
[10] M. Fortin and R. Glowinski, *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*, North–Holland, Amsterdam, 1983.
[11] J. Jahn, *Introduction to the Theory of Nonlinear Optimization*, Springer-Verlag, New York, 1983.
[12] K. Kassara, *Feedback spreading controls for semilinear parabolic systems*, J. Comput. Appl. Math., 114 (2000), pp. 41–54.
[13] K. Kassara, *Feedback spreading control laws for semilinear distributed parameter systems*, Systems Control Lett., 40 (2000), pp. 269–276.
[14] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Dunod, Paris, 1976.
[15] M. Z. Nashed, *Differentiability and related properties of nonlinear operators*, in Nonlinear Functional Analysis and Applications, L. B. Rall, ed., Academic Press, New York, 1970, pp. 103–309.
[16] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
[17] S. Shuzhong, *Viability theorems for a class of differential-operator inclusions*, J. Differential Equations, 79 (1989) pp. 232–257.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 4, pp. 1295–1314

ON THE OBSERVABILITY AND DETECTABILITY OF CONTINUOUS-TIME MARKOV JUMP LINEAR SYSTEMS*

EDUARDO F. COSTA† AND JOÃO B. R. DO VAL†

Abstract. The paper introduces a new detectability concept for continuous-time Markov jump linear systems with finite Markov space that generalizes previous concepts found in the literature. The detectability in the weak sense is characterized as mean square detectability of a certain related stochastic system, making both detectability senses directly comparable. The concept can also ensure that the solution of the coupled algebraic Riccati equation associated to the quadratic control problem is unique and stabilizing, making other concepts redundant. The paper also obtains a set of matrices that plays the role of the observability matrix for deterministic linear systems, and it allows geometric and qualitative properties. Tests for weak observability and detectability of a system are provided, the first consisting of a simple rank test, similar to the usual observability test for deterministic linear systems.

Key words. Markov jump systems, detectability and observability of stochastic systems, optimal control, stochastic systems, quadratic control

AMS subject classifications. 93E03, 93E20, 93B07, 34A30, 60J05

PII. S0363012901385460

1. Introduction.

The concepts of observability and detectability play an important role in the theory of dynamic systems. For instance, in optimal control problems, these concepts provide a connection between closed-loop stability and finiteness of the cost functional, and they ensure uniqueness of the solution to the algebraic Riccati equation. This is the scenario in the theory of deterministic time-invariant linear systems (see [14]), deterministic linear time-varying systems (see [1], [2] or [11]), and to some extent, in Markov jump linear systems (MJLS) (see [7], [9], [13], [16], and [17]).

Thanks to those developments, a number of well-established results concerning detectability and the good behavior of solutions of filtering and control problems exist today which can be found in a literature that spans more than four decades. Among the results we refer to concerning linear time-invariant deterministic systems are the following: (I) invariance of nonobserved trajectories, (II) existence of a simple rank-test condition for observability, (III) correspondence between nonobserved trajectories and stable modes of detectable systems, and (IV) relationship between observability and detectability. However, it was not known to this date how properties (I)–(IV) extend to MJLS.

Consider the continuous-time MJLS written as

$$\Phi : \begin{cases} \dot{x}(t) = A_{\theta(t)}x(t), & t \geq 0, \\ y(t) = C_{\theta(t)}x(t), & x(0) = x_0, \quad \theta(0) = \theta_0, \end{cases} \tag{1}$$

defined in a fundamental probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$, where $\mathcal{F}_t$ denotes the σ-field generated by $\{x(s), \theta(s), 0 \leq s \leq t\}$. The variables $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^q$ are the

*Received by the editors February 22, 2001; accepted for publication (in revised form) May 21, 2002; published electronically December 3, 2002. Research supported in part by FAPESP grant 98/13095-8, CNPq grant 300721/86-2(RN), PRONEX grant 015/98 "Control of Dynamical Systems," and IM-AGiMB.

http://www.siam.org/journals/sicon/41-4/38546.html

†UNICAMP - Fac. de Engenharia Elétrica e de Computação, Depto. de Telemática, C.P. 6101, CEP 13081-970, Campinas, SP, Brazil (eduardoc@dt.fee.unicamp.br, jbosco@dt.fee.unicamp.br).

continuous state and the output, respectively; x_0 is a second order random variable. The mode θ is the state of an underlying continuous-time homogeneous Markov chain $\Theta = \{\theta(t); t \geq 0\}$ having $\mathcal{S} = \{1, \ldots, N\}$ as state space and $\Lambda = [\lambda_{ij}]$, $i, j = 1, \ldots, N$, as the transition rate matrix. The initial distribution of Θ is determined by $\mu_i = P(\theta_0 = i)$, $i = 1, \ldots, N$. Matrices A_i and C_i, $1 \leq i \leq N$, belong to the collections of N real matrices: $A = (A_1, \ldots, A_N)$, $\dim(A_i) = n \times n$, and $C = (C_1, \ldots, C_N)$, $\dim(C_i) = q \times n$. Consider also the functional

$$W^t(x, \theta) = E\left\{\int_0^t x(\tau)' C'_{\theta(\tau)} C_{\theta(\tau)} x(\tau) d\tau | \mathcal{F}_0\right\} \tag{2}$$

defined for $x(0) = x$ and $\theta(0) = \theta$. Here we consider the following concept of observability, which is drawn from the observability concept for time-variant MJLS that appears, for instance, in [16]. The concept is more general than other observability concepts for MJLS, like the ones in [13].

Definition 1 (W-observability). *We say that (A, C, Λ) is weakly (W-) observable when there exist scalars $t_d \geq 0$ and $\gamma > 0$ such that $W^{t_d}(x, \theta) \geq \gamma |x|^2$ for each $x \in \mathbb{R}^n$ and $\theta \in \mathcal{S}$.*

In this paper, for the time-invariant system Φ, we present a collection of matrices $\mathcal{O} = (\mathcal{O}_1, \ldots, \mathcal{O}_N)$ associated to the W-observability concept that resembles observability matrices of deterministic linear systems. Then we provide extensions of properties (I) and (II) mentioned above, respectively: we show that nonobserved trajectories are invariant in the sense that, if $x(s)$ is in the kernel of $\mathcal{O}_{\theta(s)}$ for some $s \geq 0$, then $x(t)$ is in the kernel of $\mathcal{O}_{\theta(t)}$ for any $t \geq s$ (see Corollary 15), and we show that (A, C, Λ) is W-observable if and only if each of the matrices of the set $\mathcal{O}$ is of full rank. We also demonstrate that the largest attainable dimensionality of $\mathcal{O}$ is constrained by the system dimensions n and N (see Lemma 12) in a similar manner to observability matrices of deterministic systems.

Regarding the detectability of MJLS, before the work in [4] for discrete-time MJLS, the most general detectability concept available was the dual of the stabilizability concept, known as mean square (MS-) detectability (see [7], [9], or [16]). The concept is as follows.

Definition 2 (MS-stability). *We say that (A, Λ) is MS-stable if, for system Φ and for each $x_0 \in \mathbb{R}^n$ and $\theta_0 \in \mathcal{S}$,*

$$\lim_{t \to \infty} E\{|x(t)|^2\} = 0.$$

Definition 3 (MS-detectability). *We say that (A, C, Λ) is MS-detectable when there exists $G = \{G_1, \ldots, G_N\}$ of appropriate dimension for which $(A - GC, \Lambda)$ is MS-stable.*

In connection with the MS-detectability concept, we have that none of the well-known properties (III) and (IV) mentioned above hold. Moreover, W-observability is not comparable to MS-detectability. In Example 2, we present a system that is W-observable but is not MS-detectable. It is also simple to provide a converse example: if one takes (A, Λ) as MS-stable and $C = 0$, one has that (A, C, Λ) is MS-detectable but is not W-observable. This lack of structure sometimes compels authors to consider either a detectability or an observability hypothesis (see, for example, [9] and [16]), where these conditions appear as sufficient conditions for uniqueness of solutions to coupled algebraic Riccati equations (CAREs) arising in the optimal linear quadratic problem.

In this paper, we develop the following associate concept of W-detectability from the W-observability concept. We mention that it is analogous to a concept for time-varying systems that appears in [1].

DEFINITION 4 (W-detectability). *We say that* (A, C, Λ) *is W-detectable if there exist scalars* $t_d, s_d \geq 0$, $\gamma > 0$, *and* $0 \leq \delta < 1$ *such that* $W^{t_d}(x_0, \theta_0) \geq \gamma |x_0|^2$ *whenever* $E\{|x(s_d)|^2\} \geq \delta |x_0|^2$.

We show that W-detectability generalizes and can retrieve each of the properties (III) and (IV), respectively: for every nonobserved trajectory, a contraction condition holds, ensuring that the trajectory converges in the MS sense (see Lemma 20); W-detectability generalizes W-observability. Moreover, in one of the main results of this paper, we characterize W-detectability by means of MS-detectability as follows: (A, C, Λ) is W-detectable if and only if $(A, \mathcal{O}, \Lambda)$ is MS-detectable (see Theorem 24). This result allows us to clarify the conservativeness of MS-detectability when compared with W-detectability, and, at same time, it provides a testable condition for W-detectability; see section 4.1.

For the controlled MJLS, we show that the W-detectability concept ensures that finite cost implies stable trajectories in the MS sense and, in particular, that the solution to the CARE arising in optimal control problems is unique and stabilizing; see section 5. This result generalizes previous characterizations in [9], [13], [16], and [17].

The paper is organized as follows. In section 2 basic results and relevant definitions are introduced, and, in section 3, we introduce the observability matrices and related properties. In section 4, some characterizations of W-detectability are presented, and, in section 4.1, it is shown that W-detectability generalizes MS-detectability. In section 5, we set up the link between W-detectability and stabilizing quadratic control.

2. Notation, concepts, and basic results. Let $\mathbb{R}^n$ be the nth dimensional Euclidean space. Let $\mathcal{R}^{n,q}$ (respectively, $\mathcal{R}^n$) represent the normed linear space formed by all $n \times q$ (respectively, $n \times n$) real matrices and $\mathcal{R}^{n0}$ ($\mathcal{R}^{n+}$) the closed convex cone $\{U \in \mathcal{R}^n : U = U' \geq 0\}$ (the open cone $\{U \in \mathcal{R}^n : U = U' > 0\}$), where U' denotes the transpose of U; $U \geq V$ ($U > V$) signifies that $U - V$ is positive semidefinite (definite). For $U \in \mathcal{R}^{n,q}$, $\mathcal{N}\{U\}$ and $\mathcal{R}\{U\}$ represent the kernel and the range of U, respectively.

Let $\mathcal{M}^{n,q}$ denote the linear space formed by a number N of matrices such that $\mathcal{M}^{n,q} = \{U = (U_1, \ldots, U_N) : U_i \in \mathcal{R}^{n,q}, i = 1, \ldots, N\}$; also, $\mathcal{M}^n \equiv \mathcal{M}^{n,n}$. We denote by $\mathcal{M}^{n0}$ ($\mathcal{M}^{n+}$) the set $\mathcal{M}^n$ when it is made up by some $U_i \in \mathcal{R}^{n0}$ ($U_i \in \mathcal{R}^{n+}$) for all $i = 1, \ldots, N$. Analogously, for $U, V \in \mathcal{M}^{n0}$ $U \geq V$ ($U > V$) signifies that $U - V \in \mathcal{M}^{n0}$ ($U - V \in \mathcal{M}^{n+}$). It is known that $\mathcal{M}^{n,q}$ equipped with the inner product

$$\langle U, V \rangle = \sum_{j=1}^{N} \operatorname{tr}\{U_j' V_j\}$$

forms a Hilbert space. Let us define the norm $\|U\| = \langle U, I \rangle$ on $\mathcal{M}^{n0}$.

Consider system Φ in (1). For $i = 1, \ldots, N$, we define

$$X_i(t) = E\{x(t)x(t)' 1_{\{\theta(t)=i\}} | \mathcal{F}_0\}, \quad t \geq 0. \tag{3}$$

With this notation, we can write, for instance, $E\{|x(t)|^2 | \mathcal{F}_0\} = \langle X(t), I \rangle = \|X(t)\|$.

Now let us introduce the operators $\mathcal{L} : \mathcal{M}^n \to \mathcal{M}^n$ and their adjoint in the inner

product sense $\mathcal{T} : \mathcal{M}^n \to \mathcal{M}^n$ as

$$\mathcal{L}_i(U) = A_i'U_i + U_iA_i + \sum_{j=1}^{N} \lambda_{ij}U_j, \tag{4a}$$

$$\mathcal{T}_i(U) = A_iU_i + U_iA_i' + \sum_{j=1}^{N} \lambda_{ji}U_j, \quad i = 1, \ldots, N. \tag{4b}$$

Let also $L(t)$ and $U(t)$, $t \geq 0$, be defined by the matrix linear differential equations

$$\dot{L}_i(t) := \mathcal{L}_i(L(t)) + C_i'C_i, \quad L(0) = 0,\ t \geq 0, \tag{5a}$$

$$\dot{U}_i(t) := \mathcal{T}_i(U(t)), \quad U(0) = U \in \mathcal{M}^{n0}, \tag{5b}$$

for each $i = 1, \ldots, N$. The operators $\mathcal{L}$ and $\mathcal{T}$ are linear, and $L(t)$ and $U(t)$ defined by (5) are unique. The following results are adapted from [6] and [13]; the proof is omitted.

PROPOSITION 5. *The following assertions hold:*

(i)

$$\dot{X}_i(t) = \mathcal{T}_i(X(t)), \quad t \geq 0, \quad i = 1, \ldots, N, \tag{6}$$

for $X(0) \in \mathcal{M}^{n0}$*, such that* $X_i(0) = x_0x_0'1_{\{\theta(0)=i\}}$, $i = 1, \ldots, N$*;*

(ii)

$$W^t(x, i) = \int_0^t \langle X(\tau), C'C\rangle d\tau = \langle X(0), L(t)\rangle. \tag{7}$$

Consider the corresponding generalization of (7)

$$W^t(U) = \int_0^t \langle U(\tau), C'C\rangle d\tau = \langle U, L(t)\rangle, \tag{8}$$

where $L(\cdot)$ and $U(\cdot)$ are given by (5).

LEMMA 6. $U(\cdot) \in \mathcal{M}^{n0}$ *and* $L(s) \geq L(t)$ *whenever* $s \geq t$.

Proof. Notice that, for any $U \in \mathcal{M}^{n0}$, one can adopt the following representation (cf. Theorem 7.5.2 of [12]):

$$U_i = x_i^1x_i^{1\prime} + \cdots + x_i^{r_i}x_i^{r_i\prime},$$

where $x_i^k \in \mathbb{R}^n, k = 1, \ldots, r$ and $r_i = \operatorname{rank}(U_i) \leq n$. In connection, we can define $X_i^{j,k}(\cdot)$ as the solution of (6) with $X_i^{i,k}(0) = x_i^kx_i^{k\prime}$; it is clear from the second moment definition in (3) that $X_i^{j,k}(\cdot) \in \mathcal{M}^{n0}$. Also, from the linearity of the operator $\mathcal{T}$, we have that

$$U_i(t) = \sum_{j=1}^{N}\sum_{k=1}^{r_j} X_i^{j,k}(t), \quad t \geq 0,$$

and $U(\cdot) \in \mathcal{M}^{n0}$, which proves the first assertion.

From the expression (8) and the first assertion, it is simple to check the result for $L(\cdot)$. In fact, whenever $s \geq t$, one has that $W^s(U) \geq W^t(U)$ for each $U \in \mathcal{M}^{n0}$, and thus $\langle U, L(s) - L(t) \rangle \geq 0$. The assertion follows from the Fejer's trace theorem; cf. [12]. □

It is well known that the MS-stability of A is equivalent to the requirement that $\mathrm{Re}\{\lambda(\mathcal{T})\} < 0$; see, for instance, [6]. Then we can rewrite the MS-stability concept as follows.

DEFINITION 7 (MS-stability). *We say that* (A, Λ) *is MS-stable if*

$$\lim_{t \to \infty} \|X(t)\| = 0 \quad \forall X \in \mathcal{M}^{n0}.$$

Remark 1. Feng et al. in [10] have shown that the MS-stability concept is equivalent to other second moment stability concepts, such as exponential stability. Thus the system is MS-stable if and only if there exist $0 < \xi < 1$ and $\alpha \geq 1$ such that $\|X(t)\| \leq \alpha \xi^t \|X(0)\|$ for every $X(0) \in \mathcal{M}^{n0}$. It is also known that, if (A, Λ) is not MS-stable, then there exists $X(0) \in \mathcal{M}^{n0}$ such that $\|X(t)\| \geq \beta \zeta^t \|X(0)\|$ for some $\zeta \geq 1$ and $0 < \beta \leq 1$.

3. W-observability and observability matrices. In this section, we introduce a collection of observability matrices, and, in one of the main results, we derive a test for observability based on the rank of these matrices, in a parallel with the observability test for deterministic linear systems. We also derive a counterpart for MJLS for the well-known result for linear deterministic systems that nonobserved trajectories are invariant. An illustrative example is also provided.

Let us introduce the matrices $\mathcal{O}_i \in \mathcal{R}^{n(n^2N),n}$, defined for each $i = 1, \ldots, N$, as

$$\mathcal{O}_i := [O_i(0)\, O_i(1) \cdots O_i(n^2N - 1)]', \tag{9}$$

where each matrix $O_i(\cdot)$ belongs to the sequence of matrices on $\mathcal{M}^{n0}$ defined as

$$O_i(k) := \mathcal{L}_i(O(k-1)), \quad k > 0, \tag{10}$$

with $O_i(0) := C_i' C_i$, for each $i = 1, \ldots, N$. Notice by inspection of (5a) that

$$O_i(k) = \frac{d^{k+1} L}{dt^{k+1}}(0). \tag{11}$$

The collection of matrices $\mathcal{O} \in \mathcal{M}^{n0}$ is called the set of observability matrices of system Φ. In fact, $\mathcal{O}$ resembles the observability matrices of linear deterministic systems in many aspects, as we shall see in this section. We can mention in passing that, for an isolated Markov state i, namely, $\lambda_{ji} = 0$, $j = 1, \ldots, N$, a direct equivalence is retrieved: the pair (A_i, C_i) is observable in the deterministic sense if and only if $\mathcal{O}_i$ is a full rank matrix.

Next we present some preliminary results.

For $V \in \mathcal{R}^n$, let us identify the columns of $V = [v_1 \vdots v_2 \vdots \cdots \vdots v_n]$. For $U = (U_1, \ldots, U_N)$ and following [5], we introduce the linear and invertible operator $\widehat{\varphi} : \mathcal{M}^{n0} \to \mathbb{R}^{n^2 N}$ as

$$\widehat{\varphi}(U) = \begin{bmatrix} \varphi(U_1) \\ \vdots \\ \varphi(U_N) \end{bmatrix}, \quad \text{where} \quad \varphi(V) = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}.$$

Let $V \otimes Z$ represent the Kronecker tensor product of matrices V and Z. From (4a), using basic properties of the Kronecker product [3], we obtain

$$\varphi(\mathcal{L}_i(U)) = (I_n \otimes A_i')\varphi(U_i) + (A_i' \otimes I_n)\varphi(U_i) + \sum_{j=1}^{N} \lambda_{ij}\varphi(U_j),$$

and one can check that

$$\widehat{\varphi}(\mathcal{L}(U)) = \mathcal{A}\widehat{\varphi}(U), \tag{12}$$

where $\mathcal{A} \in \mathcal{R}^{n^2N}$ is the matrix defined by

$$\begin{bmatrix} \hat{A}_1 + \lambda_{11}I_{n^2} & \lambda_{12}I_{n^2} & \cdots & \lambda_{1N}I_{n^2} \\ \lambda_{21}I_{n^2} & \hat{A}_2 + \lambda_{22}I_{n^2} & & \\ \vdots & & \ddots & \\ \lambda_{N1}I_{n^2} & & & \hat{A}_N + \lambda_{NN}I_{n^2} \end{bmatrix}$$

and $\hat{A}_i = (I_n \otimes A_i' + A_i' \otimes I_n)$. Applying the operator $\widehat{\varphi}$ in (5a) and employing (12), we obtain

$$\begin{aligned} \dot{\ell}(t) &= \widehat{\varphi}[C'C + \mathcal{L}(L(t))] \\ &= q + \mathcal{A}\widehat{\varphi}(L(t)) = q + \mathcal{A}\ell(t), \quad t \geq 0, \end{aligned} \tag{13}$$

where $\ell(t) \in \mathbb{R}^{n^2N}$ and $q \in \mathbb{R}^{n^2N}$ are defined by

$$\ell(t) = \widehat{\varphi}(L(t)), \quad q = \widehat{\varphi}(C'C).$$

Notice by inspection of (13) that

$$\frac{d^k\ell(0)}{dt^k} = \mathcal{A}^k q. \tag{14}$$

We also introduce the following representation for the expression $\langle U, L(t)\rangle$:

$$\langle U, L(t)\rangle = \widehat{\varphi}(U)'\ \ell(t). \tag{15}$$

LEMMA 8. *Consider $x \in \mathbb{R}^n$ and $i \in \mathcal{S}$; define $X \in \mathcal{M}^{n0}$ as $X_i = xx'$ and $X_j = 0$ for all $j \neq i$. Set $w \in \mathbb{R}^{n^2N}$ as $w = \hat{\varphi}(X)$. The following assertions are equivalent:*

(i) *$x'L_i(s)x = 0$ or, equivalently, $w'\ell(s) = 0$ for some $s > 0$;*

(ii) *$w'd^m\ell/dt^m(0) = 0$ for $m = 1, \ldots, n^2N$;*

(iii) *$w'\mathcal{A}^{m-1}q = 0$ for $m = 1, \ldots, n^2N$;*

(iv) *$x \in \mathcal{N}(L_i(t))$ or, equivalently, $w'\ell(t) = 0$ for all $t \geq 0$;*

(v) *$x \in \mathcal{N}(\mathcal{O}_i)$.*

Proof. (i) $\Rightarrow$ (ii): From Lemma 6, $L(t) \leq L(s)$ for $t \leq s$; from (15), we evaluate $w'\ell(t) = \langle\widehat{\varphi}^{-1}(w), L(t)\rangle \leq \langle\widehat{\varphi}^{-1}(w), L(s)\rangle = w'\ell(s) = 0$, $t \leq s$. In addition, noticing that $w'\ell(t) \geq 0$ and recalling that $L(0) = 0$, we can write $w'\ell(t) = 0$ for all $0 \leq t \leq s$, which leads to

$$w'\frac{d^m\ell}{dt^m}(0) = 0 \quad \forall m \geq 0.$$

(ii) ⇒ (iii): The result follows immediately from (14).

(iii) ⇒ (iv): For the linear deterministic system in (13), we can write for any $t \geq 0$

$$\ell(t) = \int_0^t e^{\mathcal{A}(t-\tau)} q d\tau = \int_0^t \sum_{m=1}^{n^2N} \alpha_m(\tau) \mathcal{A}^{m-1} q d\tau$$
$$= \sum_{m=1}^{n^2N} \mathcal{A}^{m-1} q \int_0^t \alpha_m(\tau) d\tau = \sum_{m=1}^{n^2N} \hat{\alpha}_m(t) \mathcal{A}^{m-1} q,$$

where α_m and $\hat{\alpha}_m$ are scalar functions. Then we get that

$$w'\ell(t) = \sum_{m=1}^{n^2N} \hat{\alpha}_m(t) w' \mathcal{A}^{m-1} q = 0.$$

(iv) ⇒ (i): This part of the proof is trivial.

(ii) ⇔ (v): Employing (15), we write

$$w' \frac{d^m \ell}{dt^m}(0) = 0 \Leftrightarrow \left\langle X, \frac{d^m L}{dt^m}(0) \right\rangle = 0 \Leftrightarrow \frac{d^m L_i}{dt^m}(0) x = 0 \tag{16}$$

for $m = 1, \ldots, n^2N$. The proof is easily completed by noticing from (11) that

$$\mathcal{O}_i = \begin{bmatrix} O_i(0) \\ \vdots \\ O_i(n^2N-1) \end{bmatrix} = \begin{bmatrix} \frac{d^1 L_i}{dt}(0) \\ \vdots \\ \frac{d^{n^2N} L_i}{dt^{n^2N}}(0) \end{bmatrix}. \qquad \square$$

The next corollary restates some assertions in Lemma 8 for further use.

COROLLARY 9. *The following assertions are equivalent:*

(i) $x \in \mathcal{N}\{\mathcal{O}_i\}$;

(ii) $W^s(x,i) = 0$ *for some* $s > 0$;

(iii) $W^t(x,i) = 0$ *for all* $t \geq 0$.

Remark 2. Notice from Corollary 9 that if the conditions in Definitions 1 or 4 hold for some $t_d \geq 0$, then they hold for all $t \geq 0$.

The next theorem provides a rank test on the set of observability matrices $\mathcal{O}$. First, let us rewrite the W-observability concept in terms of the notation introduced in section 2.

DEFINITION 10 (W-observability). *We say that* (A, C, Λ) *is W-observable when there exist scalars* $t_d \geq 0$ *and* $\gamma > 0$ *such that* $W^{t_d}(X) \geq \gamma \|X\|$ *for each initial condition* X.

THEOREM 11. *Consider system* Φ. (A, C, Λ) *is W-observable if and only if* $\mathcal{O}_i$ *has full rank for each* $i = 1, \ldots, N$.

Proof. From (8), we can write the condition in Definition 1 equivalently as

$$\langle X, L(t_d) \rangle \geq \gamma \|X\| \quad \forall X \in \mathcal{M}^{n0}.$$

This is equivalent to requiring that $L_i(t_d)$ be positive definite for each $i = 1, \ldots, N$. The equivalencies (i) and (v) of Lemma 8 complete the proof. $\square$

Example 1. Let $N = 2$, $n = 2$, and set

$$A_1 = I_2;\ A_2 = \begin{bmatrix} 1 & 1 \\ 0 & 1; \end{bmatrix};\ C_1 = [1\ 0];\ C_2 = 0;\ \Lambda = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}.$$

From (9), one evaluates $\operatorname{rank}(\mathcal{O}_1) = \operatorname{rank}(\mathcal{O}_2) = 2$, and Theorem 11 ensures that (A, C, Λ) is W-observable.

Remark 3. It is known that (A, C, Λ) is W-observable if each pair (A_i, C_i), $i = 1, \ldots, N$, is observable; see, e.g., [16]. However, this condition is not necessary; for instance, in Example 1, none of the pairs (A_i, C_i) are observable.

3.1. Properties of the observability matrices and pathwise invariance of nonobserved trajectories. The next lemma establishes a counterpart for the well-known result about the largest attainable dimensionality of observability matrices.

LEMMA 12.

$$\mathcal{N}\{\mathcal{O}_i\} = \mathcal{N}\{[O_i(0)\vdots\cdots\vdots O_i(k)]'\} \quad \forall k \geq n^2N - 1.$$

Proof. For $x \in \mathcal{N}(\mathcal{O}_i)$, let $X_i = xx'$ and $X_j = 0$ for all $j \neq i$, and let $w = \widehat{\varphi}(X)$. From Lemma 8 (iii) and (v), we have that $\mathcal{O}_i x = 0$ is equivalent to $w'\mathcal{A}^{r-1}q = 0$, $r = 1, \ldots, n^2N$. From the Cayley–Hamilton lemma, $\mathcal{A}^m = \sum_{r=0}^{n^2N-1} \alpha_r \mathcal{A}^r$ for each $m \geq 0$, and we obtain

$$\begin{cases} w'q = 0, \\ w'\mathcal{A}q = 0, \\ \vdots \\ w'\mathcal{A}^m q = 0, \end{cases} \tag{17}$$

which, from (14), is equivalent to $w'd^m\ell(0)/dt^m = 0$, $m \geq 0$. Finally, applying (16) for a generic $m \geq 0$, we obtain $d^m L_i(0)/dt^m x = 0$, and from (11) we write $O_i(m)x = d^{m+1}L_i(0)/dt^{m+1}x = 0$ for $m \geq 0$ and, in particular, for $m \geq n^2N - 1$. Thus $\mathcal{N}\{\mathcal{O}_i\} \subset \mathcal{N}\{[O_i(0)\vdots\cdots\vdots O_i(k)]'\}$ for all $k \geq n^2N - 1$; the opposite relation holds trivially. ∎

Next we present a relation between the null spaces of the observability matrices which will be useful in what follows. The following preliminary result is needed.

PROPOSITION 13. *For each scalar $M > 0$, there exists $t_M > 0$ for which $\|x(t) - x_0\| \leq M\|x_0\|$ almost surely (a.s.) for all $t \leq t_M$.*

LEMMA 14. *Assume that the Markov state j is accessible from the state i. Then $\mathcal{N}\{\mathcal{O}_i\} \subset \mathcal{N}\{\mathcal{O}_j\}$.*

Proof. Let us deny the assertion of the lemma; that is, we assume that there exist a scalar $m > 0$ and $x_0 \in \mathbb{R}^n$ such that

$$x_0 \in \mathcal{N}\{\mathcal{O}_i\} \tag{18}$$

for which $|x_0 - x| \geq m$, for all $x \in \mathcal{N}\{\mathcal{O}_j\}$. Notice that $x_0 \neq 0$, and let x_0 and $\theta_0 = i$ be initial conditions.

We start the proof by setting $M = m/(2|x_0|)$ in Proposition 13 to obtain that there exists t_M for which $x(t) \in B_{m/2}(x_0)$, $t \leq t_M$, where $B_{m/2}(x_0) = \{x : |x - x_0| \leq m/2\}$. Let $\tilde{x}(t_M)$ and $\hat{x}(t_M)$ be the orthogonal projection of $x(t_M)$ on $\mathcal{N}\{\mathcal{O}_j\}$ and $\mathcal{R}\{\mathcal{O}_j'\}$, respectively. Notice that $\tilde{x}(t_M) \perp \hat{x}(t_M)$ and $|\tilde{x}(t_M)| \geq m/2$; see Figure 1.

From Lemma 8 (i) and (v), one has that $\mathcal{N}\{O_j\} = \mathcal{N}\{L_j(s)\}$ for $s \geq 0$, and thus $\mathcal{R}\{\mathcal{O}_j'\} = \mathcal{R}\{L_j'(s)\}$. In this situation, we can write

$$x(t_M)'L_j(s)x(t_M) = \hat{x}(t_M)'L_j(s)\hat{x}(t_M) \geq \mu|\hat{x}(t_M)|^2 \geq \mu(m/2)^2,$$

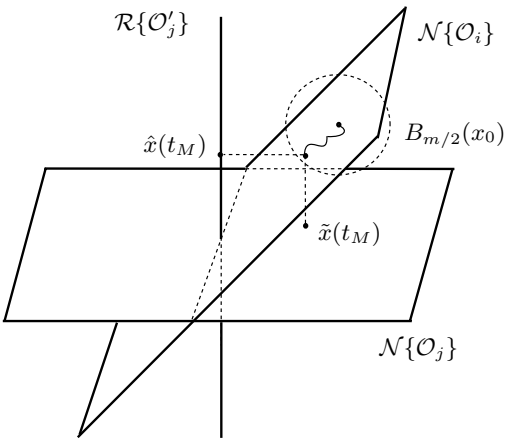


FIG. 1. *The geometry of Lemma* 14.

where μ is the smallest strictly positive eigenvalue of $L_j(s)$, and Proposition 5 leads to

$$W^s(x(t_M), j) \geq \mu m^2/4 \quad \text{a.s.}$$

Now we evaluate

$$\begin{aligned} E\{W^s(x(t_M), \theta(t_M))|\mathcal{F}_0\} &\geq E\{W^s(x(t_M), \theta(t_M))1_{\{\theta(t_M)=j\}}|\mathcal{F}_0\} \\ &\geq \frac{\mu m^2}{4} E\{1_{\{\theta(t_M)=j\}}|\mathcal{F}_0\} > 0, \end{aligned}$$

where the last inequality comes from the assumption of the lemma. Finally, we can write that

$$\begin{aligned} W^{s+t_M}(x_0, \theta_0) &= E\{W^{s+t_M}(x_0, \theta_0)|\mathcal{F}_0\} \\ &= E\{W^{t_M}(x_0, \theta_0)|\mathcal{F}_0\} + E\{W^s(x(t_M), \theta(t_M))|\mathcal{F}_0\} \\ &\geq E\{W^s(x(t_M), \theta(t_M))|\mathcal{F}_0\} > 0, \end{aligned}$$

and, from Corollary 9 (i) and (iii), it follows that $x_0 \notin \mathcal{N}\{\mathcal{O}_{\theta_0}\}$, which is a contradiction in view of (18). □

The next corollary establishes that nonobserved trajectories are pathwise invariant.

COROLLARY 15. *If* $x(t) \in \mathcal{N}\{\mathcal{O}_{\theta(t)}\}$, *then* $x(s) \in \mathcal{N}\{\mathcal{O}_{\theta(s)}\}$ *a.s. for all* $s \geq t$.

4. W-detectability. Let us start this section by rewriting the concept of W-detectability in terms of the notation introduced in section 2.

DEFINITION 16 (W-detectability). *We say that* (A, C, Λ) *is W-detectable if there exist scalars* $t_d, s_d \geq 0$, $\gamma > 0$, *and* $0 \leq \delta < 1$ *such that* $W^{t_d}(X) \geq \gamma \|X\|$ *whenever* $\|X(s_d)\| \geq \delta \|X\|$, *with* $X(0) = X$.

Notice that W-detectability requires positivity of $W^{t_d}(\cdot)$ only when the condition $\|X(s_d)\| \geq \delta \|X\|$, related to stability of the system, is satisfied. The next result is immediate; the proof is omitted.

LEMMA 17. *If* (A, C, Λ) *is W-observable, then* (A, C, Λ) *is W-detectable.*

The concept of W-detectability resembles standard concepts of detectability for linear discrete time-varying systems; see, e.g., [1] or [11]. As we shall see in Lemma 19, the concept retrieves the idea that every nonobserved state corresponds to stable modes of the system. Notice that every MS-stable MJLS is W-detectable with t_d and γ arbitrary and δ and s_d such that $\delta = \alpha\xi^{s_d} < 1$, where α and ξ are as in Remark 1.

In what follows, basic properties of W-detectability are derived. We start with some properties of the functional $W^t(X)$. In this section, the initial condition $X(0) \in \mathcal{M}^{n0}$ is denoted by X; $W^t(X(s)) = W^t(X)$ whenever $X(s) = X$.

LEMMA 18. *Let $T > 0$. The following assertions hold:*

(i) *$W^t(X)$ is continuous on X.*

(ii) *Assume that $W^T(X) = 0$ for some $T \geq 0$; then $W^t(X(s)) = 0$ for all $t, s \geq 0$.*

Proof. (i) The assertion follows immediately from the representation in (8), $W^t(X) = \langle X, L(t)\rangle$, and the continuity of the inner product.

(ii) Since $W^T(X) = 0$ for some $T > 0$, from Corollary 9 (ii) and (iii), we conclude that $W^{s+t}(X) = 0$. Now let us define $U(0) = U = X(s)$; since $U(t)$ is defined by $\dot{U}_i(t) = \mathcal{T}_i(U(t))$ and, from Proposition 5 (i), it holds that $\dot{X}_i(t) = \mathcal{T}_i(X(t))$, we have that $U(\tau) = X(\tau + s)$ for $0 \leq \tau \leq t$. Then, from the definition of W in (8), we can write that

$$\begin{aligned} W^t(X(s)) &= \int_0^t \langle U(\tau), C'C\rangle d\tau \\ &= \int_s^{s+t} \langle X(\tau), C'C\rangle d\tau \\ &\leq \int_0^{s+t} \langle X(\tau), C'C\rangle d\tau = W^{s+t}(X) = 0. \end{aligned}$$ □

The result in the next lemma parallels the known result in deterministic linear systems theory that every nonobserved trajectory corresponds to stable modes of the system. The proof is a counterpart of the discrete-time case presented in [4, Lemma 8].

LEMMA 19. *Consider system Φ, and let $T > 0$. (A, C, Λ) is W-detectable if and only if $\|X(t)\| \to 0$ as $t \to \infty$ whenever $W^T(X) = 0$.*

Proof. Sufficiency. Let us consider the set

$$\mathbb{Z} = \{Z : \|Z\| = 1, W^T(Z) = 0\}, \tag{19}$$

and let us denote as $Z(t)$ the trajectory corresponding to an initial condition $Z \in \mathbb{Z}$. By hypothesis, $\|Z(t)\| \to 0$ as $t \to \infty$, and we can write as in Remark 1 that there exist $0 < \xi < 1$ and $\alpha \geq 1$ such that $\|Z(t)\| \leq \alpha\xi^t$. Consequently, there exist $s_d \geq 0$ and $0 \leq \delta < 1$ such that $\|Z(s_d)\| < \delta$ for all $Z \in \mathbb{Z}$, and we can write

$$\mathbb{Z} \subset \bar{\mathbb{C}} = \{Z : \|Z\| = 1, \|Z(s_d)\| < \delta\}.$$

In this proof, we shall demonstrate that there exists $\gamma > 0$ such that, whenever $\|X(s_d)\| \geq \delta$, then $W^T(X) \geq \gamma \|X\|$, and consequently (A, C, Λ) is W-detectable. Let us deny the assertion and suppose that, for each $\gamma > 0$, there exists X, $\|X\| = 1$ such that $W^T(X) < \gamma$ and $\|X(s_d)\| \geq \delta$, i.e., $X \in \mathbb{C}$, where

$$\mathbb{C} = \{X : \|X\| = 1, \|X(s_d)\| \geq \delta\}.$$

Notice that, since $X(s_d)$ is solution of the differential equation (6), $X(s_d)$ is continuous on the initial condition X, and hence the set $\mathbb{C}$ is a compact set. Then we can take a

sequence $X_n \in \mathbb{C}$ with $\gamma_n \to 0$ as $n \to \infty$ in such a manner that, from the compactness of $\mathbb{C}$, there exists a subsequence X_m, which converges to some $\widehat{X} \in \mathbb{C}$, and, from the continuity of W^T (see Lemma 18),

$$\lim_{m\to\infty} W^T(X_m) = W^T(\widehat{X}) = 0.$$

In view of (19), $\widehat{X} \in \mathbb{Z} \subset \bar{\mathbb{C}}$, which completes the proof by contradiction.

Necessity. We shall show that, under W-detectability of (A, C, Λ), $\|X(t)\| \to 0$ as $t \to \infty$ when $W^T(X) = 0$. Since $W^T(X) = 0$, from Lemma 18, we have that $W^T(X(t)) = 0$ for all $t \geq 0$. Then, in view of the W-detectability of (A, C, Λ), we have that $\|X(t + s_d)\| < \delta \|X(t)\|$ for all $t \geq 0$ and some $s_d \geq 0$ and $0 \leq \delta < 1$; consequently, $\|X(t + ns_d)\| < \delta^n \|X(t)\|$, and hence

$$\lim_{n\to\infty} \sup_{0\leq t\leq s_d-1} \|X(t + ns_d)\| \leq \lim_{n\to\infty} \delta^n \sup_{0\leq t\leq s_d-1} \|X(t)\| = 0,$$

and the result follows in a straightforward manner. □

The next lemma presents a second version of the previous result, coined here in terms of the set of observability matrices $\mathcal{O}$.

LEMMA 20. *(A, C, Λ) is W-detectable if and only if $\lim_{t\to\infty} E\{|x(t)|^2\} = 0$ whenever $x_0 \in \mathcal{N}(\mathcal{O}_{\theta_0})$.*

Proof. Necessity. We show that $\lim_{t\to\infty} \|X(t)\| = 0$ whenever $x_0 \in \mathcal{N}(\mathcal{O}_{\theta_0})$, provided (A, C, Λ) is W-detectable. For the initial condition x_0, θ_0, we have that $X_j(0) = 0$, $j \neq \theta_0$, and $X_{\theta_0}(0) = x_0 x_0'$, and since $x_0 \in \mathcal{N}(\mathcal{O}_{\theta_0})$, Corollary 9 yields $W^t(X(0)) = 0$; Lemma 19 completes the proof.

Sufficiency. Let us assume that $W^T(X) = 0$. Any such $X \in \mathcal{M}^{n0}$ can be written in the following form (see Theorem 7.5.2 of [12]):

$$X_i = x_i^1 x_i^{1\prime} + \cdots + x_i^{r_i} x_i^{r_i\prime}, \tag{20}$$

where $x_i^k \in \mathbb{R}^n, k = 1, \ldots, r$, and $r_i = \operatorname{rank}(X_i) \leq n$. From (8), we have that $\langle X, L(T)\rangle = W^T(X) = 0$, and we can write that $W^T(x_i^k, i) \leq \langle X, L(T)\rangle = 0$ for any i and k. Thus (i) and (ii) of Corollary 9 provide

$$x_i^k \in \mathcal{N}(\mathcal{O}_i), \quad i = 1, \ldots, N, \;\; k = 1, \ldots, r_i.$$

Now let $v^{i,k}(0) = x_i^k \in \mathcal{N}(\mathcal{O}_i)$. Let $v^{i,k}(t) \in \mathbb{R}^n$, $k = 1, \ldots, r_i$, be given by the differential equation $\dot{v}^{i,k}(t) = A_{\theta(t)} v^{i,k}(t)$, $\theta(0) = i$. Since $x^{i,k}(0) \in \mathcal{N}(\mathcal{O}_{\theta(0)})$, from the assumption of the lemma, we have that

$$\lim_{t\to\infty} E\{|v^{i,k}(t)|^2\} = 0, \quad i = 1, \ldots, N, \;\; k = 1, \ldots, r_i. \tag{21}$$

Let $X^{i,k}(t) \in \mathcal{M}^{n0}$ be the second moment matrix $X_j^{i,k}(t) = E\{v^{i,k}(t) v^{i,k}(t)' 1_{\{\theta(t)=j\}}\}$, $j = 1, \ldots, N$. Notice that $X_i^{i,k}(0) = x_i^k x_i^{k\prime}$ and $X_j^{i,k}(0) = 0$ for $j \neq i$; in view of (20), we can write $X_i = \sum_{j=1}^N \sum_{k=1}^{r_j} X_i^{j,k}$. Then, from (6) and the linearity of the operator $\mathcal{T}$, we have that

$$X_i(t) = \sum_{j=1}^{N} \sum_{k=1}^{r_j} X_i^{j,k}(t),$$

and from (21) we evaluate

$$\lim_{t\to\infty}\|X(t)\| \le \sum_{j=1}^{N}\sum_{k=1}^{r_j}\lim_{t\to\infty}\|X^{j,k}(t)\| = \sum_{j=1}^{N}\sum_{k=1}^{r_j}\lim_{t\to\infty}E\{|v^{j,k}(t)|^2\} = 0.$$

We have shown, under the assumption of the lemma, that $\|X(t)\| \to 0$ as $t \to \infty$ for each $X \in \mathcal{M}^{n0}$ such that $W^T(X) = 0$; Lemma 19 provides that (A, C, Λ) is W-detectable. □

COROLLARY 21. *If the triplet (A, C, Λ) is not W-detectable, then there exist $i \in \mathcal{S}$ and $x_0 \in \mathcal{N}(\mathcal{O}_i)$ such that $\lim_{t\to\infty} E\{|x(t)|^2\} \neq 0$, for the initial condition $x(0) = x_0$ and $\theta(0) = i$.*

4.1. W-detectability and MS-detectability. This section deals with the relation between the concepts of MS-detectability and W-detectability. In the main result of this section, we show that W-detectability of (A, C, Λ) is equivalent to MS-detectability of $(A, \mathcal{O}, \Lambda)$. From the main result, we also derive a computational test for W-observability.

We start by dealing with the following closed-loop version of the MJLS:

$$\Phi_o : \dot{x}(t) = (A_{\theta(t)} + G_{\theta(t)}\mathcal{O}_{\theta(t)})x(t), \quad x(0) = x_0,\ \theta(0) = \theta_0. \tag{22}$$

For each $i = 1, \ldots, N$, we set

$$G_i = (-A_i - I)\mathcal{O}_i^+, \tag{23}$$

where $\mathcal{O}_i^+$ denotes the pseudoinverse of $\mathcal{O}_i$.

Let us present some properties of system Φ_o with G given in (23). First, one has that $\mathcal{O}_i^+\mathcal{O}_i x$ is the orthogonal projection of x onto $\mathcal{R}\{\mathcal{O}_i'\}$ and $I - \mathcal{O}_i^+\mathcal{O}_i$ is the projection onto $\mathcal{N}\{\mathcal{O}_i\}$. Notice that we can write $x(t) = \hat{x}(t) + \tilde{x}(t)$, where $\hat{x}(t) = \mathcal{O}_{\theta(t)}^+\mathcal{O}_{\theta(t)}x(t)$ and $\tilde{x}(t) = (I - \mathcal{O}_{\theta(t)}^+\mathcal{O}_{\theta(t)})x(t)$, and one can easily check that

$$\hat{x}(t) \perp \tilde{x}(t). \tag{24}$$

In what follows, we study each component $\hat{x}$ and $\tilde{x}$ separately. For ease of notation, we denote $\mathcal{O}_{\theta(t^-)} = \lim_{s\uparrow t}\mathcal{O}_{\theta(s)}$ and similarly for $\hat{x}(\cdot)$ and $\tilde{x}(\cdot)$. Let us define the sequence of jump times $t_1, t_2, \ldots$, as

$$\begin{cases} t_0 = 0, \\ t_{m+1} = \inf\{t > t_m : \mathcal{N}\{\mathcal{O}_{\theta(t^-)}\} \neq \mathcal{N}\{\mathcal{O}_{\theta(t)}\}\}, \quad m \ge 0. \end{cases} \tag{25}$$

LEMMA 22. *Consider system Φ_o with G given in (23). Then $|\hat{x}(t)| \le e^{-t}|\hat{x}(0)|$ a.s.*

Proof. From (23) and (22), it is a simple matter to check that, for $t_{m-1} \le t < t_m$, $\dot{\hat{x}}(t) = -\hat{x}(t)$ with a given condition $\hat{x}(t_{m-1})$ due to the strong Markov property of MJLS [8] and the linearity of Φ_o; this means that

$$\hat{x}(t) = e^{-(t-t_{m-1})}\hat{x}(t_{m-1}), \quad t_{m-1} \le t < t_m \text{ a.s.} \tag{26}$$

Regarding the sequence of jump times, from (25) and Lemma 14, we have that $\mathcal{N}\{\theta(t_m^-)\} \subset \mathcal{N}\{\theta(t_m)\}$ holds strictly, which yields that the projection of $\tilde{x}(t_m^-)$ onto $\mathcal{R}\{\mathcal{O}_{\theta(t_m)}'\}$ is null and $\hat{x}(t_m)$ is simply the result of the orthogonal projection of $\hat{x}(t_m^-)$

onto $\mathcal{R}\{\mathcal{O}'_{\theta(t_m)}\}$. Then the value of the Euclidean norm of $\hat{x}(\cdot)$ decreases at the sequence of jump times,

$$|\hat{x}(t_m)| \leq |\hat{x}(t_m^-)| \text{ a.s.} \tag{27}$$

The result follows from (26) and (27). ∎

LEMMA 23. *Consider system Φ_o with G given in* (23), *and assume that (A, C, Λ) is W-detectable. Then $E\{|\tilde{x}(t)|^2\} \to 0$ as $t \to \infty$.*

Proof. In this proof, $P_{t_m} = (I - \mathcal{O}^+_{\theta(t_m)}\mathcal{O}_{\theta(t_m)})$ stands for the orthogonal projection onto $\mathcal{N}\{\mathcal{O}_{\theta(t_m)}\}$. We start showing inductively that $E\{|\tilde{x}(t_m)|^2\} < \infty$. For $m = 0$, the result is immediate since $E\{|\tilde{x}(0)|^2\} \leq |x_0|^2$. Now we assume that $E\{|\tilde{x}(t_{m-1})|^2\} < \infty$. At time instant t_m, the orthogonal projection of $\hat{x}(t_m^-)$ onto $\mathcal{N}\{\mathcal{O}_{\theta(t_m)}\}$ is added to $\tilde{x}$; that is,

$$\tilde{x}(t_m) = \tilde{x}(t_m^-) + P_{t_m}\hat{x}(t_m^-). \tag{28}$$

Notice that $P_{t_{m-1}}P_{t_m} = P_{t_{m-1}}$ since $\mathcal{N}\{\mathcal{O}_{\theta(t_{m-1})}\} \subset \mathcal{N}\{\mathcal{O}_{\theta(t_m)}\}$, and we write $P_{t_{m-1}}P_{t_m}\hat{x}(t_m^-) = P_{t_{m-1}}\hat{x}(t_m^-) = 0$; notice that $P_{t_m}\hat{x}(t_m^-) \in \mathcal{R}\{\mathcal{O}'_{\theta(t_{m-1})}\}$ or, equivalently,

$$\tilde{x}(t_m^-) \perp P_{t_m}\hat{x}(t_m^-). \tag{29}$$

On the other hand, from (22) and (23), it is easy to check that, for $t_{m-1} \leq t < t_m$, $\dot{\tilde{x}}(t) = A_{\theta(t)}\tilde{x}(t)$ with given condition $\tilde{x}(t_{m-1})$ due to the strong Markov property of MJLS [8] and the linearity of Φ_o. Recalling that $\tilde{x}(t) \in \mathcal{N}\{\mathcal{O}_{\theta(t)}\}$ for $t_m \leq t < t_{m+1}$, Lemma 20, Remark 1, and the strong Markov property yield

$$E\{|\tilde{x}(t)|^2 1_{\{t_m \leq t < t_{m+1}\}}\} \leq \alpha E\{\xi^{t-t_m}|\tilde{x}(t_m)|^2 1_{\{t_m \leq t < t_{m+1}\}}\}, \tag{30}$$

where $0 < \xi < 1$ and $\alpha \geq 1$. Then, from (28), by employing (29) and (30) and from Lemma 22, we evaluate

$$\begin{aligned} E\{|\tilde{x}(t_m)|^2\} &= E\{|\tilde{x}(t_m^-) + P_{t_m}\hat{x}(t_m^-)|^2\} = E\{|\tilde{x}(t_m^-)|^2\} + E\{|P_{t_m}\hat{x}(t_m^-)|^2\} \\ &\leq \alpha E\{\xi^{t-t_{m-1}}|\tilde{x}(t_{m-1})|^2\} + E\{|\hat{x}(t_m^-)|^2\} \\ &< \alpha E\{|\tilde{x}(t_{m-1})|^2\} + E\{|\hat{x}(t_m^-)|^2\} < \infty, \end{aligned} \tag{31}$$

and the induction is complete. From (30) and (31), we can find $o(t) > 0$ for which

$$E\{|\tilde{x}(t)|^2 1_{\{t_m \leq t < t_{m+1}\}}\} \leq \alpha E\{\xi^{t-t_m}|\tilde{x}(t_m)|^2 1_{\{t_m \leq t < t_{m+1}\}}\} \leq o(t)$$

holds for each interval $t_{m-1} < t < t_m$, where $o(t) \to 0$ as $t \to \infty$. Then we can write

$$\begin{aligned} E\{|\tilde{x}(t)|^2\} &= E\{|\tilde{x}(t)|^2 1_{\{t_0 \leq t < t_1\}}\} + E\{|\tilde{x}(t)|^2 1_{\{t_1 \leq t < t_2\}}\} + \cdots \\ &\leq \alpha E\{\xi^{t-t_0}|\tilde{x}(t_0)|^2 1_{\{t_0 \leq t < t_1\}}\} + \alpha E\{\xi^{t-t_1}|\tilde{x}(t_1)|^2 1_{\{t_1 \leq t < t_2\}}\} + \cdots \\ &\leq o(t) + o(t) + \cdots . \end{aligned} \tag{32}$$

Finally, it can be checked that (32) has at most n elements. Indeed, from (25) and Lemma 14, it is simple to check that $\mathcal{N}\{\theta(t_0)\} \subset \mathcal{N}\{\theta(t_1)\} \subset \cdots \subset \mathcal{N}\{\theta(t_m)\}$ strictly, which yields $m \leq \dim\mathcal{N}\{\mathcal{O}_{\theta(t_m)}\} \leq n$, where the limit comes from the fact that $\mathcal{O}_i \in \mathcal{R}^{n(n^2N),n}$ for all i. Hence

$$\lim_{t\to\infty} E\{|\tilde{x}(t)|^2\} \leq n \lim_{t\to\infty} o(t) = 0. \qquad \blacksquare$$

Now we are ready to present the main result of the section.

THEOREM 24. *The triplet* (A, C, Λ) *is W-detectable if and only if the triplet* $(A, \mathcal{O}, \Lambda)$ *is MS-detectable.*

Proof. Necessity. Consider system Φ_o, let G be defined as in (23), and assume that (A, C, Λ) is W-detectable. From (24) and Lemmas 22 and 23, we evaluate

$$\lim_{t\to\infty} E\{|x(t)|^2\} = \lim_{t\to\infty} E\{|\hat{x}(t)|^2\} + \lim_{t\to\infty} E\{|\tilde{x}(t)|^2\}$$
$$\leq \lim_{t\to\infty} e^{-2t} E\{|x(0)|^2\} + \lim_{t\to\infty} E\{|\tilde{x}(t)|^2\} = 0.$$

Thus $(A + G\mathcal{O}, \Lambda)$ is MS-stable, which implies that $(A, \mathcal{O}, \Lambda)$ is MS-detectable.

Sufficiency. We show that $(A, \mathcal{O}, \Lambda)$ is not MS-detectable provided the triplet (A, C, Λ) is not W-detectable. Consider $i \in \mathcal{S}$ and $x_0 \in \mathcal{N}(\mathcal{O}_i)$ as in Corollary 21. For the initial condition $x(0) = x_0$ and $\theta(0) = i$, we have from Corollary 15 that $\mathcal{O}_{\theta(t)} x(t) = 0$, $t \geq 0$. Then the term $G_{\theta(t)}\mathcal{O}_{\theta(t)}x(t)$ vanishes in (22), and $x(t)$ evolves according to $\dot{x}(t) = A_{\theta(t)}x(t)$ for any $G \in \mathcal{M}^n$ in such a manner that the system Φ_o behaves as its open-loop version Φ no matter how G is chosen. Finally, from Corollary 21, we have that $\lim_{t\to\infty} E\{|x(t)|^2\} \neq 0$, and we conclude that there is no $G \in \mathcal{M}^n$ such that $A + G\mathcal{O}$ is MS-stable; hence $(A, \mathcal{O}, \Lambda)$ is not MS-detectable. □

The relationship between MS-detectability and W-detectability is established as follows.

THEOREM 25. *If* (A, C, Λ) *is MS-detectable, then* (A, C, Λ) *is W-detectable.*

Proof. It follows from the definition of $\mathcal{O}$ that $\mathcal{N}\{\mathcal{O}_i\} \subset \mathcal{N}\{C_i\}$. In view of this fact, it is simple to check that, given $K \in \mathcal{M}^{n,q}$, there always exists $G \in \mathcal{M}^{n,n^3N}$ such that $G_i\mathcal{O}_i = K_iC_i$, $i = 1, \ldots, N$; hence we have that MS-detectability of (A, C, Λ) implies MS-detectability of $(A, \mathcal{O}, \Lambda)$. Theorem 24 completes the proof. □

Notice that the converse of Theorem 25 does not hold in general since it is a simple matter to find situations for which $\mathcal{N}\{C_i\} \subset \mathcal{N}\{\mathcal{O}_i\}$ strictly.

Remark 4. Theorem 24 allows one to test the W-detectability of the triplet (A, C, Λ) by checking the MS-detectability of the triplet $(A, \mathcal{O}, \Lambda)$. For a downsizing in the dimensionality, one can check alternatively if the triplet $(A, \mathcal{O}'\mathcal{O}, \Lambda)$ is MS-detectable. The following computational form for the MS-detectability test, posed in terms of linear matrix inequalities, is an adaptation of the results in [17]: the MS-detectability of (A, C, Λ) is equivalent to the feasibility of the set

$$A_i'X_i + X_iA_i + C_i'L_i' + L_iC_i + \mathcal{E}_i(X) < 0, \quad i = 1, \ldots, N,$$

in the unknowns $X_i \in \mathcal{R}^{n0}$ and L_i of appropriate dimensions.

Example 2. Let $N = 2$, $n = 1$, and set

$$A_1 = -2;\ A_2 = 2;\ C_1 = 1;\ C_2 = 0;\ \Lambda = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}.$$

From (9), we evaluate $\mathcal{O}_1 = [1\ -5]'$ and $\mathcal{O}_2 = [0\ 1]'$, and Theorem 11 ensures that (A, C, Λ) is W-observable. Notice also that the condition in Lemma 20 is trivially satisfied. On the other hand, one can check by employing the result in Remark 4 that (A, C, Λ) is not MS-detectable.

Remark 5. It can be shown that matrix $\mathcal{O}_i$ is full rank if the pair (A_i, C_i) is observable. From this result and the result in Lemma 14, we conclude that a sufficient condition for W-observability of (A, C, Λ) is that the pair (A_i, C_i) is observable and $\lambda_{ji} > 0$ for all $j \neq i$. For instance, this is the scenario in Example 2.

Remark 6. For the deterministic linear system described by the pair (A_i, C_i), let N_i stand for the observability matrix $N_i = [C_i \vdots A_i C_i \vdots \cdots \vdots A_i^{n-1} C_i]$. It is widely known that the detectability of (A_i, N_i) is equivalent to the detectability of (A_i, C_i). This property is not mirrored by the MS-detectability concept since Theorem 24 states that the MS-detectability of $(A, \mathcal{O}, \Lambda)$ is equivalent to the W-detectability of (A, C, Λ), which is more general than the MS-detectability of (A, C, Λ). The scenario of MJLS with W-detectability is as follows:

$$\text{W-detectable } (A, \mathcal{O}, \Lambda) \;\Leftrightarrow\; \text{W-detectable } (A, C, \Lambda) \;\Leftrightarrow\; \text{MS-detectable } (A, \mathcal{O}, \Lambda),$$

where the first equivalence follows from Lemma 12.

Remark 7. For the degenerate case $\Lambda = 0$, we can show that W-detectability is equivalent to MS-detectability. This relation is exposed by the following equivalencies; most of them are simple to verify and are stated without further reference. We use a concise but self-evident notation; as in the remark above, N_i denotes the observability matrix of the system described by (A_i, C_i).

(i) MS-detec$(A, C, \Lambda) \;\Leftrightarrow\;$ detec(A_i, C_i) for all $i \;\Leftrightarrow\;$ detec(A_i, N_i) for all i;

(ii) $\mathcal{N}(N_i) \equiv \mathcal{N}(\mathcal{O}_i)$ for all i;

(iii) detec$(A_i, N_i) \;\Leftrightarrow$ detec$(A_i, \mathcal{O}_i)$;

(iv) detec$(A_i, \mathcal{O}_i)$ for all $i \;\Leftrightarrow$ MS-detec$(A, \mathcal{O}, \Lambda) \;\Leftrightarrow$ W-detec(A, C, Λ) (Theorem 24).

5. W-detectability and the LQ problem. In this section, we consider the linear quadratic control for system Φ. Under the W-detectability assumption, we show that the closed-loop system is MS-stable if a set of coupled algebraic matrix equations associated with the closed-loop system has a solution. In particular, we conclude that the solution to the CARE arising in the LQ problem is a unique stabilizing solution. Thus W-detectability not only generalizes MS-detectability but also plays the same role of MS-detectability in optimal LQ problems; see [7] and [9].

We start with some preliminary results for the open-loop system Φ. Consider the cost functional

$$J^T(X) = \int_0^T \langle U(\tau), S\rangle d\tau = E\left\{\int_0^T x(\tau)' S_{\theta(\tau)} x(\tau) d\tau | \mathcal{F}_0\right\} \tag{33}$$

defined whenever $U(0) = X$, where $S \in \mathcal{M}^{n0}$. Notice that the functionals J^T and W^T are closely related. In fact, it is easy to check that, when $C = S^{1/2}$, $W^T(X)$ and $J^T(X)$ coincide.

Let us consider the following coupled equation in the unknown $P \in \mathcal{M}^n$:

$$0 = \mathcal{L}_i(P) + S_i, \quad i = 1, \ldots, N, \tag{34}$$

with $S \in \mathcal{M}^{n0}$. The following results are derived from [6] and [13].

PROPOSITION 26. *Consider system* Φ *and the set of equations* (34). *Then the following assertions hold:*

(i) *If there exists* $P \in \mathcal{M}^{n0}$ *satisfying* (34), *then*

$$J^\infty(X) = \lim_{T\to\infty} J^T(X) \le \langle X, P\rangle. \tag{35}$$

(ii) *Assume that* A *is MS-stable. Then there exists a unique* P *satisfying* (34) *and* $P \in \mathcal{M}^{n0}$; *moreover,*

$$J^\infty(X) = \langle X, P\rangle.$$

LEMMA 27. *Assume that $(A, S^{1/2}, \Lambda)$ is W-detectable and that there exists $P \in \mathcal{M}^{n0}$ such that $J^\infty(X) < \langle X, P\rangle$. Then A is MS-stable.*

Proof. Let t_d, s_d, δ, and γ be as in Definition 16. Let us assume that A is not MS-stable; in this situation, there exists $X(0) \neq 0$ such that

$$\|X(t)\| \geq \beta\zeta^t \|X(0)\| \tag{36}$$

for some $0 < \beta \leq 1$ and $\zeta \geq 1$; see Remark 1. Let us define the sequence $\mathcal{N} = \{n_0, n_1, \dots\}$, where $n_0 = 0$ and each n_m, $m = 1, 2, \dots$, is the smallest integer such that $n_m > n_{m-1}$ and

$$\|X((n_m + 1)s_d)\| \geq \delta \|X(n_m s_d)\|$$

hold. It is easy to check that, if the number of elements of $\mathcal{N}$ is finite, then

$$\lim_{m\to\infty} \|X(m\, s_d)\| = 0,$$

which contradicts (36) and we conclude that $\mathcal{N}$ has infinitely many elements. Hence we can take a subsequence with infinitely many elements $\mathcal{N}' = \{n_{m_0}, n_{m_1}, \dots\}$, where $n_{m_0} = m_0 = 0$ and each m_k, $k = 1, 2, \dots$, is the smallest integer, such that $n_{m_k} \geq n_{m_{k-1}} + \max\{1, t_d/s_d\}$. In view of the W-detectability, we can evaluate

$$\begin{aligned} J^T(X) &= \int_0^T \langle X(\tau), S\rangle d\tau \geq \sum_{k=0}^{k'} \int_{n_{m_k} s_d}^{n_{m_k} s_d + t_d} \langle X(\tau), S\rangle d\tau \\ &= \sum_{k=0}^{k'} W^{t_d}(X(n_{m_k} s_d)) \geq \sum_{k=0}^{k'} \gamma \|X(n_{m_k} s_d)\| \\ &\geq \gamma \sum_{k=0}^{k'} \beta\zeta^{(n_{m_k} s_d)} \|X(0)\| \geq \gamma\, \beta\zeta^{(n_{m_0} s_d)}(k' + 1) \|X(0)\|, \end{aligned}$$

where k' is the largest integer for which $n_{m_{k'}} s_d + t_d \leq T$, in such a manner that $k' \to \infty$ as $T \to \infty$, and we conclude that $J^\infty(X) = \infty$, which, from Proposition 26 (i), contradicts the hypothesis of the lemma. □

Now we consider the system Φ in closed-loop form. Recall that we assumed in section 1 that both the state x and the jump variable θ are accessible for control. In this situation, it is well known that the optimal control is in linear state feedback form; see, e.g., [13]. Then we consider the following closed-loop version of system Φ:

$$\Phi_c \ : \ \dot{x}(t) = (A_{\theta(t)} + B_{\theta(t)} G_{\theta(t)})x(t), \tag{37}$$

where $B \in \mathcal{M}^{n,r}$ is given and $G \in \mathcal{M}^{r,n}$ can be regarded as a linear state feedback control. The associated infinite horizon cost functional is

$$J^\infty(X) = \lim_{t\to\infty} \int_0^T \langle U(\tau), Q + G'RG\rangle d\tau, \tag{38}$$

where $Q \in \mathcal{M}^{n0}$ and $R \in \mathcal{M}^{r+}$, defined whenever $U(0) = X$. The system Φ_c is said to be MS-stabilizable when there exists $G \in \mathcal{M}^{r,n}$ such that $A + BG$ is MS-stable. In what follows, $\mathcal{L}_G$ refers to the operator $\mathcal{L}$ associated to the closed-loop system with gain G; namely,

$$\mathcal{L}_{Gi}(U) = \widehat{A}_i' U_i + U_i \widehat{A}_i + \sum_{j=1}^{N} \lambda_{ij} U_j, \tag{39}$$

where $\widehat{A}_i = A_i + B_iG_i$ for each i. The same notation applies to $\mathcal{T}_G$.

A question that arises is whether a W-detectable open-loop triplet (A, C, Λ) can turn into a non-W-detectable closed-loop triplet $((A + BG), C, \Lambda)$. The next lemma gives an answer to this conjecture.

LEMMA 28. *If $(A, Q^{1/2}, \Lambda)$ is W-detectable, then $(A + BG, (Q + G'RG)^{1/2}, \Lambda)$ is W-detectable for any $G \in \mathcal{M}^{r,n}$ and $R \in \mathcal{M}^{r+}$.*

Proof. In this proof, $\mathcal{T}$ and W refer to the system Φ, and $\mathcal{T}_G$ and W_G refer to Φ_c; $X(\cdot)$ represents the trajectory of system Φ_c. We show that $\|X(t)\| \to 0$ as $t \to \infty$ whenever $W_G^T(X) = 0$. From Lemma 18, we can write for all $t \geq 0$ that

$$0 = W_G^t(X) = \int_0^t \langle X(\tau), (Q + G'RG)'^{1/2}(Q + G'RG)^{1/2}\rangle d\tau \tag{40}$$
$$= \int_0^t \langle X(\tau), Q + G'RG\rangle d\tau \geq \int_0^t \langle X(\tau), G'RG\rangle d\tau.$$

From the continuity of $X(t)$, we can evaluate

$$\langle X(t), G'RG\rangle = \langle R^{1/2}GX(t)^{1/2}, R^{1/2}GX(t)^{1/2}\rangle = 0$$

for all $t \geq 0$, and, since $R_i > 0$, we get that $G_iX(t) = 0$ for all $t \geq 0$ and i. Then we can write

$$\mathcal{T}_{Gi}(X(t)) = \widehat{A}_iX_i(t) + X_i\widehat{A}_i' + \sum_{j=1}^N \lambda_{ji}X_j$$
$$= A_iX_i(t) + X_iA_i' + \sum_{j=1}^N \lambda_{ji}X_j = \mathcal{T}_i(X(t))$$

for all $t \geq 0$, and, in view of Proposition 5 with $\widehat{A}_i = (A_i + B_iG_i)$ for $i = 1, \ldots, N$, we have that such trajectories of systems Φ_c and Φ coincide whenever the initial conditions coincide. We set $C = Q^{1/2}$ in system Φ to conclude, similarly to (40), that

$$W^T(X) = \int_0^T \langle X(\tau), C'C\rangle d\tau = \int_0^T \langle X(\tau), Q\rangle d\tau \leq W_G^T(X) = 0$$

for any $s > 0$, and the detectability of $(A, Q^{1/2}, \Lambda)$ ensures that $\|X(t)\| \to 0$ as $t \to \infty$. ∎

THEOREM 29. *Consider the closed-loop system Φ_c with a linear state feedback control $G \in \mathcal{M}^{r,n}$. Assume that $(A, Q^{1/2}, \Lambda)$ is W-detectable. If there exists a solution $P \in \mathcal{M}^{n0}$ of*

$$\mathcal{L}_{Gi}(P) + Q_i + G_i'R_iG_i = 0, \quad i = 1, \ldots, N, \tag{41}$$

then $A + BG$ is MS-stable.

Proof. Set $S = Q + G'RG$, and notice, from Lemma 28 and the assumption in the theorem, that $(\hat{A}, S^{1/2}, \Lambda)$ is W-detectable, where $\widehat{A} = (A + BG)$. Then, from Proposition 26 (i), we have that $J_G^\infty(X) \leq \langle X, P\rangle$, and Lemma 27 states that $\widehat{A}$ is MS-stable. ∎

THEOREM 30. *Consider the system*

$$\dot{x}(t) = A_{\theta(t)}x(t) + B_{\theta(t)}u(t), \tag{42}$$

the associated infinite-horizon linear quadratic cost $J^\infty(X)$, *and the CARE in the unknown* $P \in \mathcal{M}^{n0}$:

$$A_i'P_i + P_iA_i + \sum_{j=1}^{N} \lambda_{ij}P_j - P_iB_iR_i^{-1}B_i'P_i + Q_i = 0. \tag{43}$$

Assume that $(A, Q^{1/2}, \Lambda)$ *is W-detectable. Then the following assertions hold:*

(i) *There exists a solution* $P \in \mathcal{M}^{n0}$ *of* (43) *if and only if the system is MS-stabilizable;*

(ii) *If* P *is a solution of* (43), *then it is unique. The optimal state feedback control*

$$u(t) = -R_i^{-1}B_i'P_ix(t) \quad \textit{whenever } \theta(t) = i \tag{44}$$

is such that

$$\lim_{t\to\infty} E\{|x(t)|^2\} = 0.$$

Proof. The sufficiency part of assertion (i) is a well-known result; see, e.g., [17, Theorem 3.1]. Notice that, if we denote $G_i = -R_i^{-1}B_i'P_i$, we can write (43) equivalently as

$$\mathcal{L}_{Gi}(P) + G_i'R_iG_i + Q_i = 0, \quad i = 1, \dots, N, \tag{45}$$

and, from Theorem 29, we have that $A + BG$ is MS-stable. This argument completes the proof of assertion (i) and also part of the assertion (ii) regarding the MS-stability of the closed-loop system defined by (44). Let us show now that P is unique. Suppose that $\bar{P} \in \mathcal{M}^{n0}$ is a solution of (43). In a similar fashion to (45), we can write

$$\mathcal{L}_{\bar{G}i}(\bar{P}) + \bar{G}_i'R_i\bar{G}_i + Q_i = 0, \quad i = 1, \dots, N, \tag{46}$$

where $\bar{G}_i = -R_i^{-1}B_i'\bar{P}_i$; notice that, from Theorem 29, the system with gain $\bar{G}$ is also MS-stable. Subtracting (46) from (45), we get, after some manipulations, that

$$\mathcal{L}_{Gi}(P - \bar{P}) + (G_i - \bar{G}_i)'R_i(G_i - \bar{G}_i) = 0, \quad i = 1, \dots, N,$$

and we identify $S_i = (G_i - \bar{G}_i)'R_i(G_i - \bar{G}_i)$ in (34) to get, from Proposition 26 (ii), that $P - \bar{P} \in \mathcal{M}^{n0}$; that is, $P_i - \bar{P}_i \geq 0$, $i = 1, \dots, N$. Now, subtracting (45) from (46), we get similarly that $\bar{P}_i - P_i \geq 0$, $i = 1, \dots, N$, and we conclude that $\bar{P} = P$. It remains only to show that the feedback control (44) is optimal. Let us suppose that there exist $X \in \mathcal{M}^{n0}$ and $\bar{G} \in \mathcal{M}^{r,n}$ such that $J_{\bar{G}}^\infty(X) < J_G^\infty(X)$. From Proposition 26 (ii), we have that $J_{\bar{G}}^\infty(X) < J_G^\infty(X) = \langle X, P\rangle$, and, since $(A + B\bar{G}, Q + \bar{G}'R\bar{G}, \Lambda)$ is W-detectable (see Lemma 28), Lemma 27 ensures that the closed-loop system with gain $\bar{G}$ is MS-stable. Then, from Proposition 26 (ii), we have that there exists a unique solution $\bar{P}$ to (46) and $J_{\bar{G}}^\infty(X) = \langle X, \bar{P}\rangle$. Once again, subtracting (45) from (46), we obtain $\bar{P} \geq P$, and we conclude that $J_{\bar{G}}^\infty(X) = \langle X, \bar{P}\rangle \geq \langle X, P\rangle = J_G^\infty(X)$, which denies the initial hypothesis, and hence

$$J_G^\infty(X) \leq J_K^\infty(X) \quad \forall K \in \mathcal{M}^{r,n}.$$

The fact that the optimal control action is in linear state feedback form is a well-established result which comes from dynamic programming arguments and from the fact that the system is Markovian; see, for instance, [13] and [15]. ∎

6. Conclusions. This paper introduces the concept of W-detectability and the set of observability matrices $\mathcal{O}$ that is related to the concept of W-observability for continuous-time MJLS.

We show that the concepts of W-observability and W-detectability reproduce geometric and qualitative properties of the deterministic concepts within the MJLS setting. In particular, we show how the properties (I)–(IV) mentioned in section 1 extend to MJLS; respectively, we have shown that

- if $x(t) \in \mathcal{N}\{\mathcal{O}_{\theta(t)}\}$, then $x(s) \in \mathcal{N}\{\mathcal{O}_{\theta(s)}\}$ a.s. for all $s \geq t$;
- (A, C, Λ) is W-observable if and only if $\mathcal{O}_i$ has full rank for each $i = 1, \ldots, N$;
- (A, C, Λ) is W-detectable if and only if $\lim_{t\to\infty} E\{|x(t)|^2\} = 0$ whenever $x_0 \in \mathcal{N}(\mathcal{O}_{\theta_0})$; and
- (A, C, Λ) is W-detectable provided (A, C, Λ) is W-observable.

We also show that those concepts generalize the previous concepts encountered in the literature and that they play the same role in the quest for stabilizing solutions of quadratic control problems. Regarding the concept of MS-detectability, in one of the main results of this paper, we show that (A, C, Λ) is W-detectable if and only if $(A, \mathcal{O}, \Lambda)$ is MS-detectable. The result provides a testable condition for W-detectability. Moreover, the kernel of $\mathcal{O}$ is in general smaller than that of the original matrices C, henceforth making W-detectability and MS-detectability directly comparable.

Testable conditions for the concept of W-observability is also developed in terms of the set of observability matrices $\mathcal{O}$. The test of W-observability in Theorem 11 for MJLS and the observability test for N deterministic time-invariant linear systems, each with dimension n, are alike.

REFERENCES

[1] B. D. O. Anderson and J. B. Moore, *Detectability and stabilizability of time-varying discrete-time linear systems*, SIAM J. Control Optim., 19 (1981), pp. 20–32.

[2] S. Bittanti, A. J. Laub, and J. C. Willems, *The Riccati Equation*, Springer-Verlag, New York, 1991.

[3] J. W. Brewer, *Kronecker products and matrix calculus in system theory*, IEEE Trans. Circuits Systems I Fund. Theory Appl. 25 (1979), pp. 772–781.

[4] E. F. Costa and J. B. R. do Val, *On the detectability and observability of discrete-time Markov jump linear systems*, Systems Control Lett., 44 (2001), pp. 135–145.

[5] O. L. Costa and M. D. Fragoso, *Stability results for discrete-time linear systems with Markovian jumping parameters*, J. Math. Anal. Appl., 179 (1993), pp. 154–178.

[6] O. L. V. Costa, J. B. R. do Val, and J. C. Geromel, *Continuous-time state-feedback H_2-control of Markovian jump linear systems via convex analysis*, Automatica J. IFAC, 35 (1999), pp. 259–268.

[7] O. L. V. Costa and M. Fragoso, *Discrete-time LQ-optimal control problems for infinite Markov jump parameter systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 2076–2088.

[8] M. H. A. Davis, *Markov Models and Optimization*, Chapman and Hall, London, 1993.

[9] J. B. R. do Val, J. C. Geromel, and O. L. V. Costa, *Solutions for the linear quadratic control problem of Markov jump linear systems*, J. Optim. Theory Appl., 103 (1999), pp. 283–311.

[10] X. Feng, K. A. Loparo, Y. Ji, and H. J. Chizeck, *Stochastic stability properties of jump linear systems*, IEEE Trans. Automat. Control, 37 (1992), pp. 38–53.

[11] W. W. Hager and L. L. Horowitz, *Convergence and stability properties of the discrete Riccati operator equation and the associated optimal control and filtering problems*, SIAM J. Control Optim., 14 (1976), pp. 295–312.

[12] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1990.

[13] Y. Ji and H. J. Chizeck, *Controllability, stabilizability and continuous time Markovian jump linear quadratic control*, IEEE Trans. Automat. Control, 35 (1990), pp. 777–788.
[14] T. Kailath, *Linear Systems*, Prentice–Hall, Englewood Cliffs, NJ, 1980.
[15] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*, Prentice–Hall, Englewood Cliffs, NJ, 1986.
[16] T. Morozan, *Stability and control for linear systems with jump Markov perturbations*, Stochastic Anal. Appl., 13 (1995), pp. 91–110.
[17] M. A. Rami and L. E. Ghaoui, *LMI optimization for nonstandard Riccati equations arising in stochastic control*, IEEE Trans. Automat. Control, 41 (1996), pp. 1666–1671.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 4, pp. 1315–1319

ERRATUM: SENSITIVITY ANALYSIS OF THE VALUE FUNCTION FOR OPTIMIZATION PROBLEMS WITH VARIATIONAL INEQUALITY CONSTRAINTS*

YVES LUCET† AND JANE J. YE‡

Abstract. In our paper [*SIAM J. Control Optim.*, 40 (2001), pp. 699–723], due to an error in the proof, an additional assumption is needed for the conclusion of Theorem 3.6 to hold. In this erratum, we restate and prove Theorem 3.6 and correct other related mistakes accordingly.

PII. S036301290139926X

In our paper [1], due to an error in the proof, an additional assumption is needed for the conclusion of Theorem 3.6 to hold. As a consequence, Theorem 4.2 does not hold, each of Theorems 4.4, 4.8, 4.11, and 4.13 requires an additional assumption, and the last two lines on page 701 and the first two lines on page 702 should be changed to

$$M^1 = M^1_{CD}(\Sigma), M^1_C(\Sigma), M^1_S(\Sigma),$$
$$M^0 = M^0_{CD}(\Sigma), M^0_C(\Sigma), M^0_S(\Sigma).$$

We first correct Theorem 3.6 by adding the additional assumption (0.1) as follows.

THEOREM 3.6. *In addition to the basic assumption* (BH), *assume that there exists* $\delta > 0$ *such that the set*

$$\{(x,y) \in C \ : \ \Psi(x,y,\bar\alpha) \le p, H(x,y,\bar\alpha) = q, r \in F(x,y,\bar\alpha) + N_\Omega(y), f(x,y,\bar\alpha) \le M, \\ (p,q,r) \in B(0;\delta)\}$$

is bounded for each M *and the following assumption holds:*

$$(\gamma,\beta,\eta,0) \in M^0(\bar x,\bar y,\bar\alpha) \text{ implies } \gamma = 0,\ \beta = 0,\ \eta = 0. \tag{0.1}$$

Then the value function $V(\alpha)$ *is lower semicontinuous near* $\bar\alpha$, *and*

$$\partial V(\bar\alpha) \subseteq \bigcup_{(\bar x,\bar y)\in\Sigma(\bar\alpha)} \{-\zeta : (\gamma,\beta,\eta,\zeta) \in M^1(\bar x,\bar y,\bar\alpha)\},$$
$$\partial^\infty V(\bar\alpha) \subseteq \bigcup_{(\bar x,\bar y)\in\Sigma(\bar\alpha)} \{-\zeta : (\gamma,\beta,\eta,\zeta) \in M^0(\bar x,\bar y,\bar\alpha)\},$$

where $M^\lambda(\bar x,\bar y,\bar\alpha)$ *is the set of index* λ *multipliers for problem* GP(p,q,r,α) *at* $(0,0,0,\bar\alpha)$, *i.e., vectors* $(\gamma,\beta,\eta,\zeta)$ *in* $R^d \times R^l \times R^m \times R$ *satisfying*

$$\begin{cases} 0 \in \lambda\partial f(\bar x,\bar y,\bar\alpha) + \partial\langle\Psi,\gamma\rangle(\bar x,\bar y,\bar\alpha) + \partial\langle H,\beta\rangle(\bar x,\bar y,\bar\alpha) + \partial\langle F,\eta\rangle(\bar x,\bar y,\bar\alpha) \\ +\{0\}\times D^*N_\Omega(\bar y,-F(\bar x,\bar y,\bar\alpha))(\eta)\times\{0\} + \{(0,0,\zeta)\} + N_C(\bar x,\bar y)\times\{0\}, \\ \gamma \ge 0 \text{ and } \langle\Psi(\bar x,\bar y,\bar\alpha),\gamma\rangle = 0, \end{cases}$$

*Received by the editors December 6, 2001; accepted for publication (in revised form) May 20, 2002; published electronically December 3, 2002. This work was partly supported by an NSERC research grant.

http://www.siam.org/journals/sicon/41-4/39926.html

†Center for Experimental and Constructive Mathematics, Simon Fraser University, 8888 University Dr., Burnaby, BC, Canada, V5A 1S6 (lucet@cecm.sfu.ca).

‡Department of Mathematics and Statistics, University of Victoria, Victoria, BC, Canada V8W 3P4 (janeye@Math.UVic.CA).

and $\Sigma(\bar{\alpha})$ *is the set of solutions of problem* $GP(\bar{\alpha})$.

We now make the correct statements for Theorems 4.4, 4.8, 4.11, and 4.13 by translating assumption (0.1) to the case of CD, C, P, and S multipliers, respectively. Unless otherwise indicated, we denote by $\nabla f(x,y,\alpha)$ the gradient of function f with respect to (x,y,α) and not the gradient of f with respect to (x,y) as in section 4 of [1].

THEOREM 4.4. *Assume that there exists* $\delta > 0$ *such that the set*

$$\begin{aligned}\{(x,y) \in C :{} & (p,q,r) \in B(0;\delta), \Psi(x,y,\bar{\alpha}) \le p, H(x,y,\bar{\alpha}) = q,\\ & y \ge 0,\ F(x,y,\bar{\alpha}) \ge r, \langle y, F(x,y,\bar{\alpha}) - r\rangle = 0, f(x,y,\bar{\alpha}) \le M\}\end{aligned}$$

is bounded for each M. *Assume also that*

$$\begin{aligned}&0 \in \nabla\Psi(\bar{x},\bar{y},\bar{\alpha})^\top\gamma + \nabla H(\bar{x},\bar{y},\bar{\alpha})^\top\beta + \nabla F(\bar{x},\bar{y},\bar{\alpha})^\top\eta + (0,\xi,0) + N_C(\bar{x},\bar{y}) \times \{0\},\\ &\gamma \ge 0 \text{ and } \langle\Psi(\bar{x},\bar{y},\bar{\alpha}),\gamma\rangle = 0,\\ &\xi_i = 0 \qquad \text{if } \bar{y}_i > 0 \text{ and } F_i(\bar{x},\bar{y},\bar{\alpha}) = 0,\\ &\eta_i = 0 \qquad \text{if } \bar{y}_i = 0 \text{ and } F_i(\bar{x},\bar{y},\bar{\alpha}) > 0,\\ &\text{either } \xi_i < 0, \eta_i < 0 \text{ or } \xi_i\eta_i = 0 \qquad \text{if } \bar{y}_i = 0 \text{ and } F_i(\bar{x},\bar{y}) = 0\end{aligned}$$

implies that $\gamma = 0$, $\beta = 0$, $\eta = 0$. *Then the value function* V *is lower semicontinuous near* $\bar{\alpha}$, *and*

$$\begin{aligned}\partial V(\bar{\alpha}) \subseteq \bigcup_{(\bar{x},\bar{y})\in\Sigma(\bar{\alpha})} &\{\nabla_\alpha f(\bar{x},\bar{y},\bar{\alpha}) + \nabla_\alpha\Psi(\bar{x},\bar{y},\bar{\alpha})^\top\gamma + \nabla_\alpha H(\bar{x},\bar{y},\bar{\alpha})^\top\beta\\ &+\nabla_\alpha F(\bar{x},\bar{y},\bar{\alpha})^\top\eta : (\gamma,\beta,\eta) \in M^1_{CD}(\bar{x},\bar{y})\},\end{aligned} \tag{0.2}$$

$$\begin{aligned}\partial^\infty V(\bar{\alpha}) \subseteq \bigcup_{(\bar{x},\bar{y})\in\Sigma(\bar{\alpha})} &\{\nabla_\alpha\Psi(\bar{x},\bar{y},\bar{\alpha})^\top\gamma + \nabla_\alpha H(\bar{x},\bar{y},\bar{\alpha})^\top\beta\\ &+\nabla_\alpha F(\bar{x},\bar{y},\bar{\alpha})^\top\eta : (\gamma,\beta,\eta) \in M^0_{CD}(\bar{x},\bar{y})\}.\end{aligned} \tag{0.3}$$

If the set in the right-hand side of inclusion (0.3) *contains only the zero vector, then the value function* V *is Lipschitz near* $\bar{\alpha}$. *If the set in the right-hand side of inclusion* (0.3) *contains only the zero vector and the set in the right-hand side of inclusion* (0.2) *is a singleton, then the value function is strictly differentiable at* $\bar{\alpha}$.

THEOREM 4.8. *Assume that there exists* $\delta > 0$ *such that the set*

$$\begin{aligned}\{(x,y) \in C :{} & (p,q,q^m) \in B(0;\delta), \Psi(x,y,\bar{\alpha}) \le p, H(x,y,\bar{\alpha}) = q,\\ & \min\{y_i, F_i(x,y,\bar{\alpha})\} = q^m_i, i = 1,\dots,m, f(x,y,\bar{\alpha}) \le M\}\end{aligned}$$

is bounded for each M. *Assume also that*

$$\begin{aligned}&0 \in \nabla\Psi(\bar{x},\bar{y},\bar{\alpha})^\top\gamma + \nabla H(\bar{x},\bar{y},\bar{\alpha})^\top\beta + \nabla F(\bar{x},\bar{y},\bar{\alpha})^\top\eta + (0,\xi,0) + N_C(\bar{x},\bar{y}) \times \{0\},\\ &\gamma \ge 0, \langle\Psi,\gamma\rangle(\bar{x},\bar{y},\bar{\alpha}) = 0,\end{aligned}$$

where

$$\begin{aligned}&\eta_i = 0 \quad \forall i \in I_+,\\ &\xi_i = 0 \quad \forall i \in L,\\ &\eta_i = r_i(1-\bar{t}_i), \xi_i = r_i\bar{t}_i \text{ for some } \bar{t}_i \in [0,1], \quad \forall i \in I_0\end{aligned}$$

implies that $\gamma = 0$, $\beta = 0$, $\eta = 0$, $r_i = 0$, $i = 1, \ldots, m$. *Then the value function* V *is lower semicontinuous near* $\bar{\alpha}$, *and*

$$\partial V(\bar{\alpha}) \subseteq \bigcup_{(\bar{x},\bar{y})\in\Sigma(\bar{\alpha})} \{\nabla_\alpha f(\bar{x},\bar{y},\bar{\alpha}) + \nabla_\alpha \Psi(\bar{x},\bar{y},\bar{\alpha})^\top \gamma + \nabla_\alpha H(\bar{x},\bar{y},\bar{\alpha})^\top \beta$$
$$+\nabla_\alpha F(\bar{x},\bar{y},\bar{\alpha})^\top \eta : (\gamma,\beta,\eta) \in M_C^1(\bar{x},\bar{y})\}, \tag{0.4}$$
$$\partial^\infty V(\bar{\alpha}) \subseteq \bigcup_{(\bar{x},\bar{y})\in\Sigma(\bar{\alpha})} \{\nabla_\alpha \Psi(\bar{x},\bar{y},\bar{\alpha})^\top \gamma + \nabla_\alpha H(\bar{x},\bar{y},\bar{\alpha})^\top \beta$$
$$+\nabla_\alpha F(\bar{x},\bar{y},\bar{\alpha})^\top \eta : (\gamma,\beta,\eta) \in M_C^0(\bar{x},\bar{y})\}. \tag{0.5}$$

If the set in the right-hand side of inclusion (0.5) *contains only the zero vector, then the value function* V *is Lipschitz near* $\bar{\alpha}$. *If the set in the right-hand side of inclusion* (0.5) *contains only the zero vector and the set in the right-hand side of inclusion* (0.4) *is a singleton, then the value function is strictly differentiable at* $\bar{\alpha}$.

THEOREM 4.11. *Assume that there exists* $\delta > 0$ *such that, for* $(\bar{x},\bar{y}) \in \Sigma(\bar{\alpha})$ *and each index set* $\sigma \subseteq I_0(\bar{x},\bar{y})$, *the set in Proposition* 4.10 *is bounded for each* M *and*

$$\begin{cases} 0 = \nabla\Psi(\bar{x},\bar{y},\bar{\alpha})^\top \gamma + \nabla H(\bar{x},\bar{y},\bar{\alpha})^\top \beta + \nabla F(\bar{x},\bar{y},\bar{\alpha})^\top \eta + (0,\xi,0) + N_C(\bar{x},\bar{y}) \times \{0\}, \\ \gamma_{J(\Psi)} = 0, \eta_{I_+} = 0, \xi_L = 0, \xi_\sigma \le 0, \eta_{I_0\setminus\sigma} \le 0, \end{cases}$$

implies that $\gamma = 0$, $\beta = 0$, $\eta = 0$. *Then the value function* V *is lower semicontinuous near* $\bar{\alpha}$, *and*

$$\partial V(\bar{\alpha}) \subseteq \bigcup_{(\bar{x},\bar{y})\in\Sigma(\bar{\alpha})} \{\nabla_\alpha f(\bar{x},\bar{y},\bar{\alpha}) + \nabla_\alpha \Psi(\bar{x},\bar{y},\bar{\alpha})^\top \gamma + \nabla_\alpha H(\bar{x},\bar{y},\bar{\alpha})^\top \beta$$
$$+\nabla_\alpha F(\bar{x},\bar{y},\bar{\alpha})^\top \eta : (\gamma,\beta,\eta) \in \cup_{\sigma\subseteq I_0} M_\sigma^1(\bar{x},\bar{y})\}, \tag{0.6}$$
$$\partial^\infty V(\bar{\alpha}) \subseteq \bigcup_{(\bar{x},\bar{y})\in\Sigma(\bar{\alpha})} \{\nabla_\alpha \Psi(\bar{x},\bar{y},\bar{\alpha})^\top \gamma + \nabla_\alpha H(\bar{x},\bar{y},\bar{\alpha})^\top \beta$$
$$+\nabla_\alpha F(\bar{x},\bar{y},\bar{\alpha})^\top \eta : (\gamma,\beta,\eta) \in \cup_{\sigma\subseteq I_0} M_\sigma^0(\bar{x},\bar{y})\}. \tag{0.7}$$

If the set in the right-hand side of inclusion (0.7) *contains only the zero vector, then the value function* V *is Lipschitz near* $\bar{\alpha}$. *If the set in the right-hand side of inclusion* (0.7) *contains only the zero vector and the set in the right-hand side of inclusion* (0.6) *is a singleton, then the value function is strictly differentiable at* $\bar{\alpha}$.

THEOREM 4.13. *In addition to the assumptions of Theorem* 4.11, *assume that* $C = R^n \times R^a \times R^b$ *and, for all* $(\bar{x},\bar{z},\bar{u}) \in \Sigma(\bar{\alpha})$, *the partial MPEC linear independence constraint qualification is satisfied; i.e.,*

$$\begin{cases} 0 = \nabla_{x,y}\Psi(\bar{x},\bar{y},\bar{\alpha})^\top \gamma + \nabla_{x,y} H(\bar{x},\bar{y},\bar{\alpha})^\top \beta + \nabla_{x,y} F(\bar{x},\bar{y},\bar{\alpha})^\top \eta + (0,\xi), \\ \gamma_{J(\Psi)} = 0, \eta_{I_+} = 0, \xi_L = 0, \end{cases}$$

implies that $\eta_{I_0} = 0, \xi_{I_0} = 0$, *where* $J(\Psi) := \{i : \Psi_i(\bar{x},\bar{y},\bar{\alpha}) < 0\}$. *Further assume that*

$$\begin{cases} 0 = \nabla\Psi(\bar{x},\bar{y},\bar{\alpha})^\top \gamma + \nabla H(\bar{x},\bar{y},\bar{\alpha})^\top \beta + \nabla F(\bar{x},\bar{y},\bar{\alpha})^\top \eta + (0,\xi,0), \\ \gamma_{J(\Psi)} = 0, \eta_{I_+} = 0, \xi_L = 0, \eta_{I_0} \le 0, \xi_{I_0} \le 0, \end{cases}$$

implies that $\gamma = 0$, $\beta = 0$, $\eta = 0$. *Then the value function* V *is lower semicontinuous near* $\bar{\alpha}$, *and*

$$\begin{aligned}\partial V(\bar{\alpha}) \subseteq \bigcup_{(\bar{x},\bar{y})\in\Sigma(\bar{\alpha})} &\{\nabla_\alpha f(\bar{x},\bar{y},\bar{\alpha}) + \nabla_\alpha \Psi(\bar{x},\bar{y},\bar{\alpha})^\top \gamma + \nabla_\alpha H(\bar{x},\bar{y},\bar{\alpha})^\top \beta \\ &+\nabla_\alpha F(\bar{x},\bar{y},\bar{\alpha})^\top \eta : (\gamma,\beta,\eta) \in M_S^1(\bar{x},\bar{y})\},\\ \partial^\infty V(\bar{\alpha}) \subseteq \bigcup_{(\bar{x},\bar{y})\in\Sigma(\bar{\alpha})} &\{\nabla_\alpha \Psi(\bar{x},\bar{y},\bar{\alpha})^\top \gamma + \nabla_\alpha H(\bar{x},\bar{y},\bar{\alpha})^\top \beta \\ &+\nabla_\alpha F(\bar{x},\bar{y},\bar{\alpha})^\top \eta : (\gamma,\beta,\eta) \in M_S^0(\bar{x},\bar{y})\}.\end{aligned}$$

Note that the additional assumption (0.1) and its corresponding assumptions in Theorems 4.4., 4.8, 4.11, and 4.13 are automatically satisfied in the case in which the perturbation is additive. In the case of nonadditive perturbations, they are needed even in the case of nonlinear programming, i.e., when $\Omega = R^m$ in Theorem 3.6.

The main error occurs in the proof of Theorem 3.6 when we applied [1, Proposition 2.6] to obtain the partial subdifferentials from the subdifferentials of the fully perturbed value function. The positions of vectors ζ and 0 were switched by mistake. Instead of proving that $(\zeta, 0) \in \partial^\infty \widetilde{V}(0,\bar{\alpha})$ implies $\zeta = 0$, we proved that $(0,\zeta) \in \partial^\infty \widetilde{V}(0,\bar{\alpha})$ implies $\zeta = 0$. Hence, on page 709 in lines 13–18, "For any $(0,0,0,\zeta) \in \partial^\infty \widetilde{V}(0,0,0,\bar{\alpha})$, we have $(0,0,0,\zeta) \in -M^0(\bar{x},\bar{y},\bar{\alpha})$ for some point $(\bar{x},\bar{y},\bar{\alpha}) \in \Sigma(0,0,0,\bar{\alpha})$. Therefore,

$$(0,0,\zeta) \in N_C(\bar{x},\bar{y}) \times \{0\},$$

which implies that $\zeta = 0$" should be changed to "For any $(-\gamma,-\beta,-\eta,0) \in \partial^\infty \widetilde{V}(0,0,0,\bar{\alpha})$, we have $(-\gamma,-\beta,-\eta,0) \in -M^0(\bar{x},\bar{y},\bar{\alpha})$ for some point $(\bar{x},\bar{y},\bar{\alpha}) \in \Sigma(0,0,0,\bar{\alpha})$. Hence $(\gamma,\beta,\eta,0) \in M^0(\bar{x},\bar{y},\bar{\alpha})$, which implies $\gamma = 0$, $\beta = 0$, $\eta = 0$ by assumption (0.1)."

Consider the nonlinear programming formulation of (OPCC) in [1, section 4.1]. Assumption (0.1) amounts to the nonexistence of a nonzero vector $(\gamma,\beta,r^F,r^y,\mu)$ such that

$$\begin{aligned}&0 \in \nabla\Psi(\bar{x},\bar{y},\bar{\alpha})^\top \gamma + \nabla H(\bar{x},\bar{y},\bar{\alpha})^\top \beta\\ &-\nabla F(\bar{x},\bar{y},\bar{\alpha})^\top r^F - \{(0,r^y,0)\} + \mu\nabla\langle y,F\rangle(\bar{x},\bar{y},\bar{\alpha}) + N_C(\bar{x},\bar{y}) \times \{(0)\},\\ &\gamma \geq 0, \langle\gamma,\Psi(\bar{x},\bar{y},\bar{\alpha})\rangle = 0,\\ &r^F \geq 0, r^y \geq 0, \langle r^F, F(\bar{x},\bar{y},\bar{\alpha})\rangle = 0, \langle r^y,\bar{y}\rangle = 0.\end{aligned}$$

However, using [1, Proposition 4.16] with x replaced by (x,α), the above assumption will never be satisfied, and hence [1, Theorem 4.2] does not hold. Consider the following example, which is the example in [1] with the extra constraint $(x,y) \in [-1,1]\times[-1,1]$:

$$\begin{aligned}\text{minimize} \quad & -y\\ \text{subject to} \quad & x - y = 0,\\ & x \geq 0, y \geq 0, xy = 0, (x,y) \in [-1,1]\times[-1,1].\end{aligned}$$

Note that the growth hypothesis holds since the set $[-1,1]\times[-1,1]$ is compact. The normal multiplier set $M_{NLP}^1(0,0) = \emptyset$. So [1, Theorem 4.2] is not true for this example.

That is, the nonlinear programming multipliers may not be useful in the sensitivity analysis.

Acknowledgments. The authors would like to thank the anonymous referees, in particular, referee 2, and the associate editor who suggested listing the statements that should be corrected in the beginning of the erratum.

REFERENCE

[1] Y. Lucet and J. J. Ye, *Sensitivity analysis of the value function for optimization problems with variational inequality constraints*, SIAM J. Control Optim., 40 (2001), pp. 699–723.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 5, pp. 1321–1349

ADAPTIVE FINITE ELEMENT APPROXIMATION FOR DISTRIBUTED ELLIPTIC OPTIMAL CONTROL PROBLEMS*

RUO LI†, WENBIN LIU‡, HEPING MA§, AND TAO TANG¶

Abstract. In this paper, sharp a posteriori error estimators are derived for a class of distributed elliptic optimal control problems. These error estimators are shown to be useful in adaptive finite element approximation for the optimal control problems and are implemented in the adaptive approach. Our numerical results indicate that the sharp error estimators work satisfactorily in guiding the mesh adjustments and can save substantial computational work.

Key words. mesh adaptivity, optimal control, a posteriori error estimate, finite element method

AMS subject classifications. 49J20, 65N30

PII. S0363012901389342

1. Introduction.

Finite element approximation of optimal control problems has long been an important topic in engineering design work and has been extensively studied in the literature. There have been extensive theoretical and numerical studies for finite element approximation of various optimal control problems; see [2, 12, 13, 15, 20, 23, 37, 44]. For instance, for the optimal control problems governed by some linear elliptic or parabolic state equations, a priori error estimates of the finite element approximation were established long ago; see, for example, [12, 13, 15, 20, 23, 37]. Furthermore, a priori error estimates were established for the finite element approximation of some important flow control problems in [17] and [11]. A priori error estimates have also been obtained for a class of state constrained control problems in [43], though the state equation is assumed to be linear. In [29], this assumption has been removed by reformulating the control problem as an abstract optimization problem in some Banach spaces and then applying nonsmooth analysis. In fact, the state equation there can be a variational inequality.

In recent years, the adaptive finite element method has been extensively investigated. Adaptive finite element approximation is among the most important means to boost the accuracy and efficiency of finite element discretizations. It ensures a higher density of nodes in a certain area of the given domain, where the solution is more difficult to approximate. At the heart of any adaptive finite element method is an a posteriori error estimator or indicator. The literature in this area is extensive. Some of the techniques directly relevant to our work can be found in [1, 5, 6, 7, 28, 32, 34, 42, 47]. It is our belief that adaptive finite element enhancement is one of the future directions to pursue in developing sophisticated numerical methods for optimal design problems.

*Received by the editors May 14, 2001; accepted for publication (in revised form) April 28, 2002; published electronically December 11, 2002. This work was supported in part by Hong Kong Baptist University, Hong Kong Research Grants Council, and the British EPSRC.

http://www.siam.org/journals/sicon/41-5/38934.html

†School of Mathematical Science, Peking University, Beijing 100871, People's Republic of China.

‡CBS & Institute of Mathematics and Statistics, The University of Kent, Canterbury, CT2 7NF England (W.B.Liu@ukc.ac.uk).

§Department of Mathematics, Shanghai University, Shanghai 200436, People's Republic of China (hpma@guomai.sh.cn).

¶Department of Mathematics, The Hong Kong Baptist University, Kowloon Tong, Hong Kong (ttang@math.hkbu.edu.hk).

Although adaptive finite element approximation is widely used in numerical simulations, it has not yet been *fully* utilized in optimal control. Initial attempts in this aspect have been reported only recently for some design problems; see, e.g., [3, 4, 38, 41]. However, a posteriori error indicators of a heuristic nature are widely used in most applications. For instance, in some existing work on adaptive finite element approximation of optimal design, the mesh refinement is guided by a posteriori error estimators *solely* from the state equation (or the displacement) for a fixed control (or design). Thus error information from approximation of the control (design) is not utilized. Although these methods may work well in some particular applications, they cannot be applied confidently in general. It is unlikely that the potential power of adaptive finite element approximation has been fully utilized due to the lack of more sophisticated a posteriori error indicators.

Very recently, some error estimators of residual type were developed in [8, 9, 30, 31, 33]. These error estimators are based on a posteriori estimation of the discretization error for the state *and* the control (design). When there is no constraint in a control problem, normally the optimality conditions consist of coupled partial differential equations only. Consequently, one may be able to write down the dual system of the *whole* optimality conditions and then apply the weighted a posteriori error estimation technique to obtain a posteriori estimators for the *objective functional* approximation error of the control problem; see [8, 9]. In many applications (like parameter estimation), it is more interesting to obtain a posteriori error estimators for the control approximation error [22]. Furthermore, there frequently exist some constraints for the control in applications. In such cases, the optimality conditions often contain a variational inequality and then have some very different properties. Thus it does not always seem to be possible to apply the techniques used in [8, 9] to constrained control problems.

In our work, constrained cases are studied via residual estimation using the norms of energy type. A posteriori error estimators are derived for quite general constrained control problems governed by the elliptic equations (see [30, 31, 33]) with upper error bounds. However, these error estimators have yet to be applied to adaptive finite element methods. Indeed, numerical experiments indicated that these estimators tend to over-refine the computational meshes. Thus the resulting computational meshes may not be efficient in reducing approximation errors. It seems that one has to derive sharper error estimators in order to obtain more efficient meshes. This seems to be possible at least for a class of control problems, which are frequently met in applications. More details on these will be given in section 3.

In this paper, we consider the convex optimal control problem

$$\begin{cases} \min\limits_{u\in K}\{g(y)+h(u)\}, \\ \\ -\mathrm{div}(A\nabla y) = f + Bu \quad \text{in } \Omega,\ y|_{\partial\Omega} = 0, \end{cases} \tag{1.1}$$

where g and h are given convex functionals, K is a closed convex set, and B is a continuous linear operator. The details will be specified later. The main objective of this work is to derive sharp a posteriori error estimators for some frequently met optimal control problems. A number of new techniques have to be introduced in order to obtain such estimators. Our numerical tests indicate that these improved error estimators indeed lead to efficient computational meshes.

The paper is organized as follows: In section 2, we describe the finite element approximation for the convex optimal control problem (1.1). In section 3, we derive

error estimates for the problem with an obstacle constraint. Both upper bounds and lower bounds are established with attention on their equivalence. In section 4, numerical experiments will be carried out, with particular attention to testing the influence of various indicators on the mesh construction.

2. The elliptic optimal control problem and finite element approximation. In this section, we describe the elliptic optimal control problem and its finite element approximation. Let Ω and Ω_U be two bounded open sets in $\mathbf{R}^n (n \leq 3)$ with the Lipschitz boundaries $\partial\Omega$ and $\partial\Omega_U$. We denote by $C^0(\bar{\Omega})$ the space of continuous functions on $\bar{\Omega}$. We adopt the standard notation $W^{m,q}(\Omega)$ for Sobolev spaces on Ω with norm $\|\cdot\|_{m,q,\Omega}$ and seminorm $|\cdot|_{m,q,\Omega}$(see (1.2) of [16, p. 2]). For $q = 2$, we denote $W^{m,2}(\Omega)$ by $H^m(\Omega)$ with norm $\|\cdot\|_{m,\Omega} := \|\cdot\|_{m,2,\Omega}$ and seminorm $|\cdot|_{m,\Omega} := |\cdot|_{m,2,\Omega}$. We set $H_0^1(\Omega) = \{v \in H^1(\Omega) : v|_{\partial\Omega} = 0\}$. In addition, c or C denotes a general positive constant independent of h.

In the rest of the paper, we shall take the state space $Y = H_0^1(\Omega)$, the control space $U = L^2(\Omega_U)$ with the inner product $(\cdot,\cdot)_U$, and $H = L^2(\Omega)$ with the inner product $(\cdot,\cdot)$. We wish to study the finite element approximation of the distributed elliptic convex optimal control problem (1.1). Assume that g and h are convex functionals which are continuously differentiable on $H = L^2(\Omega)$ and $U = L^2(\Omega_U)$, respectively, and h is further strictly convex. Suppose that K is a closed convex set in the control space U, $f \in L^2(\Omega)$, B is a continuous linear operator from U to $H \subset Y'$ (the dual space of Y), and

$$A(\cdot) = (a_{i,j}(\cdot))_{n\times n} \in (L^\infty(\Omega))^{n\times n}$$

such that there is a constant $c > 0$ satisfying, for any vector $\xi \in \mathbf{R}^n$,

$$(A\xi)\cdot\xi \geq c|\xi|^2.$$

We further assume that $h(u) \to +\infty$ as $\|u\|_{0,\Omega_U} \to \infty$, the functional $g(\cdot)$ is bounded below, and

$$|(g'(v) - g'(w), q)| \leq C\|v - w\|_{1,\Omega}\|q\|_{1,\Omega} \qquad \forall v, w, q \in Y. \tag{2.1}$$

To consider the finite element approximation of the above optimal control problem, here we give it a weak formula

$$\text{(CCP)} \begin{cases} \min\limits_{u\in K}\{g(y) + h(u)\}, \\ \\ a(y, w) = (f + Bu, w) \quad \forall w \in Y = H_0^1(\Omega), \end{cases} \tag{2.2}$$

where

$$a(v, w) = \int_\Omega (A\nabla v)\cdot\nabla w \quad \forall v, w \in H^1(\Omega),$$
$$(f, w) = \int_\Omega fw \quad \forall f, w \in L^2(\Omega).$$

Under these assumptions, the control problem (CCP) has a unique solution (y, u), and a pair (y, u) is the solution of (CCP) if and only if there is a costate $p \in Y$ such

that the triplet (y, p, u) satisfies the following optimality conditions (see [27]):

$$\text{(CCP-OPT)}\begin{cases} a(y,w) = (f + Bu, w) \quad \forall w \in Y = H_0^1(\Omega), \\ a(q,p) = (g'(y), p) \quad \forall q \in Y = H_0^1(\Omega), \\ (h'(u) + B^*p, v - u)_U \geq 0 \quad \forall v \in K \subset U = L^2(\Omega_U), \end{cases} \tag{2.3}$$

where B^* is the adjoint operator of B and g' and h' are the derivatives of g and h. Here g' and h' have been viewed as functions in $H = L^2(\Omega)$ and $U = L^2(\Omega_U)$, respectively, using the well-known representation theorem in a Hilbert space.

Let us consider the finite element approximation of the above control problem. For ease of exposition, we consider only n-simplex, conforming Lagrange elements. Also, we assume that Ω and Ω_U are polygonal. Let T^h be a partitioning of Ω into disjoint open regular n-simplices τ so that $\bar{\Omega} = \cup_{\tau \in T^h} \bar{\tau}$. Each element has at most one face on $\partial\Omega$, and the adjoining elements $\bar{\tau}$ and $\bar{\tau}'$ have either only one common vertex or a whole edge or a whole face if τ and $\tau' \in T^h$. Let h_τ denote the diameter of the element τ in T^h. Associated with T^h is a finite dimensional subspace S^h of $C^0(\bar{\Omega})$ such that $v_h|_\tau$ are polynomials of k-order ($k \geq 1$) for all $v_h \in S^h$ and $\tau \in T^h$. Denote $\{P_i\}$ $(i = 1, 2, \ldots, J)$ the vertex set associated with T^h. Let $Y^h = V_0^h := S^h \cap Y$.

Similarly, we have a regular partitioning of Ω_U, and we use the following corresponding notation: T_U^h, τ_U, h_{τ_U} and P_i^U $(i = 1, 2, \ldots, J_U)$. Associated with T_U^h is another finite dimensional subspace W_U^h of $L^2(\Omega_U)$ such that $v_h|_{\tau_U}$ are polynomials of k-order ($k \geq 0$) for all $v_h \in W_U^h$ and $\tau_U \in T_U^h$. Note here that there is no requirement for the continuity or boundary conditions. Let $U^h = W_U^h \subset U = L^2(\Omega_U)$.

Due to the limited regularity of the optimal control u (at most in $H^1(\Omega_U)$ in general), here we will consider only the piecewise constant space for the control approximation, while higher-order finite spaces may be used for the state and costate.

Then a possible finite element approximation of (CCP) is as follows:

$$(\text{CCP})^h \begin{cases} \min\limits_{u_h \in K^h} \{g(y_h) + h(u_h)\}, \\ a(y_h, w_h) = (f + Bu_h, w_h) \quad \forall w_h \in Y^h, \end{cases} \tag{2.4}$$

where K^h is a closed convex set in U^h such that there are $v_h \in K^h$ converging to an element $v \in K$ in U. It follows that the control problem $(\text{CCP})^h$ has a unique solution (y_h, u_h) and that a pair $(y_h, u_h) \in Y^h \times U^h$ is the solution of $(\text{CCP})^h$ if and only if there is a costate $p_h \in Y^h$ such that the triplet (y_h, p_h, u_h) satisfies the following optimality conditions:

$$(\text{CCP-OPT})^h \begin{cases} a(y_h, w_h) = (f + Bu_h, w_h) \quad \forall w_h \in Y^h \subset Y = H_0^1(\Omega), \\ a(q_h, p_h) = (g'(y_h), q_h) \quad \forall q_h \in Y^h \subset Y = H_0^1(\Omega), \\ (h'(u_h) + B^*p_h, v_h - u_h)_U \geq 0 \quad \forall v_h \in K^h \subset U = L^2(\Omega_U). \end{cases} \tag{2.5}$$

It follows that (y_h, p_h, u_h) is uniformly bounded in $Y \times Y \times U$. This is because $g(y_h) + h(u_h)$ is uniformly bounded due to the above assumption on K^h. Thus $\|u_h\|_U$ is also uniformly bounded. Then it follows from (2.5) and (2.1) that $\|y_h\|_Y$ and $\|p_h\|_Y$ are uniformly bounded.

The finite element approximation solution must be solved by using some mathematical programming algorithms such as the conjugate gradient method, the interior point method, and the SQP algorithms. This is a very active research area and is too large to be reviewed here even very briefly. Some of the recent progress in this area has been summarized in [14].

3. Sharp a posteriori error estimators. Deriving a posteriori error estimators for the finite element approximation of the control problem (CCP) is not an easy task since the triplet (y, p, u) is the solution of the coupled system (CCP-OPT). Although there is much work on a priori error estimates for finite element approximation of optimal control problems, as seen in section 1, there are substantial differences between a priori error estimates and a posteriori error estimates for such control problems. Only very recently, some a posteriori error estimators have been derived in the literature. For the control problem (CCP), for instance, the following error estimators have been derived in [31] and [35], assuming that $(h'(u_h) + B^*p_h)|_{\tau_U} \in H^1(\tau_U)$ for any $\tau_U \in T_U^h$:

$$\|u_h - u\|_{0,\Omega_U}^2 + \|y_h - y\|_{1,\Omega}^2 + \|p_h - p\|_{1,\Omega}^2 \le C\left(\bar{\eta}_1^2 + \sum_{i=2}^{5} \eta_i^2\right) = C\bar{\eta}^2, \tag{3.1}$$

where

$$\begin{aligned}
\bar{\eta}_1^2 &= \sum_{\tau_U \in T_U^h} h_{\tau_U}^2 \|\nabla(h'(u_h) + B^*p_h)\|_{0,\tau_U}^2, \\
\eta_2^2 &= \sum_{\tau \in T^h} h_\tau^2 \int_\tau (f + Bu_h + \operatorname{div}(A\nabla y_h))^2, \\
\eta_3^2 &= \sum_{l \in \partial T^h} h_l \int_l [(A\nabla y_h) \cdot \mathbf{n}]^2, \\
\eta_4^2 &= \sum_{\tau \in T^h} h_\tau^2 \int_\tau (g'(y_h) + \operatorname{div}(A^*\nabla p_h))^2, \\
\eta_5^2 &= \sum_{l \in \partial T^h} h_l \int_l [(A^*\nabla p_h) \cdot \mathbf{n}]^2,
\end{aligned} \tag{3.2}$$

where h_l is the diameter of the face l, and the A-normal derivative jump over the interior face l is defined by

$$[(A\nabla v_h) \cdot \mathbf{n}]|_l = ((A\nabla v_h)|_{\partial\tau_l^1} - (A\nabla v_h)|_{\partial\tau_l^2}) \cdot \mathbf{n},$$

with $\mathbf{n}$ being the unit outer normal vector of τ_l^1 on $l = \bar{\tau}_l^1 \cap \bar{\tau}_l^1$. The A^*-normal derivative jump is similarly defined for the transposed matrix A^* of A.

However, major improvements on these error estimators are much needed in order that they can be used to guide mesh adaptivity efficiently in solving the optimal control problem numerically. For example, it does not seem that they are *always* sharp for the constrained cases, and this can be seen from Figures 4.4 and 4.6 in Example 4.1 of section 4, where it is clear that $|u - u_h|$ has a very different profile (the left of Figure 4.4) from that of $\bar{\eta}_1$ (or $\bar{\eta}$) (the left of Figure 4.6). Consequently, the mesh refinement adjustment schemes based on $\bar{\eta}$ may be inefficient. In Example 4.1, the state and costate are very smooth, but the optimal control u has the gradient

jumps across the free boundary, which is the boundary of the zero set $\{x : u(x) = 0\}$, as seen in Figure 4.1. This causes large control approximation errors along the free boundary, as seen from the left of Figures 4.3 and 4.4. Thus an efficient computational mesh for the control should have a higher density of nodes around the free boundary, as those in Figure 4.2. However, the mesh adjustment guided by $\bar{\eta}_1$ did not achieve this goal well, as seen from Figure 4.5. In fact, the resulting mesh even produced a larger approximation error for the control than the uniform mesh of the same size. A sharp error estimator will lead to much more efficient computational meshes, as seen in Figure 4.2.

In this section, we study sharp error estimates for finite element approximation of the convex control problem (CCP). It follows that $\bar{\eta}$ consists of three parts: The part $\bar{\eta}_1$ is contributed from the approximation error of the variational inequality, and $\eta_2^2+\eta_3^2$, $\eta_4^2+\eta_5^2$ result from the approximation error of the state and costate equations. It is well known that $\eta_2^2+\eta_3^2$ and $\eta_4^2+\eta_5^2$ are sharp error estimators for the state and costate equations. Therefore, the key to our purpose is to improve $\bar{\eta}_1^2$. However, it is difficult to derive improved estimates without knowing explicit structures of the control constraint sets K and K^h; the methods and techniques to be developed will depend heavily on these details. Here we derive a posteriori error estimators with both upper bounds and lower bounds for a class of convex sets K of obstacle type, which are most frequently met in real applications. We achieved this by exploring the special structure of the constraint sets. The ideas are applicable to some other control problems, e.g., the boundary control problems.

We shall first consider the constraint of a single obstacle

$$K = \{v \in U : v \geq \phi\}, \qquad K^h = U^h \cap K,$$

and then we will extend the results to more general cases.

We define the coincidence set (contact set) Ω_U^- and the noncoincidence set (noncontact set) Ω_U^+ as follows:

$$\Omega_U^- = \{x \in \Omega_U : u(x) = \phi(x)\}, \quad \Omega_U^+ = \{x \in \Omega_U : u(x) > \phi(x)\}.$$

It can be seen that the inequality in (2.3) is equivalent to the following:

$$h'(u) + B^*p \geq 0, \quad u \geq \phi, \quad (h'(u) + B^*p)(u - \phi) = 0, \quad \text{a.e. in } \Omega_U. \tag{3.3}$$

We shall show that the quantity $(h'(u_h) + B^*p_h)|_{\Omega_U^-}$ can be mostly removed from the error indicator $\bar{\eta}$ in this case, which enables us to obtain sharp error estimates. To make the presentation of our approach clearer and less technical, we shall first derive sharp error estimators containing an a priori quantity and then approximate it using an a posteriori quantity so that the estimators can be easily applied in numerical computations. Let us note that some approximations of a priori quantities are also used in [9].

In the following, we assume that there is a constant $c > 0$ such that

$$(h'(v) - h'(w), v - w)_U \geq c\|v - w\|_{0,\Omega_U}^2 \qquad \forall v, w \in U. \tag{3.4}$$

3.1. Upper error bounds. We first consider the case of a constant obstacle $\phi(x) \equiv \phi_0$. We define

$$\Omega_h^+ = \{\cup\bar{\tau}_U \ : \ \tau_U \subset \Omega_U^+, \ \tau_U \in T_U^h\}, \qquad \Omega_h^- = \{\cup\bar{\tau}_U \ : \ \tau_U \subset \Omega_U^-, \ \tau_U \in T_U^h\},$$
$$\Omega_h^b = \Omega_U \backslash (\Omega_h^+ \cup \Omega_h^-), \qquad \Omega_h^{+b} = \Omega_h^+ \cup \Omega_h^b, \qquad \Omega_h^{-b} = \Omega_h^- \cup \Omega_h^b,$$

and denote by χ_Q the characteristic function of Q. Let ∂T^h be the set consisting of all of the faces l of any $\tau \in T^h$ such that l is not on $\partial\Omega$. Let h_l be the diameter of the face l. We need the following lemmas in deriving residual-type a posteriori error estimates.

LEMMA 3.1 (see [10]). *Let $\pi_h : C^0(\bar{\Omega}) \to S^h$ be the standard Lagrange interpolation operator such that*

$$\pi_h v := \sum_i v(\mathbf{a}_i)\varphi_i,$$

where $\mathbf{a}_i$ are the nodes on $\bar{\Omega}$ and φ_i are the corresponding shape functions. Then, for $m = 0, 1$ and $n/2 < q \le \infty$,

$$\|v - \pi_h v\|_{m,q,\tau} \le C h_\tau^{2-m} |v|_{2,q,\tau} \quad \forall v \in W^{2,q}(\Omega), \tag{3.5}$$

where the constant C depends only on Ω and the minimum angle of the simplices in T^h.

LEMMA 3.2 (see [21]). *For all $v \in W^{1,q}(\Omega), 1 \le q \le \infty$,*

$$\|v\|_{0,q,\partial\tau} \le C(h_\tau^{-1/q}\|v\|_{0,q,\tau} + h_\tau^{1-1/q}|v|_{1,q,\tau}). \tag{3.6}$$

We need another operator $\hat{\pi}_h$: the local averaging interpolation operator defined in [42], which can be applied to functions not necessarily continuous, preserves the homogeneous boundary conditions and is stable in the $W^{1,q}$-norm. The full definition of $\hat{\pi}_h$ is rather long. Thus the readers are referred to [42]. Fortunately, we need only to use one of its properties, which is stated in the following lemma.

LEMMA 3.3. *Let $\hat{\pi}_h : W^{1,q}(\Omega) \to S^h$ be the local averaging interpolation operator defined in* (2.13) *of* [42]. *For $m = 0, 1$ and $1 \le q \le \infty$,*

$$|v - \hat{\pi}_h v|_{m,q,\tau} \le C \sum_{\bar{\tau}' \cap \bar{\tau} \ne \emptyset} h_\tau^{1-m} |v|_{1,q,\tau'} \qquad \forall v \in W^{1,q}(\Omega). \tag{3.7}$$

LEMMA 3.4. *Let $\pi_h^a : L^1(\Omega_U) \to W_U^h$ be the integral averaging operator such that*

$$(\pi_h^a v)|_{\tau_U} := \frac{1}{|\tau_U|} \int_{\tau_U} v \qquad \forall \tau_U \in T_U^h.$$

Then, for $m = 0, 1$ and $1 \le q \le \infty$,

$$\|v - \pi_h^a v\|_{0,q,\tau_U} \le C h_{\tau_U}^m |v|_{m,q,\tau_U} \qquad \forall v \in W^{m,q}(\Omega_U). \tag{3.8}$$

Proof. The result is trivial for $m = 0$. For $m = 1$, we note that $\pi_h^a v|_{\tau_U} = v|_{\tau_U}$ if v is a constant on τ_U. Thus (3.8) can be proved by the standard techniques in the finite element method [10]. ∎

We first give some upper bounds for $u - u_h$ in the L^2-norm and for $y - y_h$, $p - p_h$ in the H^1-norm. We shall use the following inequality:

$$|(Bv, w)| = |(v, B^* w)_U| \le C\|v\|_{0,\Omega_U}\|w\|_{1,\Omega} \qquad \forall v \in U, w \in Y, \tag{3.9}$$

which is held from our assumptions on the operator B.

THEOREM 3.1. *Let (y, p, u) and (y_h, p_h, u_h) be the solutions of* (2.3) *and* (2.5), *respectively. Let the obstacle ϕ be a constant ϕ_0. Assume that conditions* (3.4), (2.1), *and* (3.9) *hold, and $(h'(u_h) + B^* p_h)|_{\tau_U} \in H^1(\tau_U)$ for any $\tau_U \in T_U^h$. Then*

$$\|u_h - u\|_{0,\Omega_U}^2 + \|y_h - y\|_{1,\Omega}^2 + \|p_h - p\|_{1,\Omega}^2 \le C \sum_{i=1}^5 \eta_i^2, \tag{3.10}$$

where

$$\eta_1^2 = \sum_{\tau_U \in T_U^h} h_{\tau_U}^2 \|\nabla(h'(u_h) + B^* p_h)\chi_{\Omega_h^{+b}}\|_{0,\tau_U}^2, \tag{3.11}$$

$$\eta_2^2 = \sum_{\tau \in T^h} h_\tau^2 \int_\tau (f + Bu_h + \operatorname{div}(A\nabla y_h))^2,$$

$$\eta_3^2 = \sum_{l \in \partial T^h} h_l \int_l [(A\nabla y_h)\cdot \mathbf{n}]^2,$$

$$\eta_4^2 = \sum_{\tau \in T^h} h_\tau^2 \int_\tau (g'(y_h) + \operatorname{div}(A^*\nabla p_h))^2,$$

$$\eta_5^2 = \sum_{l \in \partial T^h} h_l \int_l [(A^*\nabla p_h)\cdot \mathbf{n}]^2.$$

Proof. We first estimate the error $\|u - u_h\|_{0,\Omega_U}^2$. It follows from the assumption (3.4) and the inequalities (2.3) and (2.5) that, for any $v_h \in K^h$,

$$\begin{aligned} & c\|u - u_h\|_{0,\Omega_U}^2 \\ &\le (h'(u), u - u_h)_U - (h'(u_h), u - u_h)_U \\ &\le (-B^* p, u - u_h)_U - (h'(u_h), u - u_h)_U + (h'(u_h) + B^* p_h, v_h - u_h)_U \\ &= (h'(u_h) + B^* p_h, v_h - u)_U + (B^*(p_h - p), u - u_h)_U. \end{aligned} \tag{3.12}$$

We introduce y_{u_h} and p_{u_h}, defined by

$$a(y_{u_h}, w) = (f + Bu_h, w) \qquad \forall w \in Y, \tag{3.13}$$

$$a(q, p_{u_h}) = (g'(y_{u_h}), q) \qquad \forall q \in Y. \tag{3.14}$$

It follows from (2.3), (3.13), and (3.14) that

$$a(y_{u_h} - y, w) = (B(u_h - u), w) \qquad \forall w \in Y, \tag{3.15}$$

$$a(q, p_{u_h} - p) = (g'(y_{u_h}) - g'(y), q) \qquad \forall q \in Y. \tag{3.16}$$

Taking $w = p_{u_h} - p$ in (3.15) and $q = y_{u_h} - y$ in (3.16), we have, due to the convexity of g,

$$(B(u_h - u), p_{u_h} - p) = (g'(y_{u_h}) - g'(y), y_{u_h} - y) \ge 0.$$

Using (3.12) together with (3.9) gives

$$\begin{aligned} & c\|u - u_h\|_{0,\Omega_U}^2 \\ &\le (h'(u_h) + B^* p_h, v_h - u)_U + (B^*(p_h - p_{u_h}), u - u_h)_U - (p_{u_h} - p, B(u_h - u)) \\ &\le \sum_{\tau_U \in T_U^h} (h'(u_h) + B^* p_h, v_h - u)_{\tau_U} + C\|p_h - p_{u_h}\|_{1,\Omega}^2 + \frac{c}{2}\|u - u_h\|_{0,\Omega_U}^2. \end{aligned}$$

Now take $v_h = \pi_h^a u \in K^h$ defined in Lemma 3.4. Then we have

$$\begin{aligned} (h'(u_h) + B^* p_h, v_h - u)_{\tau_U} &= ((I - \pi_h^a)(h'(u_h) + B^* p_h), (\pi_h^a - I)(u - u_h))_{\tau_U} \\ &\le Ch_{\tau_U}\|\nabla(h'(u_h) + B^* p_h)\|_{0,\tau_U}\|u - u_h\|_{0,\tau_U} \\ &\le Ch_{\tau_U}^2\|\nabla(h'(u_h) + B^* p_h)\|_{0,\tau_U}^2 + \frac{c}{4}\|u - u_h\|_{0,\tau_U}^2. \end{aligned}$$

Noting that $(v_h - u)|_{\tau_U} = (\pi_h^a - I)u|_{\tau_U} = 0$ for any $\tau_U \in \Omega_U \backslash \Omega_h^{+b}$, we obtain

(3.17)
$$\begin{aligned} \|u - u_h\|_{0,\Omega_U}^2 &\leq C \sum_{\tau_U \in \Omega_h^{+b}} h_{\tau_U}^2 \|\nabla(h'(u_h) + B^* p_h)\|_{0,\tau_U}^2 + C\|p_h - p_{u_h}\|_{1,\Omega_U}^2 \\ &= C\eta_1^2 + C\|p_h - p_{u_h}\|_{1,\Omega_U}^2. \end{aligned}$$

The second step is to estimate the error $\|p_{u_h} - p_h\|_{1,\Omega}$. Let $e_p = p_{u_h} - p_h$. Then it follows from $(2.4)_2$, (3.14), and (2.1) that

$$\begin{aligned} & c\|p_{u_h} - p_h\|_{1,\Omega}^2 \leq a(e_p, p_{u_h}) - a(e_p, p_h) \\ = \; & (g'(y_{u_h}), e_p) - a(e_p - \hat{\pi}_h e_p, p_h) - (g'(y_h), \hat{\pi}_h e_p) \\ = \; & \sum_{\tau \in T^h} \int_\tau (g'(y_h) + \operatorname{div}(A^* \nabla p_h))(e_p - \hat{\pi}_h e_p) \\ & - \sum_{l \in \partial T^h} \int_l [(A^* \nabla p_h) \cdot \mathbf{n}](e_p - \hat{\pi}_h e_p) + (g'(y_{u_h}) - g'(y_h), e_p) \\ \leq \; & C \sum_{\tau \in T^h} h_\tau^2 \int_\tau (g'(y_h) + \operatorname{div}(A^* \nabla p_h))^2 + C \sum_{l \in \partial T^h} h_l \int_l [(A^* \nabla p_h) \cdot \mathbf{n}]^2 \\ & + C\|y_{u_h} - y_h\|_{1,\Omega}^2 + \frac{c}{2}\|e_p\|_{1,\Omega}^2, \end{aligned}$$

where we have used Lemma 3.3 to obtain

$$\|e_p - \hat{\pi}_h e_p\|_{0,\tau} \leq C h_\tau \left(\sum_{\bar{\tau}' \cap \bar{\tau} \neq \emptyset} |e_p|_{1,\tau'}^2 \right)^{1/2} \tag{3.18}$$

and Lemmas 3.2 and 3.3 to have, assuming $l \subset \bar{\tau}$,

$$\begin{aligned} \|e_p - \hat{\pi}_h e_p\|_{0,l} &\leq C(h_\tau^{-1/2}\|e_p - \hat{\pi}_h e_p\|_{0,\tau} + h_\tau^{1/2}|e_p - \hat{\pi}_h e_p|_{1,\tau}) \\ &\leq C h_\tau^{1/2} \left(\sum_{\bar{\tau}' \cap \bar{\tau} \neq \emptyset} |e_p|_{1,\tau'}^2 \right)^{1/2}. \end{aligned} \tag{3.19}$$

Thus we have

$$\|p_{u_h} - p_h\|_{1,\Omega}^2 \leq C(\hat{\eta}_4^2 + \hat{\eta}_5^2) + C\|y_{u_h} - y_h\|_{1,\Omega}^2. \tag{3.20}$$

The third step is thus to estimate the error $\|y_{u_h} - y_h\|_{1,\Omega}$. Let $e_y = y_{u_h} - y_h$, and let $\hat{\pi}_h$ be the interpolator in Lemma 3.3. It can be seen that $a(e_y, \hat{\pi}_h e_y) = 0$ due to the Galerkin orthogonality $a(e_y, w_h) = 0 \; \forall w_h \in Y^h$ from $(2.5)_1$ and (3.13). Then it follows from (2.5), (3.13), (3.6), and (3.7) that

$$
\begin{aligned}
c\|y_{u_h}-y_h\|_{1,\Omega}^2 &\le a(e_y,e_y)=a(e_y,e_y-\hat{\pi}_h e_y)\\
&=\sum_{\tau\in T^h}\int_\tau (f+Bu_h+\operatorname{div}(A\nabla y_h))(e_y-\hat{\pi}_h e_y)\\
&\qquad -\sum_{l\in\partial T^h}\int_l [(A\nabla y_h)\cdot\mathbf{n}](e_y-\hat{\pi}_h e_y)\\
&\le C\sum_{\tau\in T^h} h_\tau^2\int_\tau (f+Bu_h+\operatorname{div}(A\nabla y_h))^2\\
&\qquad +C\sum_{l\in\partial T^h} h_l\int_l [(A\nabla y_h)\cdot\mathbf{n}]^2+\frac{c}{2}\|e_y\|_{1,\Omega}^2,
\end{aligned}
$$

where we have bounded $\|e_y-\hat{\pi}_h e_y\|_{0,\tau}$ and $\|e_y-\hat{\pi}_h e_y\|_{0,l}$ as in (3.18) and (3.19). Thus we have

$$
\|y_{u_h}-y_h\|_{1,\Omega}^2\le C(\hat{\eta}_2^2+\hat{\eta}_3^2). \tag{3.21}
$$

Finally, by noting that, from (3.15), (3.16), (3.9), and (2.1), we have

$$
\|y_{u_h}-y\|_{1,\Omega}\le C\|u_h-u\|_{0,\Omega_U}, \tag{3.22}
$$

$$
\|p_{u_h}-p\|_{1,\Omega}\le C\|y_{u_h}-y\|_{1,\Omega}\le C\|u_h-u\|_{0,\Omega_U}, \tag{3.23}
$$

we combine (3.17), (3.20), and (3.21) to obtain

$$
\begin{aligned}
&\|u-u_h\|_{0,\Omega_U}^2+\|y-y_h\|_{1,\Omega}^2+\|p-p_h\|_{1,\Omega}^2\\
\le\ &\|u-u_h\|_{0,\Omega_U}^2+2\left(\|y-y_{u_h}\|_{1,\Omega}^2+\|p-p_{u_h}\|_{1,\Omega}^2\right)\\
&+2\left(\|y_{u_h}-y_h\|_{1,\Omega}^2+\|p_{u_h}-p_h\|_{1,\Omega}^2\right)\\
\le\ &C\|u-u_h\|_{0,\Omega_U}^2+2\left(\|y_{u_h}-y_h\|_{1,\Omega}^2+\|p_{u_h}-p_h\|_{1,\Omega}^2\right)\le\ C\sum_{i=1}^5\eta_i^2.
\end{aligned}
$$

Therefore, the proof is completed. ∎

In many applications, we are mostly interested in computing the values of the state and the control. In such cases, it is more useful to bound the errors in the L^2-norm to derive sharper estimators, which are given in the following theorem. We shall use the following condition:

$$
|(Bv,w)|=|(v,B^*w)_U|\le C\|v\|_{0,\Omega_U}\|w\|_{0,\Omega}\qquad\forall v\in U, w\in Y, \tag{3.24}
$$

which is held from our assumptions. We shall assume the following condition:

$$
|(g'(v)-g'(w),q)|\le C\|v-w\|_{0,\Omega}\|q\|_{2,\Omega}\qquad\forall v,w\in Y,\ q\in H^2(\Omega). \tag{3.25}
$$

THEOREM 3.2. *Assume that all of the conditions of Theorem* 3.1 *and* (3.25) *are satisfied except that* (3.9) *is replaced with* (3.24). *Assume that* Ω *is convex. Then*

$$
\|u-u_h\|_{0,\Omega_U}^2+\|y-y_h\|_{0,\Omega}^2+\|p-p_h\|_{0,\Omega}^2\le C\left(\eta_1^2+\sum_{i=2}^5\hat{\eta}_i^2\right),
$$

where η_1^2 *is defined in Theorem* 3.1 *and*

$$\hat{\eta}_2^2 = \sum_{\tau\in T^h} h_\tau^4 \int_\tau (f + Bu_h + \operatorname{div}(A\nabla y_h))^2,$$

$$\hat{\eta}_3^2 = \sum_{l\in\partial T^h} h_l^3 \int_l [(A\nabla y_h)\cdot \mathbf{n}]^2,$$

$$\hat{\eta}_4^2 = \sum_{\tau\in T^h} h_\tau^4 \int_\tau (g'(y_h) + \operatorname{div}(A^*\nabla p_h))^2,$$

$$\hat{\eta}_5^2 = \sum_{l\in\partial T^h} h_l^3 \int_l [(A^*\nabla p_h)\cdot \mathbf{n}]^2.$$

Proof. Again, we first estimate the error $\|u-u_h\|^2_{0,\Omega_U}$. By the same argument as in the proof of Theorem 3.1 but using (3.24), we have

$$\|u-u_h\|^2_{0,\Omega_U} \le C\eta_1^2 + C\|p_h - p_{u_h}\|^2_{0,\Omega}. \tag{3.26}$$

To estimate $\|p_h - p_{u_h}\|^2_{0,\Omega}$, we use the dual technique. Consider the following auxiliary problems: Find $\xi \in H_0^1(\Omega)$ and $\zeta \in H_0^1(\Omega)$ such that

$$a(w,\xi) = (f_1, w) \quad \forall w \in Y, \tag{3.27}$$

$$a(\zeta, q) = (f_2, q) \quad \forall q \in Y. \tag{3.28}$$

It follows from the well-known regularity results that

$$\|\xi\|_{2,\Omega} \le C\|f_1\|_{0,\Omega}, \quad \|\zeta\|_{2,\Omega} \le C\|f_2\|_{0,\Omega}.$$

Let $f_2 = p_{u_h} - p_h$ in (3.28) and denote by $\pi_h : C^0(\bar{\Omega}) \to Y^h$ the standard Lagrange interpolation operator. It follows from $(2.5)_2$ and (3.14) that

$$\begin{aligned}
\|p_{u_h} - p_h\|^2_{0,\Omega} &= (f_2, p_h(u_h) - p_h) = a(\zeta, p_{u_h}) - a(\zeta, p_h)\\
&= (g'(y_{u_h}), \zeta) - a(\zeta - \pi_h\zeta, p_h) - (g'(y_h), \pi_h\zeta)\\
&= \sum_{\tau\in T^h}\int_\tau \operatorname{div}(A^*\nabla p_h)(\zeta - \pi_h\zeta) - \sum_{l\in\partial T^h}\int_l [(A^*\nabla p_h)\cdot\mathbf{n}](\zeta - \pi_h\zeta)\\
&\qquad + (g'(y_{u_h}), \zeta) - (g'(y_h), \pi_h\zeta)\\
&= \sum_{\tau\in T^h}\int_\tau (g'(y_h) + \operatorname{div}(A^*\nabla p_h))(\zeta - \pi_h\zeta)\\
&\quad - \sum_{l\in\partial T^h}\int_l [(A^*\nabla p_h)\cdot\mathbf{n}](\zeta - \pi_h\zeta) + (g'(y_{u_h}) - g'(y_h), \zeta).
\end{aligned}$$

By using Lemmas 3.1 and 3.2,

$$\|\zeta - \pi_h\zeta\|_{0,\tau} \le Ch_\tau^2|\zeta|_{2,\tau}, \tag{3.29}$$

$$\|\zeta - \pi_h\zeta\|_{0,l} \le C(h_\tau^{-1/2}\|\zeta - \pi_h\zeta\|_{0,\tau} + h_\tau^{1/2}|\zeta - \pi_h\zeta|_{1,\tau}) \le Ch_\tau^{3/2}|\zeta|_{2,\tau}, \tag{3.30}$$

where $l \subset \bar{\tau}$. Then it follows from (3.25) that

$$\begin{aligned}\|p_{u_h} - p_h\|_{0,\Omega}^2 &\le C \sum_{\tau\in T^h} h_\tau^2 \|g'(y_h) + \operatorname{div}(A^*\nabla p_h)\|_{0,\tau}\|\zeta\|_{2,\tau} \\ &\quad + C \sum_{l\in\partial T^h} h_l^{3/2} \left(\int_l [(A^*\nabla p_n)\cdot \mathbf{n}]^2\right)^{1/2} \|\zeta\|_{2,\tau} + C\|y_{u_h} - y_h\|_{0,\Omega}\|\zeta\|_{2,\Omega} \\ &\le C \sum_{\tau\in T^h} h_\tau^4 \int_\tau (g'(y_h) + \operatorname{div}(A^*\nabla p_h))^2 + C \sum_{l\in\partial T^h} h_l^3 \int_l [(A^*\nabla p_n)\cdot \mathbf{n}]^2 \\ &\quad + C\|y_{u_h} - y_h\|_{0,\Omega}^2 + \frac{1}{2}\|f_2\|_{0,\Omega}^2.\end{aligned}$$

Therefore, we have

$$\|p_{u_h} - p_h\|_{0,\Omega}^2 \le C(\hat{\eta}_4^2 + \hat{\eta}_5^2) + C\|y_{u_h} - y_h\|_{0,\Omega}^2. \tag{3.31}$$

The second step is again to estimate $\|y_{u_h} - y_h\|_{0,\Omega}^2$. Similarly, letting $f_1 = y_{u_h} - y_h$ in (3.27) gives

$$\begin{aligned}\|y_{u_h} - y_h\|_{0,\Omega}^2 = (f, y_{u_h} - y_h) &= a(y_{u_h} - y_h, \xi) = a(y_{u_h} - y_h, \xi - \pi_h\xi) \\ &= \sum_{\tau\in T^h} \int_\tau (f + Bu_h + \operatorname{div}(A\nabla y_h))(\xi - \pi_h\xi) \\ &\quad - \sum_{l\in\partial T^h} \int_l [(A\nabla y_h)\cdot \mathbf{n}](\xi - \pi_h\xi) \\ &\le C \sum_{\tau\in T^h} h_\tau^2 \int_\tau \|f + Bu_h + \operatorname{div}(A\nabla y_h)\|_{0,\tau}\|\xi\|_{2,\tau} \\ &\quad + C \sum_{l\in\partial T^h} h_l^{3/2} \left(\int_l [(A\nabla y_h)\cdot \mathbf{n}]^2\right)^{1/2} \|\xi\|_{2,\tau} \\ &\le C \sum_{\tau\in T^h} h_\tau^4 \int_\tau (f + Bu_h + \operatorname{div}(A\nabla y_h))^2 \\ &\quad + C \sum_{l\in\partial T^h} h_l^3 \int_l [(A\nabla y_h)\cdot \mathbf{n}]^2 + \frac{1}{2}\|f_1\|_{0,\Omega}^2,\end{aligned}$$

where we have estimated $\|\xi - \pi_h\xi\|_{0,\tau}$ and $\|\xi - \pi_h\xi\|_{0,l}$ as in (3.29) and (3.30). The above result leads to

$$\|y_{u_h} - y_h\|_{0,\Omega}^2 \le C(\hat{\eta}_2^2 + \hat{\eta}_3^2). \tag{3.32}$$

Then it follows from (3.26), (3.31), and (3.32) that

$$\|u - u_h\|_{0,\Omega_U}^2 \le C\left(\eta_1^2 + \sum_{i=2}^5 \hat{\eta}_i^2\right). \tag{3.33}$$

Finally, we estimate $\|y_h - y\|_{0,\Omega}$ and $\|p_h - p\|_{0,\Omega}$. It follows from (3.15), (3.16), and (2.1) that

$$\begin{aligned}\|y_h - y\|_{0,\Omega} &\le \|y_h - y_{u_h}\|_{0,\Omega} + \|y_{u_h} - y\|_{0,\Omega} \\ &\le \|y_h - y_{u_h}\|_{0,\Omega} + C\|u_h - u\|_{0,\Omega_U}\end{aligned}$$

and

$$\begin{aligned}\|p_h - p\|_{0,\Omega} &\le \|p_h - p_{u_h}\|_{0,\Omega} + \|p_{u_h} - p\|_{0,\Omega}\\ &\le \|p_h - p_{u_h}\|_{0,\Omega} + C\|y_{u_h} - y\|_{1,\Omega}\\ &\le \|p_h - p_{u_h}\|_{0,\Omega} + C\|u_h - u\|_{0,\Omega_U}.\end{aligned}$$

The above results, together with (3.31)–(3.33), yield

$$\|y_h - y\|_{0,\Omega}^2 + \|p_h - p\|_{0,\Omega}^2 \le C\left(\eta_1^2 + \sum_{i=2}^{5}\hat{\eta}_i^2\right). \tag{3.34}$$

Hence the proof is completed by combining (3.33) and (3.34). ∎

3.2. Lower error bounds. In this subsection, we wish to demonstrate that the error estimates obtained above are quite sharp by establishing lower error bounds for the finite element approximation. We start with the following lemma about the bubble functions, the proof of which can be found in [1, 45].

LEMMA 3.5. *Let $\tau \in T^h$. Let τ_l^1, τ_l^2 be two elements in T^h with a common edge (face) $l = \bar{\tau}_l^1 \cap \bar{\tau}_l^2$. For any constants B_τ and D_l, there exist polynomials $w_\tau \in H_0^1(\tau)$ and $w_l \in H_0^1(\tau_l^1 \cup \tau_l^2)$ such that, for $m = 0, 1$,*

$$\int_\tau B_\tau w_\tau = h_\tau^2 \int_\tau B_\tau^2, \qquad |w_\tau|_{m,\tau}^2 \le C h_\tau^{2(1-m)+2} \int_\tau B_\tau^2,$$

$$\int_l D_l w_l = h_l \int_l D_l^2, \qquad |w_l|_{m,\tau_l^1\cup\tau_l^2}^2 \le C h_l^{2(1-m)+1} \int_l D_l^2.$$

For ease of exposition, we assume that A is a constant matrix and Y^h is the piecewise linear finite element space. We also assume that there exists an integer $k \ge 0$ independent of h such that, for any $\tau_U \in T_U^h$, $(h'(u_h) + B^* p_h)|_{\tau_U}$ is a polynomial of k-order on τ_U. This assumption is needed to apply the inverse property in our proof below, and it may impose an implicit relationship between the meshes for the state and the control. We further assume that

$$\|h'(v) - h'(w)\|_{0,\Omega_U} \le C\|v - w\|_{0,\Omega_U} \qquad \forall v, w \in Y. \tag{3.35}$$

THEOREM 3.3. *Let (y, p, u) and (y_h, p_h, u_h) be the solutions of (2.3) and (2.5), respectively. Assume that A is a constant matrix, Y^h is the piecewise linear finite element space, $f \in L^2(\Omega)$, $\phi \equiv \phi_0$, $(h'(u_h) + B^* p_h)|_{\tau_U}$ is a polynomial of k-order on τ_U for any $\tau_U \in T_U^h$ with $k \ge 0$ independent of h, and the conditions (2.1), (3.24), and (3.35) hold. Then there exists a constant C depending on the matrix A and those constants in (2.1), (3.24), (3.35), and Lemma 3.5 such that*

$$\begin{aligned}\sum_{i=1}^{5}\eta_i^2 \;\le\; & C(\|u - u_h\|_{0,\Omega_U}^2 + \|y - y_h\|_{1,\Omega}^2 + \|p - p_h\|_{1,\Omega}^2)\\ &+ C\sum_{\tau\in T^h} h_\tau^2(\|F - \bar{F}\|_{0,\tau}^2 + \|G - \bar{G}\|_{0,\tau}^2)\\ &+ C\sum_{\tau_U\in T_U^h} h_{\tau_U}^2 \|\nabla(h'(u_h) + B^* p_h)\chi_{\Omega_h^b}\|_{0,\tau_U}^2,\end{aligned}$$

where η_i $(1 \le i \le 5)$ *are defined in Theorem* 3.1, $F = f + Bu_h$, $G = g'(y_h)$, $\bar{F}|_\tau = \int_\tau F/|\tau|$, *and* $\bar{G}|_\tau = \int_\tau G/|\tau|$.

Proof. From the optimality conditions (2.3), we deduce that $(h'(u)+B^*p)|_{\Omega_U^+} = 0$. It follows from the inverse property [10], (3.35), and (3.24) that

$$\begin{aligned}
\eta_1^2 &= \sum_{\tau_U \in T_U^h} h_{\tau_U}^2 (\|\nabla(h'(u_h) + B^*p_h)\chi_{\Omega_h^+}\|_{0,\tau_U}^2 + \|\nabla(h'(u_h) + B^*p_h)\chi_{\Omega_h^b}\|_{0,\tau_U}^2) \\
&\le C\|h'(u_h) + B^*p_h - h'(u) - B^*p\|_{0,\Omega_h^+}^2 + C \sum_{\tau_U \in T_U^h} h_{\tau_U}^2 \|\nabla(h'(u_h) + B^*p_h)\chi_{\Omega_h^b}\|_{0,\tau_U}^2 \\
&\le C(\|u - u_h\|_{0,\Omega_U}^2 + \|p - p_h\|_{1,\Omega}^2) + C \sum_{\tau_U \in T_U^h} h_{\tau_U}^2 \|\nabla(h'(u_h) + B^*p_h)\chi_{\Omega_h^b}\|_{0,\tau_U}^2.
\end{aligned}$$

To bound η_2^2, let w_τ be the bubble function as in Lemma 3.5 with $B_\tau = \bar{F}|_\tau$. It follows from (2.5) and (3.13) that

$$\begin{aligned}
\eta_2^2 &= \sum_{\tau \in T^h} h_\tau^2 \int_\tau F^2 \le 2 \sum_{\tau \in T^h} h_\tau^2 \int_\tau \{\bar{F}^2 + (F - \bar{F})^2\} \\
&= 2 \sum_{\tau \in T^h} \int_\tau \{w_\tau F + w_\tau(\bar{F} - F) + h_\tau^2 (F - \bar{F})^2\} \\
&= 2 \sum_{\tau \in T^h} \int_\tau (A\nabla(y_{u_h} - y_h)) \cdot \nabla w_\tau + 2 \sum_{\tau \in T^h} \int_\tau \{w_\tau(\bar{F} - F) + h_\tau^2 (F - \bar{F})^2\} \\
&\le C \sum_{\tau \in T^h} |y_{u_h} - y_h|_{1,\tau}^2 + \delta \sum_{\tau \in T^h} (|w_\tau|_{1,\tau}^2 + h_\tau^{-2}\|w_\tau\|_{0,\tau}^2) + C \sum_{\tau \in T^h} h_\tau^2 \int_\tau (F - \bar{F})^2 \\
&\le C(|y_{u_h} - y|_{1,\Omega}^2 + |y - y_h|_{1,\Omega}^2) + C\delta\eta_2^2 + C \sum_{\tau \in T^h} h_\tau^2 \int_\tau (F - \bar{F})^2.
\end{aligned}$$

Then it follows from this inequality and (3.22) that

$$\eta_2^2 \le C(\|u - u_h\|_{0,\Omega_U}^2 + \|y - y_h\|_{1,\Omega}^2) + C \sum_{\tau \in T^h} h_\tau^2 \int_\tau (F - \bar{F})^2. \tag{3.36}$$

To estimate η_3, we define the bubble function w_l as in Lemma 3.5 with $D_l = [(A\nabla y_h) \cdot \mathbf{n}]|_l$. By (3.13),

$$\begin{aligned}
\eta_3^2 &= \sum_{l \in \partial T^h} h_l \int_l D_l^2 = \sum_{l \in \partial T^h} \int_l w_l [(A\nabla y_h) \cdot \mathbf{n}] = \sum_{l \in \partial T^h} \int_{\tau_l^1 \cup \tau_l^2} (A\nabla y_h) \cdot \nabla w_l \\
&= \sum_{l \in \partial T^h} \int_{\tau_l^1 \cup \tau_l^2} (A\nabla(y_h - y_{u_h})) \cdot \nabla w_l + \sum_{l \in \partial T^h} \int_{\tau_l^1 \cup \tau_l^2} (f + Bu_h) w_l \\
&\le C \sum_{\tau \in T^h} |y_{u_h} - y_h|_{1,\tau}^2 + \delta \sum_{l \in \partial T^h} (|w_l|_{1,\tau_l^1 \cap \tau_l^2}^2 + h_l^{-2}\|w_l\|_{0,\tau_l^1 \cap \tau_l^2}^2) + C\eta_2^2 \\
&\le C(|y_{u_h} - y|_{1,\Omega}^2 + |y - y_h|_{1,\Omega}^2) + C\delta\eta_3^2 + C\eta_2^2.
\end{aligned}$$

It follows from the above inequality, (3.22), and (3.36) that

$$\eta_3^2 \le C\left(\|u - u_h\|_{0,\Omega_U}^2 + \|y - y_h\|_{1,\Omega}^2\right) + C \sum_{\tau \in T^h} h_\tau^2 \int_\tau (F - \bar{F})^2.$$

For η_4, let w_τ be set as in Lemma 3.5 with $B_\tau = \bar{G}|_\tau$. It follows from (3.14), (2.1), (3.23), and (3.22) that

$$\begin{aligned}
\eta_4^2 &= \sum_{\tau\in T^h} h_\tau^2 \int_\tau G^2 \le 2\sum_{\tau\in T^h} h_\tau^2 \int_\tau \{\bar{G}^2 + (G-\bar{G})^2\} \\
&= 2\sum_{\tau\in T^h} \int_\tau \{w_\tau G + w_\tau(\bar{G}-G) + h_\tau^2 (G-\bar{G})^2\} \\
&= 2\sum_{\tau\in T^h} \int_\tau \{(A\nabla w_\tau)\cdot\nabla(p_{u_h}-p_h) + w_\tau(g'(y_h)-g'(y_{u_h})) \\
&\qquad\qquad + w_\tau(\bar{G}-G) + h_\tau^2(G-\bar{G})^2\} \\
&\le C|p_{u_h}-p_h|_{1,\Omega}^2 + \delta\sum_{\tau\in T^h}(|w_\tau|_{1,\tau}^2 + h_\tau^{-2}\|w_\tau\|_{0,\tau}^2) \\
&\qquad\qquad + C\|y_h - y_{u_h}\|_{1,\Omega}^2 + C\sum_{\tau\in T^h} h_\tau^2 \int_\tau (G-\bar{G})^2 \\
&\le C\left(\|u-u_h\|_{0,\Omega_U}^2 + \|p-p_h\|_{1,\Omega}^2 + \|y-y_h\|_{1,\Omega}^2\right) + C\delta\eta_4^2 + C\sum_{\tau\in T^h} h_\tau^2 \int_\tau (G-\bar{G})^2.
\end{aligned}$$

Thus

$$\eta_4^2 \le C\left(\|u-u_h\|_{0,\Omega_U}^2 + \|y-y_h\|_{1,\Omega}^2 + \|p-p_h\|_{1,\Omega}^2\right) + C\sum_{\tau\in T^h} h_\tau^2 (G-\bar{G})^2. \tag{3.37}$$

To estimate η_5, we set w_l as in Lemma 3.5 with $D_l = [(A^*\nabla p_h)\cdot\mathbf{n}]|_l$. It follows from (3.14), (2.1), (3.23), and (3.22) that

$$\begin{aligned}
\eta_5^2 &= \sum_{l\in\partial T^h} h_l \int_l D_l^2 = \sum_{l\in\partial T^h}\int_l w_l[(A^*\nabla p_h)\cdot\mathbf{n}] = \sum_{l\in\partial T^h}\int_{\tau_l^1\cup\tau_l^2} (A^*\nabla p_h)\cdot\nabla w_l \\
&= \sum_{l\in\partial T^h}\int_{\tau_l^1\cup\tau_l^2} (A\nabla w_l)\cdot\nabla(p_h - p_{u_h}) + \sum_{l\in\partial T^h}\int_{\tau_l^1\cup\tau_l^2} g'(y_{u_h})w_l \\
&\le C|p_h - p_{u_h}|_{1,\Omega}^2 + \delta\sum_{l\in\partial T^h}(|w_l|_{1,\tau_l^1\cap\tau_l^2}^2 + h_l^{-2}\|w_l\|_{0,\tau_l^1\cap\tau_l^2}^2) \\
&\quad + C\|y_{u_h} - y_h\|_{1,\Omega}^2 + C\eta_4^2 \\
&\le C(\|u-u_h\|_{0,\Omega_U}^2 + \|p-p_h\|_{1,\Omega}^2 + \|y-y_h\|_{1,\Omega}^2) + C\delta\eta_5^2 + C\eta_4^2.
\end{aligned}$$

This inequality, combined with (3.37), implies

$$\eta_5^2 \le C(\|u-u_h\|_{0,\Omega_U}^2 + \|y-y_h\|_{1,\Omega}^2 + \|p-p_h\|_{1,\Omega}^2) + C\sum_{\tau\in T^h} h_\tau^2 (G-\bar{G})^2.$$

Thus we proved the desirable result. □

We believe that the error estimator $\eta_1^2 + \sum_{i=2}^5 \hat{\eta}_i^2$ in Theorem 3.2 is also sharp, though we are unable to establish any lower error bound for it. As a matter of fact, to our best knowledge, there exist no lower a posteriori error bounds in the L^2-norm in the literature or for any control problem.

3.3. Sharp a posteriori error estimators. In the above section, we have shown the following error bounds:

(3.38)
$$c\left(\sum_{i=1}^{5}\eta_i^2 \quad - \sum_{\tau\in T^h} h_\tau^2 \int_\tau \{(F-\bar F)^2+(G-\bar G)^2\} - \sum_{\tau\in\Omega_h^b} h_\tau^2 \|\nabla(h'(u_h)+B^*p_h)\|_{0,\tau}^2\right)$$
$$\le \|u-u_h\|_{0,\Omega_U}^2 + \|y-y_h\|_{1,\Omega}^2 + \|p-p_h\|_{1,\Omega}^2 \le C\left(\sum_{i=1}^{5}\eta_i^2\right),$$

provided that the conditions of Theorems 3.1 and 3.3 hold. We note that, if the free boundary $\partial\Omega_U^+$ is regular, for instance, if the free boundary consists of a finite number of smooth surfaces or if the total area of the free boundary is finite, then meas (Ω_h^b) is of the order h as $h\to 0$. Thus the second and third terms of the left side are of higher order as $h\to 0$ if the data are regular. Take the following typical quadratic control as one example: let $\Omega=\Omega_U$; let U^h and Y^h be the piecewise constant and linear spaces, respectively; let $Bu=u$, $f\in H^1(\Omega)$, $h(u)=\int_\Omega u^2$, and $g(y)=\int_\Omega (y-y_0)^2$ with $y_0\in H^1(\Omega)$. Then one has

$$\int_\Omega (F-\bar F)^2+(G-\bar G)^2 \le Ch^2(|f|_{H^1}^2+|y_0|_{H^1}^2)+C\int_\Omega (Bu_h-\overline{Bu_h})^2,$$

$$\sum_{\tau\in\Omega_h^b}\|\nabla(h'(u_h)+B^*p_h)\|_{0,\tau}^2 \le C\left(\int_{\Omega_h^b}|\nabla B^*p|^2+\|p_h-p\|_{H^1(\Omega)}^2\right).$$

Thus it follows that the second and third terms of the left side of (3.38) are not needed in computations. It can be seen that the above observation still holds even if f, y_0 are only piecewise smooth. For more general objective functionals, one can proceed as in Remark 3.4. Therefore, (3.38) gives equivalent a posteriori error estimates in the global sense and thus shows that the estimator $\sum_{i=1}^5\eta_i^2$ is in general quite sharp.

An obvious problem is that the characteristic function $\chi_{\Omega_h^{+b}}$ is not a posteriori in the sense that we usually do not know the position of the free boundary. Nevertheless, one can substitute it with some a posteriori quantities, thus obtaining some a posteriori error indicators, which can then be used in the adaptive finite element method.

One possible idea is to approximate $\chi_{\Omega_h^{+b}}$ by the finite element solution, as suggested in [24] and [32]. The basic idea is to approximate the characteristic function with the a posteriori quantity $\chi_{\Omega_U^+}^h$. For $\alpha>0$, let

$$\chi_{\Omega_U^+}^h = \frac{u_h-\phi_0}{h^\alpha+u_h-\phi_0}.$$

Thus, in computing, we replace η_1^2 by

$$\widetilde{\eta}_1^2 = \sum_{\tau\in T_U^h} h_{\tau_U}^2 \|\nabla(h'(u_h)+B^*p_h)\chi_{\Omega_U^+}^h\|_{0,\tau_U}^2.$$

In the following, we investigate the possible errors caused by this replacement. To this end, we separate Ω_U into three parts:

$$\Omega_h^-, \quad \Omega_U^{\alpha/2} := \{x\in\Omega_U \ : \ u_h(x)<\phi_0+h^{\alpha/2},\ u(x)>\phi_0\}, \text{ and } \Omega_h^{+b}\backslash\Omega_U^{\alpha/2}.$$

Then, for $\tau \in \Omega_h^-$, we have

$$\|\chi^h_{\Omega_U^+} - \chi_{\Omega_h^{+b}}\|_{0,\infty,\tau} = \left\| \frac{u_h - u}{h^\alpha + u_h - u} \right\|_{0,\infty,\tau} \le \min\{1, h^{-\alpha}\|u_h - u\|_{0,\infty,\tau}\},$$

and, for $\tau \in \Omega_h^{+b} \backslash \Omega_U^{\alpha/2}$,

$$\|\chi^h_{\Omega_U^+} - \chi_{\Omega_h^{+b}}\|_{0,\infty,\tau} = \left\| \frac{h^\alpha}{h^\alpha + u_h - \phi_0} \right\|_{0,\infty,\tau} \le h^{\alpha/2}.$$

Therefore, if the error $\|u_h - u\|_{0,\infty,\tau}$ for $\tau \subset \Omega_h^-$ (where $u \equiv \phi_0$) is of the order h^β with $\beta > \alpha$, then the difference caused by the replacement is a high-order small quantity locally for all $\tau \in \Omega_h^- \cup (\Omega_h^{+b} \backslash \Omega_U^{\alpha/2})$. The size of the remaining domain $\Omega_U^{\alpha/2}$ depends on the error $\|u_h - u\|_{0,\infty}$. It can be shown (see, e.g., [32]) that meas $(\Omega_U^{\alpha/2}) \to 0$, as long as $\|u_h - u\|_{0,\infty,\Omega} \to 0$, as $h \to 0$. Thus $\chi^h_{\Omega_U^+}$ is a good approximator to $\chi_{\Omega_h^{+b}}$, and this is confirmed in our numerical tests; see section 4.

Remark 3.1. Generally speaking, for the problem considered here, the costate p is more regular than the solution u. Therefore, we may use p_h instead of u_h to approximate the characteristic function. It can be seen from (2.3) that

$$u = \max\{-(h')^{-1}(B^*p), \phi_0\}.$$

Thus, for example, we can define

$$\widetilde{\chi}^h_{\Omega_U^+} = \frac{\widetilde{u}_h - \phi_0}{h^\alpha + \widetilde{u}_h - \phi_0},$$

where $\widetilde{u}_h = \max\{-(h')^{-1}(B^*p_h), \phi_0\}$. Similarly, we can show

$$\|\widetilde{\chi}^h_{\Omega_U^+} - \chi_{\Omega_h^{+b}}\|_{0,\infty,\tau} \le \begin{cases} \min\{1, h^{-\alpha}\|(h')^{-1}(B^*p_h) - (h')^{-1}(B^*p)\|_{0,\infty,\tau}\} & \forall \tau \in \Omega_h^-, \\ h^{\alpha/2} \qquad \forall \tau \in \Omega_h^{+b} \backslash \Omega_U^{\alpha/2}. \end{cases}$$

3.4. Nonconstant obstacles. If the constraint ϕ_0 is a function $\phi(x)$, one could introduce $u^*(x) = u(x) - \phi(x)$. Then the triplet (y, p, u^*) satisfies the following optimality conditions:

$$\text{(3.39)} \qquad \begin{cases} a(y, w) = (f^* + Bu^*, w) \quad \forall w \in Y = H_0^1(\Omega), \\ a(q, p) = (g'(y), p) \quad \forall q \in Y = H_0^1(\Omega), \\ ((h^*)'(u^*) + B^*p, v - u)_U \ge 0 \quad \forall v \in K \subset U = L^2(\Omega_U), \end{cases}$$

where $f^* = f + B\phi, h^*(v) = h(v + \phi)$, and $K = \{v \in U : v \ge 0\}$. Thus the problem is reduced to the case of (2.3) with $\phi_0 = 0$.

However, this strategy, although simpler, may affect the efficiency of the resulting error estimators. Let us try to explain this: the inactive data $\phi|_{\Omega_U^-}$ on the noncoincidence set does not affect the solution of (CCP) and thus is not expected to play a major role in a sharp error estimator. However, with the transformation $u - \phi$, this data may be brought into the resulting error estimators through f^*. Thus we will directly consider the error $u - u_h$ rather than $u^* - u_h^*$. Let

$$\text{(3.40)} \quad K = \{v \in U : v \ge \phi \text{ a.e. in } \Omega_U\}, \qquad K^h = \{v_h \in U^h : v_h \ge \phi^h \text{ a.e. in } \Omega_U\},$$

where $\phi^h \in U^h$ is an approximation of ϕ. Here we take $\phi^h = \pi_h^a \phi$. It should be noticed that $K^h \not\subset K$ in general.

THEOREM 3.4. *Let (y, p, u) and (y_h, p_h, u_h) be the solutions of (2.3) and (2.5), respectively. Assume that all of the conditions of Theorem 3.1 and (3.35) hold and K^h is defined as in (3.40) with $\phi \in L^2(\Omega_U)$ and $\phi^h = \pi_h^a \phi$. Then*

$$\|u_h - u\|_{0,\Omega_U}^2 + \|y_h - y\|_{1,\Omega}^2 + \|p_h - p\|_{1,\Omega}^2 \le C \sum_{i=1}^{6} \eta_i^2, \tag{3.41}$$

where η_i $(i = 1\text{–}5)$ are defined in Theorem 3.1 and

$$\eta_6^2 = \sum_{\tau_U \in T_U^h} \|(\phi^h - \phi)\chi_{\Omega_h^{-b}}\|_{0,\tau_U}^2.$$

Proof. We will give only the details for the estimation of $\|u - u_h\|_{0,\Omega_U}^2$. The other terms can be estimated similarly as in Theorem 3.1. It should be emphasized that here one cannot take $v = u_h$ in (2.3) since $u_h \ge \phi$ may not be true. It follows from (3.3) that

$$h'(u) + B^* p \ge 0, \quad (h'(u) + B^* p)\chi_{\Omega_U^+} = 0. \tag{3.42}$$

Then it follows from the assumption (3.4), the inequality (2.5), and (3.42) that, for any $v_h \in K^h$,

$$\begin{aligned}
&c\|u - u_h\|_{0,\Omega_U}^2 \\
&\le (h'(u), u - u_h)_U - (h'(u_h), u - u_h)_U + (h'(u_h) + B^* p_h, v_h - u_h)_U \\
&= (h'(u_h) + B^* p_h, v_h - u)_U + (B^*(p_h - p), u - u_h)_U + (h'(u) + B^* p, u - u_h)_U \\
&= ((h'(u_h) + B^* p_h)\chi_{\Omega_h^{+b}}, v_h - u)_U + (B^*(p_h - p), u - u_h)_U \\
&\quad + ((h'(u_h) + B^* p_h - (h'(u) + B^* p))\chi_{\Omega_h^-}, v_h - u)_U \\
&\quad + ((h'(u) + B^* p)\chi_{\Omega_h^-}, v_h - u_h)_U + ((h'(u) + B^* p)\chi_{\Omega_U^- \setminus \Omega_h^-}, u - u_h)_U \\
&\quad := \sum_{i=1}^{5} I_i.
\end{aligned} \tag{3.43}$$

Take $v_h = \pi_h^a u$. Then I_1 and I_2 can be estimated as in the proof of Theorem 3.1 such that

$$I_1 + I_2 \le C(\eta_1^2 + \|p_h - p_{u_h}\|_{1,\Omega}^2) + \delta \|u - u_h\|_{0,\Omega_U}^2,$$

where δ is a small positive constant. It follows from (3.35), (3.24), and (3.23) that

$$\begin{aligned}
I_3 &\le \sum_{\tau_U \in T_U^h} (\|h'(u_h) - h'(u)\|_{0,\tau_U} + \|B^*(p_h - p)\|_{0,\tau_U}) \|(\phi - \phi^h)\chi_{\Omega_h^-}\|_{0,\tau_U} \\
&\le \delta(\|u - u_h\|_{0,\Omega_U}^2 + \|p_h - p_{u_h}\|_{1,\Omega}^2) + C\eta_6^2.
\end{aligned}$$

We note that $I_4 \le 0$ due to (3.42) and the fact that $(v_h - u_h)|_{\Omega_h^-} = (\phi^h - u_h)|_{\Omega_h^-} \le 0$.

Finally, to estimate I_5, we use $u|_{\Omega_U^-} = \phi|_{\Omega_U^-}$, (3.42) and $u_h \geq \phi^h$ to get

$$\begin{aligned}
&((h'(u)+B^*p)\chi_{\Omega_U^-\setminus\Omega_h^-}, u-u_h)_{\tau_U} = ((h'(u)+B^*p)\chi_{\Omega_U^-\setminus\Omega_h^-}, \phi-u_h)_{\tau_U}\\
&\leq ((h'(u)+B^*p)\chi_{\Omega_U^-\setminus\Omega_h^-}, \phi-\phi^h)_{\tau_U} = (h'(u)+B^*p, (\phi-\phi^h)\chi_{\Omega_h^b})_{\tau_U}\\
&= (h'(u)+B^*p-(h'(u_h)+B^*p_h), (\phi-\phi^h)\chi_{\Omega_h^b})_{\tau_U}\\
&\quad + ((I-\pi_h^a)(h'(u_h)+B^*p_h), (\phi-\phi^h)\chi_{\Omega_h^b})_{\tau_U}.
\end{aligned}$$

Thus

$$I_5 \leq \delta(\|u-u_h\|_{0,\Omega_U}^2 + \|p_h - p_{u_h}\|_{1,\Omega}^2) + C(\eta_1^2+\eta_6^2).$$

The rest of the proof is the same as that in Theorem 3.1. □

Remark 3.2. We can approximate the characteristic functions $\chi_{\Omega_h^{+b}}$ and $\chi_{\Omega_h^{-b}}$ by

$$\chi^h_{\Omega_h^{+b}} = \frac{u_h-\phi^h}{h^{\alpha^+}+u_h-\phi^h}, \qquad \chi^h_{\Omega_h^{-b}} = \frac{h^{\alpha^-}}{h^{\alpha^-}+u_h-\phi^h},$$

where α^+ and α^- are positive parameters.

3.5. Double obstacles. We now consider the control problem with the double obstacles: $\phi_1(x) < \phi_2(x)$. Let

$$\begin{aligned}
&K = \{v \in U : \phi_1 \leq v \leq \phi_2 \text{ a.e. in } \Omega_U\},\\
&K^h = \{v_h \in U^h : \phi_1^h \leq v_h \leq \phi_2^h \text{ a.e. in } \Omega_U\},
\end{aligned} \tag{3.44}$$

where $\phi_i^h \in U^h$ is an approximation of ϕ_i $(i=1,2)$. We assume that $\phi_i^h = \pi_h^a\phi_i$ $(i=1,2)$. To generalize the ideas used in Theorem 3.4 to this case, we define

$$\begin{aligned}
&\Omega^-_{\phi_i} = \{x \in \Omega_U : u(x) = \phi_i(x),\}, && \Omega^-_{\phi} = \Omega^-_{\phi_1} \cup \Omega^-_{\phi_2}, && \Omega^+_{\phi} = \Omega_U \setminus \Omega^-_{\phi},\\
&\Omega^-_{\phi_i,h} = \{\cup\bar\tau_U : \tau_U \subset \Omega^-_{\phi_i}, \tau_U \in T_U^h\}, && \Omega^-_{\phi,h} = \Omega^-_{\phi_1,h} \cup \Omega^-_{\phi_2,h}, && \Omega^{+b}_{\phi,h} = \Omega_U \setminus \Omega^-_{\phi,h},\\
&\Omega^{-b}_{\phi_i,h} = \{\cup\bar\tau_U : \bar\tau_U \cap \Omega^-_{\phi_i,h} \neq \emptyset, \tau_U \in T_U^h\}.
\end{aligned}$$

THEOREM 3.5. *Let (y,p,u) and (y_h,p_h,u_h) be the solutions of* (2.3) *and* (2.5)*, respectively. Assume that all of the conditions of Theorem* 3.4 *hold and K and K^h are defined as in* (3.44) *with $\phi_i \in L^2(\Omega_U)$ and $\phi_i^h = \pi_h^a\phi_i$ $(i=1,2)$. Then*

$$\|u_h-u\|_{0,\Omega_U}^2 + \|y_h-y\|_{1,\Omega}^2 + \|p_h-p\|_{1,\Omega}^2 \leq C\sum_{i=1}^6 \eta_i^2, \tag{3.45}$$

where η_i $(i=2$–$5)$ are defined in Theorem 3.1 *and*

$$\eta_1^2 = \sum_{\tau_U \in T_U^h} h_{\tau_U}^2 \|\nabla(h'(u_h)+B^*p_h)\chi_{\Omega^{+b}_{\phi,h}}\|_{0,\tau_U}^2,$$

$$\eta_6^2 = \sum_{\tau_U \in T_U^h} \sum_{i=1,2} \|(\phi_i^h-\phi_i)\chi_{\Omega^{-b}_{\phi_i,h}}\|_{0,\tau_U}^2.$$

Proof. Again, we give only the details for estimation of the error $\|u-u_h\|^2_{0,\Omega_U}$. In this case, we have

$$(h'(u)+B^*p)\chi_{\Omega^-_{\phi_1}} \ge 0, \quad (h'(u)+B^*p)\chi_{\Omega^-_{\phi_2}} \le 0, \quad (h'(u)+B^*p)\chi_{\Omega^+_{\phi}} = 0. \tag{3.46}$$

As in (3.43), it follows from the assumption (3.4), the inequality (2.5), and (3.46) that, for any $v_h \in K^h$,

$$\begin{aligned}
&c\|u-u_h\|^2_{0,\Omega_U} \\
&\le (h'(u),u-u_h)_U - (h'(u_h),u-u_h)_U + (h'(u_h)+B^*p_h, v_h-u_h)_U \\
&= (h'(u_h)+B^*p_h, v_h-u)_U + (B^*(p_h-p),u-u_h)_U + (h'(u)+B^*p,u-u_h)_U \\
&= ((h'(u_h)+B^*p_h)\chi_{\Omega^{+b}_{\phi,h}}, v_h-u)_U + (B^*(p_h-p),u-u_h)_U \\
&\quad +((h'(u_h)+B^*p_h-(h'(u)+B^*p))\chi_{\Omega^-_{\phi,h}}, v_h-u)_U \\
&\quad +((h'(u)+B^*p)\chi_{\Omega^-_{\phi,h}}, v_h-u_h)_U + ((h'(u)+B^*p)\chi_{\Omega^-_{\phi}\setminus\Omega^-_{\phi,h}}, u-u_h)_U \\
&:= \sum_{i=1}^5 J_i.
\end{aligned} \tag{3.47}$$

It is easy to see that for $1 \le i \le 3$, J_i can be estimated as I_i. Thanks to (3.46), we still have $J_4 \le 0$. Also, J_5 can be treated similarly to I_5. For instance, let us consider the case that $\tau_U \subset (\Omega^{-b}_{\phi_2,h}\setminus\Omega^-_{\phi_2,h})$. Assume that $\Omega^{-b}_{\phi_2,h} \cap \Omega^-_{\phi_1} = \emptyset$ for simplicity. We then have, from $u|_{\Omega^-_{\phi_2}} = \phi_2|_{\Omega^-_{\phi_2}}$, (3.46), and $u_h \le \phi^h_2$, that

$$\begin{aligned}
&((h'(u)+B^*p)\chi_{\Omega^{-b}_{\phi_2,h}\setminus\Omega^-_{\phi_2,h}}, u-u_h)_{\tau_U} \\
&\quad = ((h'(u)+B^*p)\chi_{\Omega^{-b}_{\phi_2,h}\setminus\Omega^-_{\phi_2,h}}, \phi_2-u_h)_{\tau_U} \\
&\qquad \le ((h'(u)+B^*p)\chi_{\Omega^{-b}_{\phi_2,h}\setminus\Omega^-_{\phi_2,h}}, \phi_2-\phi^h_2)_{\tau_U} \\
&\quad = (h'(u)+B^*p-(h'(u_h)+B^*p_h), \phi_2-\phi^h_2)_{\tau_U} \\
&\qquad + ((I-\pi^a_h)(h'(u_h)+B^*p_h), \phi_2-\phi^h_2)_{\tau_U}.
\end{aligned}$$

The rest of the proof is the same as that of Theorem 3.4. ∎

Remark 3.3. In computing, we may approximate the characteristic functions $\chi_{\Omega^{+b}_h}$ and $\chi_{\Omega^{-b}_{\phi_i,h}}$ by

$$\chi^h_{\Omega^{+b}_h} = \frac{(u_h-\phi^h_1)(\phi^h_2-u_h)}{h^{\alpha^+}+(u_h-\phi^h_1)(\phi^h_2-u_h)}, \qquad \chi^h_{\Omega^{-b}_{\phi_i,h}} = \frac{h^{\alpha^-}}{h^{\alpha^-}+|u_h-\phi^h_i|},$$

where α^+ and α^- are positive parameters.

Remark 3.4. It is clear that the uniform monotonicity conditions and Lipschitz continuity (2.1), (3.4), (3.25), (3.35), assumed in the proofs of Theorems 3.1–3.3, are needed to hold only in a neighborhood of the true solutions. This observation is useful in some applications involving a nonquadratic objective functional like $g(y) = \int_\Omega (y-y_0)^4$.

For (2.1) and (3.25), let us assume that $g(y) = \int_\Omega j(y)$, where j is twice continuously differentiable on R^1, to just fix the idea. Then it follows from the Sobolev embedding result $H^1(\Omega) \to L^\beta(\Omega)$ ($\beta < \infty$ if $n = 2$, and $\beta = 6$ if $n = 3$) that we have, using the Hölder inequality,

$$\begin{aligned} |(g'(v) - g'(w), q)| &\le \|j''(z)\|_{0,\beta^*,\Omega}\|v - w\|_{0,\beta,\Omega}\|q\|_{0,\beta,\Omega} \\ &\le C\|j''(z)\|_{0,\beta^*,\Omega}\|v - w\|_{1,\Omega}\|q\|_{1,\Omega}, \end{aligned}$$

where $z = \theta v + (1 - \theta)w$ with $\theta \in [0, 1]$, $\beta^* = (1 - 2/\beta)^{-1}$ for any $\beta > 2$ if $n = 2$ and $\beta^* = 3/2$ if $n = 3$.

Also, by using the embedding result $H^2(\Omega) \to L^\infty(\Omega)$ for $n \le 3$, we have

$$|(g'(v) - g'(w), q)| \le \|j''(z)\|_{0,\Omega}\|v - w\|_{0,\Omega}\|q\|_{2,\Omega}.$$

For example, if $g(y) = \int_\Omega (y - y_0)^4$ with $y_0 \in L^4(\Omega)$, we have by $H^1(\Omega) \to L^{2\beta^*}(\Omega)$ that

$$\begin{aligned} |(g'(v) - g'(w), q)| &\le C(\|v^2\|_{0,\beta^*,\Omega} + \|w^2\|_{0,\beta^*,\Omega} + \|y_0^2\|_{0,\beta^*,\Omega})\|v - w\|_{1,\Omega}\|q\|_{1,\Omega}, \\ &\le C(\|v\|_{1,\Omega}^2 + \|w\|_{1,\Omega}^2 + \|y_0\|_{L^4(\Omega)}^2)\|v - w\|_{1,\Omega}\|q\|_{1,\Omega} \end{aligned}$$

and

$$\begin{aligned} |(g'(v) - g'(w), q)| &\le C(\|v^2\|_{0,\Omega} + \|w^2\|_{0,\Omega} + \|y_0^2\|_{0,\Omega})\|v - w\|_{0,\Omega}\|q\|_{2,\Omega}, \\ &\le C(\|v\|_{1,\Omega}^2 + \|w\|_{1,\Omega}^2 + \|y_0\|_{L^4(\Omega)}^2)\|v - w\|_{0,\Omega}\|q\|_{2,\Omega}. \end{aligned}$$

Thus (2.1) and (3.25) hold as long as v, w are in a bounded set of Y. One can discuss (3.4) and (3.35) similarly.

It follows from the proofs of Theorems 3.4–3.5 that Theorem 3.2 can also be generalized to the nonconstant or double obstacle cases in the same way. Thus one can just use $\eta_1^2 + \eta_6^2$ as the error indicator in adaptive finite element methods if only the values of the control and state are important in an application.

4. Numerical experiments. In this section, we carry out some numerical experiments to demonstrate possible applications of the error estimators obtained in section 3. In most control problems, the optimal control is often of prime interest. Thus it is important to develop mesh refinement schemes that efficiently reduce the error $\|u - u_h\|$. In practice, there are four major types of adaptive finite element methods—namely, the h-method (mesh refinement), the p-method (order enrichment), the r-method (mesh redistribution), and the hp-method. A posteriori error estimators can be used as error indicators to guide the mesh refinement in adaptive finite element methods. For our numerical tests, using an adaptive mesh redistribution (AMR) method is advantageous since it can keep the number of the total nodes unchanged while adjusting the distribution of the nodes.

4.1. AMR method. The general idea behind the AMR method is to adjust meshes such that the a posteriori error estimators (the monitor functions to be called) are *equally* distributed over the computational meshes, while the total number of the nodes remains the same. Clearly, this method particularly suits our purposes of testing the efficiency of the known a posteriori error estimators.

In solving the optimal control problem (1.1), we use an iterative method to move the meshes and to redistribute the solutions on the new grid points. The procedure

for the mesh moving part is described in [24, 25, 26]. The key idea here is to use some kind of equivalent error estimators as the monitor function (or moving mesh indicator). More precisely, let $(x(\xi,\eta), y(\xi,\eta))$ be the mesh map in two dimensions. Here (ξ,η) are the computational coordinates. Let $M > 0$ be the monitor function which depends on the physical solution to be adapted. By solving the Euler–Lagrange equation

$$\nabla \cdot (M^{-1}\nabla\xi) = 0, \qquad \nabla \cdot (M^{-1}\nabla\eta) = 0, \tag{4.1}$$

a map between the physical domain Ω and the logical domain Ω_c can be computed. Typically, the map transforms a uniform mesh in the logical domain to cluster grid points at the regions of physical domain where the solutions are of greater physical interest. One of the crucial issues is what monitor functions are to be used. One popular choice in the AMR method literature is a gradient-based monitor function like $M_\tau = \sqrt{1 + |\nabla y_h|^2_\tau}$, which moves more grids to the regions of the largest solution gradients. In [24], it was shown that the gradient-based monitor functions may not be suitable for free boundary problems, and a monitor function associated with a posteriori error estimators is introduced which was found particularly useful in approximating the variational inequalities with free boundaries. In this section, we will use the same solution procedures as described in [24] to obtain the numerical solutions with moving grids, except that monitor functions will be based on the error estimators developed in this work.

4.2. Numerical tests. Our numerical example is the following type of optimal control problem:

$$\text{(OCP)} \qquad \begin{aligned} &\min \frac{1}{2}\int_\Omega (y - y_0)^2 + \frac{1}{2}\int_{\Omega_U} (u - u_0)^2 \\ &\text{s.t.} \quad \begin{cases} -\Delta y = Bu + f, \\ y|_{\partial\Omega} = y_0|_{\partial\Omega} = 0, \\ u \geq 0 \quad \text{in } \Omega_U. \end{cases} \end{aligned}$$

In our example, $\Omega_U = \Omega = [0,1] \times [0,1]$ and $B = I$. We also use the same meshes for the approximation of the state and the control. Thus $\tau_U = \tau$. Let Ω^h be a polygonal approximation to Ω with boundary $\partial\Omega^h$. Let T^h be a partitioning of Ω^h into a disjoint regular triangular τ so that $\bar{\Omega}^h = \cup_{\tau\in T^h}\bar{\tau}$. Assume that the state y is approximated in the finite element space Y^h with Φ^i as basis functions and u is approximated in U^h with Ψ^i as basis functions. Thus the problem (OCP) is discretized as the following optimization problem:

$$\begin{aligned} \min \quad & \tfrac{1}{2}\left\{(Y - Y^0)^T Q (Y - Y^0) + (U - U^0)^T M (U - U^0)\right\} \\ \text{s.t.} \quad & AY = BU + F, \\ & U \geq 0, \end{aligned} \tag{4.2}$$

with

$$Q^{ij} = \int_\Omega \Phi^i \Phi^j dx, \qquad M^{ij} = \int_\Omega \Psi^i \Psi^j dx,$$
$$A^{ij} = \int_\Omega \nabla\Phi^i \nabla\Phi^j dx, \qquad B^{ij} = \int_\Omega \Phi^i \Psi^j dx,$$
$$F_i = \int_\Omega f\Phi^i dx.$$

The finite element solution (y_h, u_h) is given by $y_h = \sum_i Y_i \Phi^i$ and $u_h = \sum_i U_i \Psi^i$, and (y_0, u_0) is approximated by $y_0^h = \sum_i Y_i^0 \Phi^i$ and $u_0^h = \sum_i U_i^0 \Psi^i$.

In solving the above optimization problem, we use a projection gradient method developed by He [19]. The projection method, though simple, is by no means the most efficient algorithm for solving our problem, but the purpose of the experiments in this section is to test the efficiency of the error indicators. The idea in [19] is the first to introduce the Lagrange multiplier P and then to set

$$H = \begin{pmatrix} Q & 0 & -A^T \\ 0 & M & B^T \\ A & -B & 0 \end{pmatrix}, \qquad x = \begin{pmatrix} Y \\ U \\ P \end{pmatrix}, \qquad c = \begin{pmatrix} QY^0 \\ MU^0 \\ F \end{pmatrix}.$$

The algorithm for solving the optimization problem (4.2) is described by the following pseudocode:

```
du = beta*(Hx + c)
e = x - max(x-du,b)
error =||e||
do  while error >= TOL
    d= beta*H^T*e
    g = d + du
    beta=beta*error/||d||
    e = e + d
    rho = error^2/||e||^2
    x = max(x - gamma \rho g,b)
    du = \beta (H x + c)
    e = x - max(x-du,b)
    error = ||e||
end do
```

We now briefly describe the solution algorithm to be used for solving the numerical examples in this section.

ALGORITHM 0

(i) Solve the optimization problem (4.2) with the above optimization code on the current mesh, and calculate the error monitor function M;
(ii) move the mesh to a new location, and update the solution on new meshes using the monitor M, as described in [25].

It is important to note from Theorem 3.2 that the error $\|u - u_h\|_{L^2(\Omega_U)}$ is largely controlled by η_1. Thus, in Algorithm 0, η_1, in (3.11) will be used to construct the

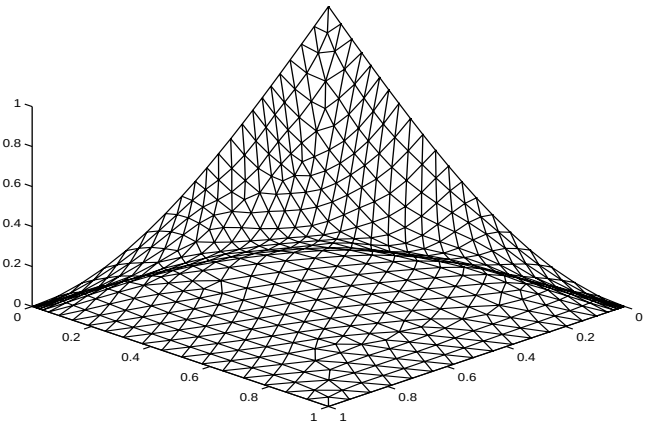


FIG. 4.1. *The surface of the solution u.*

monitor function M discussed in section 4.1,

$$M|_\tau = \sqrt{1 + \lambda \widetilde{\eta}_1^2|_\tau}, \tag{4.3}$$

where $\lambda > 0$ is a positive constant, and

$$\widetilde{\eta}_1^2|_\tau = h_\tau^2 \|\nabla(h'(u_h) + B^* p_h)\chi_{u_h}\|_{0,\tau}^2 . \tag{4.4}$$

In general, λ should be chosen such that $\lambda\|\eta_1\| >> 1$. Here we let $\lambda\|\eta_1\|_2 = 10^4$. As discussed in section 3, in our computation, we approximate the characteristic function used in η_1 by the following approximation:

$$\chi_{u_h} = \frac{u_h}{u_h + \epsilon}, \tag{4.5}$$

where $\epsilon > 0$ is a (small) positive number. In our experiments, we tried a range of values for ϵ between 0.1 and 1, and similar computational results were obtained.

Example 4.1. In this example we have

$$u_0 = 1 - \sin(\pi x_1/2) - \sin(\pi x_2/2), \quad y_0 = 0, \quad p = Z(x_1, x_2), \quad f = 4\pi^4 Z - u,$$

where $Z = \sin \pi x_1 \sin \pi x_2$. The exact solution of this problem is $y = 2\pi^2 Z, u = \max(u_0 - p, 0)$.

20×20 *nodes solution.* We first compute Example 4.1 on a 20×20 uniform mesh and then adjust the mesh by using Algorithm 0. The parameters λ and ϵ in (4.3) and (4.5) are 10^5 and 0.1, respectively. In Figure 4.1, the exact solution u is plotted. It is seen that the free boundary for this problem is just a single curve, and the maximum magnitude of the solution u is 1. The state and costate are approximated by piecewise linear elements. Both piecewise constant and piecewise linear elements are used to approximate the control in this example. In Figure 4.2, the 20×20 adaptive meshes are displayed. The control approximation errors are presented in Figures 4.3 and 4.4. It is observed that the maximum errors are distributed along the free boundary, as seen from Figures 4.3 and 4.4.

The adaptive meshes shown in Figure 4.2 are obtained by using the AMR method with the monitor function defined by (4.3). It is seen that a higher density of node

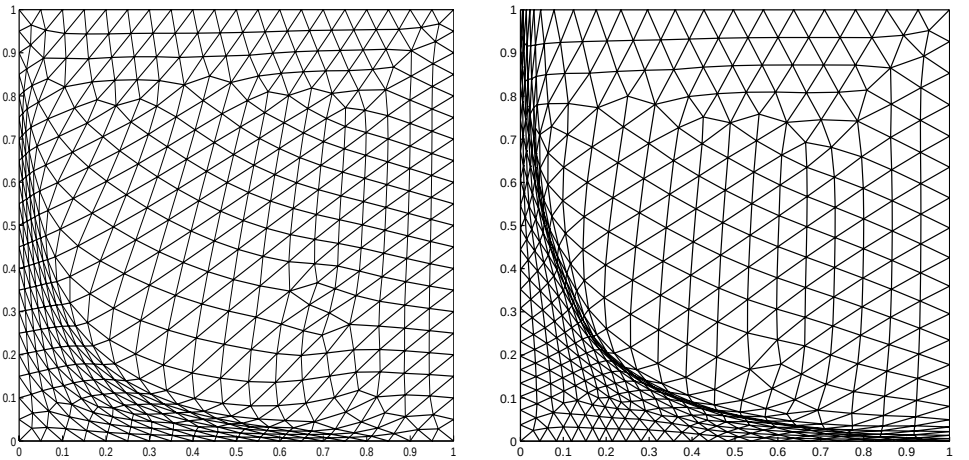


FIG. 4.2. *The adaptive mesh obtained by using piecewise constant elements (left) and piecewise linear elements (right), with* 20×20 *nodes.*

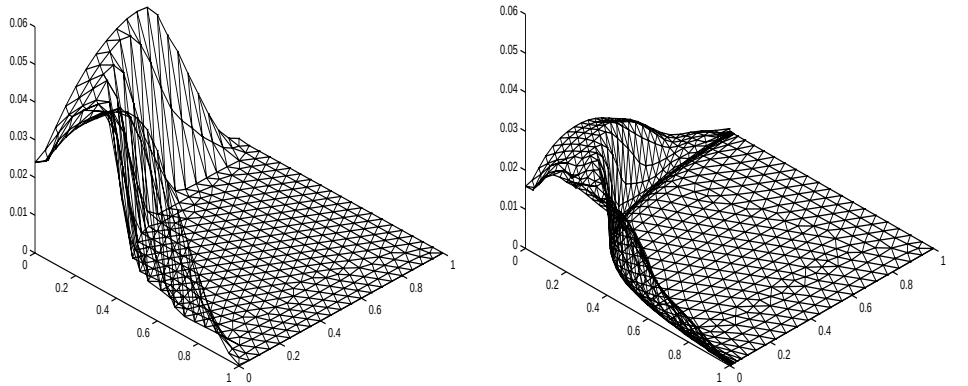


FIG. 4.3. L^2*-error* $\|u - u_h\|$ *with uniform mesh (left) and adaptive mesh (right), obtained by using piecewise constant elements with* 20×20 *nodes.*

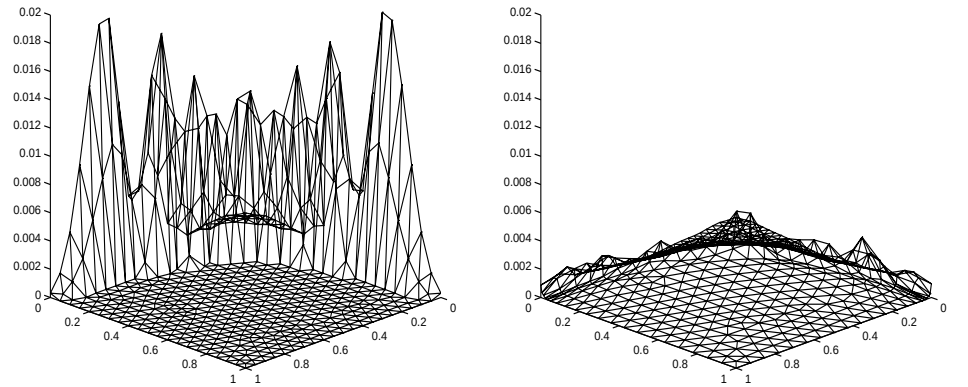


FIG. 4.4. *Same as Figure* 4.3*, except with linear elements.*

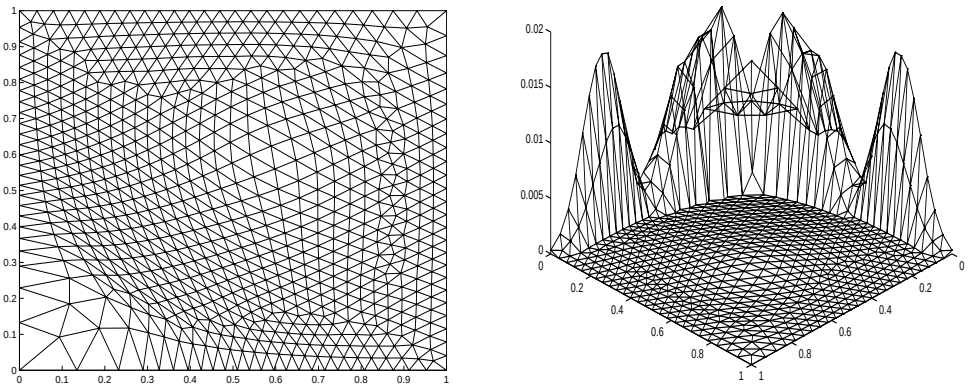


FIG. 4.5. *Example* 4.1 *with* 20×20 *nodes: Mesh (left) and error (right) obtained by using linear elements with unsharp error estimator associated with* $\bar{\eta}_1$*, as defined by* (3.2).

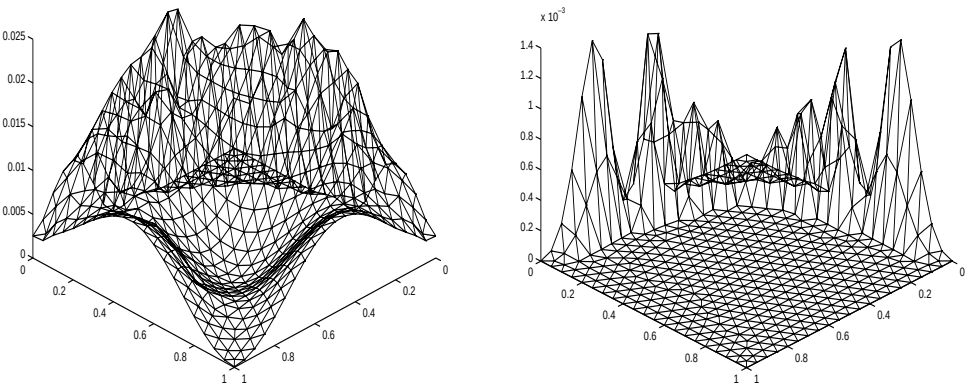


FIG. 4.6. *Profiles of unsharp estimator* $\bar{\eta}_1$ *and the sharp estimator* $\widetilde{\eta}_1$ *obtained by using the linear element with* 20×20 *nodes.*

points are now distributed along the free boundary. Furthermore, the approximation error is substantially reduced, as seen in Figures 4.3 and 4.4. In Figure 4.4, the L^2-norm of $u - u_h$ is 4.3×10^{-3} on the uniform mesh but is reduced 10 times to 4.4×10^{-4} on the adaptive mesh, while the L^2 error of the state approximation becomes slightly larger. It was found that one would need a 100×100 uniform mesh to produce such an error reduction. Thus efficient adaptive meshes can indeed save substantial computational work.

However, if we replace the estimator $\widetilde{\eta}_1$ in the monitor (4.3) with the estimator $\bar{\eta}_1$ given by (3.2), then a very different mesh is obtained; see Figure 4.5. As also seen in Figure 4.5, such a mesh is not efficient in reducing the control error; the error is virtually the same as that on the uniform mesh. The main reason is that the estimator $\bar{\eta}_1$ may not be sharp in this case. In fact, from Figure 4.6, it is clear that $\bar{\eta}_1$ and $|u - u_h|$ have very different profiles, while $\widetilde{\eta}_1$ has a profile similar to that of $|u - u_h|$.

40×40 *nodes solution.* To see the effect of mesh refinement, numerical solutions for Example 4.1 are obtained by using 40×40 linear elements. The control error distributions in this case are plotted in Figure 4.7, while the adapted mesh is plotted

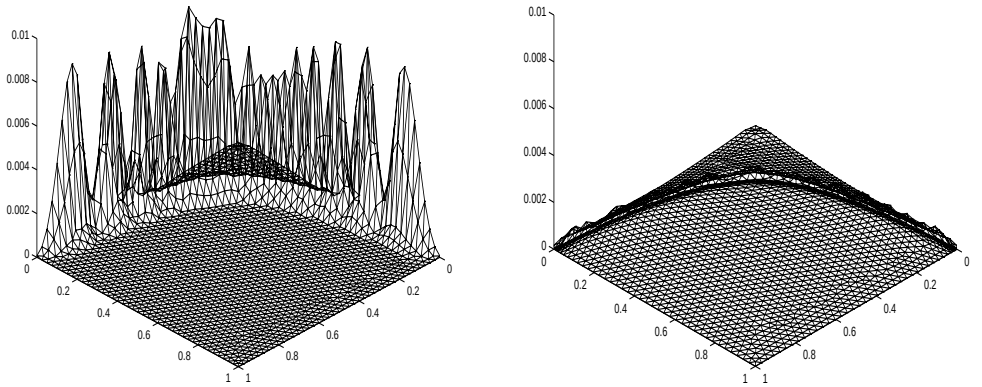


FIG. 4.7. *Same as Figure* 4.4*, except with* 40×40 *nodes.*

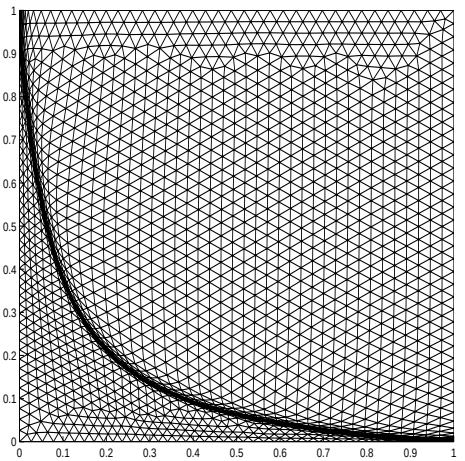


FIG. 4.8. *The adaptive mesh obtained by using piecewise linear elements with* 40×40 *nodes.*

in Figure 4.8. It is clear that the control errors are reduced with the finer mesh and with the use of the adaptive meshes.

5. Conclusion. In this work, we have derived some sharp a posteriori error indicators for the distributed elliptic optimal control problems. It is shown that the error indicators obtained can be applied in adaptive finite element computations and are found efficient in guiding mesh adjustments for our numerical examples. It is clear from the numerical experiments that the AMR methods can substantially increase the approximation accuracy. We point out that the approaches used in this work can be generalized to study other control problems.

REFERENCES

[1] M. AINSWORTH AND J. T. ODEN, *A posteriori error estimation in finite element analysis*, Comput. Methods Appl. Mech. Engrg., 142 (1997), pp. 1–88.

[2] W. ALT AND U. MACKENROTH, *Convergence of finite element approximations to state constrained convex parabolic boundary control problems*, SIAM J. Control Optim., 27 (1989), pp. 718–736.

[3] P. Alotto, P. Girdinio, O. Horigami, S. Ito, K. Iwanaga, K. Kato, T. Kuriyama, and H. Maeda, *Mesh adaption and optimization techniques in magnet design*, IEEE Trans. Magnetics, 32 (1996), pp. 1–8.

[4] N. V. Banichuk, F. J. Barthold, A. Falk, and E. Stein, *Mesh refinement for shape optimization*, Structural Optim., 9 (1995), pp. 46–51.

[5] R. E. Bank and A. Weiser, *Some a posteriori error estimators for elliptic partial differential equations*, Math. Comp., 44 (1985), pp. 283–301.

[6] J. Baranger and H. E. Amri, *A posteriori error estimators in finite element approximation of quasi-Newtonian flows*, M2AN Math. Model. Numer. Anal., 25 (1991), pp. 31–47.

[7] J. W. Barrett and W. B. Liu, *Finite element approximation of some degenerate quasilinear elliptic and parabolic problems*, in Numerical Analysis (Dundee, 1993), Pitman Res. Notes Math. Ser. 303, Longman, Harlow, UK, 1994, pp. 1–16.

[8] R. Becker and H. Kapp, *Optimization in PDE models with adaptive finite element discretization*, in Proceedings of ENUMATH-97, World Scientific, Singapore, 1998, pp. 147–155.

[9] R. Becker, H. Kapp, and R. Rannacher, *Adaptive finite element methods for optimal control of partial differential equations: Basic concept*, SIAM J. Control Optim., 39 (2000), pp. 113–132.

[10] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[11] K. Deckelnick and M. Hinze, *Error estimates in space and time for tracking-type control of the instationary Stokes system*, in Proceedings of Control and Estimation of Distributed Parameter Systems, Graz, Austria, 2001, to appear.

[12] F. S. Falk, *Approximation of a class of optimal control problems with order of convergence estimates*, J. Math. Anal. Appl., 44 (1973), pp. 28–47.

[13] D. A. French and J. T. King, *Approximation of an elliptic control problem by the finite element method*, Numer. Funct. Anal. Optim., 12 (1991), pp. 299–314.

[14] M. Hinze, *Optimal and Instantaneous Control of the Instationary Navier-Stokes Equations*, Habilitation thesis, Technische Universitat Berlin, Berlin, Germany, 2000.

[15] T. Geveci, *On the approximation of the solution of an optimal control problem governed by an elliptic equation*, RAIRO Anal. Numér., 13 (1979), pp. 313–328.

[16] V. Girault and P. A. Raviart, *Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms*, Springer-Verlag, Berlin, 1986.

[17] M. D. Gunzburger and L. S. Hou, *Finite-dimensional approximation of a class of constrained nonlinear control problems*, SIAM J. Control Optim., 34 (1996), pp.1001–1043.

[18] J. Haslinger and P. Neittaanmaki, *Finite Element Approximation for Optimal Shape Design*, John Wiley and Sons, Chichester, UK, 1989.

[19] B. S. He, *Solving a class of linear projection equations*, Numer. Math., 68 (1994), pp. 71–80.

[20] G. Knowles, *Finite element approximation of parabolic time optimal control problems*, SIAM J. Control Optim., 20 (1982), pp. 414–427.

[21] A. Kufner, O. John, and S. Fucik, *Function Spaces*, Nordhoff, Leyden, The Netherlands, 1977.

[22] K. Kunisch, W.-B. Liu, and N. N. Yan, *A posteriori error estimators for a model parameter estimation problem*, submitted to Proceedings of the European Conference on Numerical Mathematics and Advanced Applications, Ischia, Italy, 2001.

[23] I. Lasiecka, *Ritz–Galerkin approximation of the time optimal boundary control problem for parabolic systems with Dirichlet boundary conditions*, SIAM J. Control Optim., 22 (1984), pp. 477–500.

[24] R. Li, W.-B. Liu, and T. Tang, *Moving Mesh Method with Error-Estimator-Based Monitor and Its Application to Static Obstacle Problem*, manuscript.

[25] R. Li, T. Tang, and P.-W. Zhang, *Moving mesh methods in multiple dimensions based on harmonic maps*, J. Comput. Phys., 170 (2001), pp. 562–588.

[26] R. Li, T. Tang, and P.-W. Zhang, *A moving mesh finite element algorithm for singular problems in two and three space dimensions*, J. Comput. Phys., 177 (2002), pp. 365–393.

[27] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, 1971.

[28] W.-B. Liu and J. W. Barrett, *Quasi-norm error bounds for the finite element approximation of some degenerate quasilinear elliptic equations and variational inequalities*, RAIRO Modél. Math. Anal. Numér., 28 (1994), pp. 725–744.

[29] W.-B. Liu and D. Tiba, *Error estimates in the approximation of optimization problems governed by nonlinear operators*, Numer. Funct. Anal. Optim., 22 (2001), pp. 953–972.

[30] W.-B. Liu and N. N. Yan, *A posteriori error estimates for a model boundary optimal control problem*, J. Comput. Appl. Math., 120 (2000), pp. 159–173.

[31] W.-B. Liu and N. N. Yan, *A posteriori error analysis for control problem governed by nonlinear elliptic equations*, in Proceedings of EUNMATH99, Science Press, Singapore, 2000, pp. 146–153.
[32] W.-B. Liu and N. N. Yan, *A posteriori error estimators for a class of variational inequalities*, J. Sci. Comput., 35 (2000), pp. 361–393.
[33] W.-B. Liu and N. N. Yan, *A posteriori error estimates for convex boundary control problems*, SIAM. J. Numer. Anal., 39 (2001), pp. 73–99.
[34] W.-B. Liu and N. N. Yan, *Quasi-norm local error estimates for p-Laplacian*, SIAM. J. Numer. Anal., 39 (2001), pp. 100–127.
[35] W.-B. Liu and N. N. Yan, *A posteriori error estimates for distributed convex optimal control problems*, Adv. Comput. Math., 15 (2001), pp. 285–309.
[36] W.-B. Liu and N. N. Yan, *A posteriori error estimates for optimal control problems governed by parabolic equations*, Numer. Math., to appear.
[37] K. Malanowski, *Convergence of approximations vs. regularity of solutions for convex, control constrained, optimal control systems*, Appl. Math. Optim., 8 (1982), pp. 69–95.
[38] K. Maute, S. Schwarz, and E. Ramm, *Adaptive topology optimization of elastoplastic structures*, Structural Optim., 15 (1998), pp. 81–91.
[39] P. Neittaanmaki and D. Tiba, *Optimal Control of Nonlinear Parabolic Systems: Theory, Algorithms and Applications*, Marcel Dekker, New York, 1994.
[40] O. Pironneau, *Optimal Shape Design for Elliptic System*, Springer-Verlag, Berlin, 1984.
[41] A. Schleupen, K. Maute, and E. Ramm, *Adaptive FE-procedures in shape optimization*, Structural and Multidisciplinary Optimization, 19 (2000), pp. 282–302.
[42] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.
[43] D. Tiba and F. Troltzsch, *Error estimates for the discretization of state constrained convex control problems*, Numer. Funct. Anal. Optim., 17 (1996), pp. 1005–1028.
[44] F. Troltzsch, *Semidiscrete Ritz-Galerkin approximation of nonlinear parabolic boundary control problems—strong convergence of optimal control*, Appl. Math. Optim., 29 (1994), pp. 309–329.
[45] R. Verfürth, *A posteriori error estimators for the Stokes equations*, Numer. Math., 55 (1989), pp. 309–325.
[46] R. Verfürth, *The equivalence of a posteriori error estimators*, in Fast Solvers for Flow Problems (Kiel, 1994), Vieweg, Braunschweig, Germany, 1995, pp. 273–283.
[47] R. Verfürth, *A Review of Posteriori Error Estimation and Adaptive Mesh Refinement Techniques*, Wiley-Teubner, Chichester, UK, 1996.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 5, pp. 1350–1371

SPECTRAL ANALYSIS AND SINGULAR VALUE COMPUTATIONS OF THE NONCOMPACT FREQUENCY RESPONSE AND COMPRESSION OPERATORS IN SAMPLED-DATA SYSTEMS*

TOMOMICHI HAGIWARA†

Abstract. This paper is motivated by the problem of computing the frequency response gain of general sampled-data systems with noncompact frequency response operators. We first show that, with the J-unitary transformation, the computation in the noncompact operator case can be reduced, in principle, to that in the compact operator case, to which an existing efficient and reliable bisection method can be applied. At the same time, however, we point out that there arise some critical problems in this reduction to the compact case which could be serious enough to invalidate the apparent success in the reduction. Through some spectral analysis of operators involving or related to the frequency response operators, we eventually prove that these critical problems can be circumvented after all, and we give an explicit result that shows how to compute the frequency response gain with a bisection method dealing only with finite-dimensional matrices. Extending the arguments, we also give a bisection method to compute the singular values of the frequency response operators and the associated compression operators.

Key words. sampled-data system, frequency response, spectral analysis, singular values, bisection method, numerical computation

AMS subject classifications. 47A10, 47N70, 93B28, 93B40, 93C57

PII. S0363012901394802

1. Introduction.

1.1. Background and motivation of the study. It is quite common these days to control a continuous-time plant with a digital controller. In such a case, the measurement output y of the plant is detected and the control input u is changed at every sampling period, while the disturbance input w is a continuous-time signal, and it affects the controlled output z, which is also a continuous-time signal. This situation is shown in Figure 1, where P and Ψ denote the continuous-time plant and the digital controller, respectively, and $\mathcal{S}$ and $\mathcal{H}$ denote the ideal sampler and the hold device with sampling period h, respectively. Also, the solid lines denote continuous-time signals, while the dashed lines denote discrete-time signals. Such a system is called a sampled-data system, especially when close attention is paid to the intersample behavior of the continuous-time signals w and z.

The analysis and synthesis of sampled-data systems scored a great success in the past decade, and, for example, the servo problem, the H_∞ control problem, the H_2 control problem, as well as the robust stability problem have been studied (see, e.g., [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]). In many of these problems, the frequency response operators of sampled-data systems (with their intersample behavior taken into account), introduced in [13, 14], play an important role and/or provide an insight from the frequency domain, as well as those closely related studies [15, 16] also do.

This paper is motivated by a study on the computation of the gain characteristics of the frequency response of sampled-data systems. By computing the gain (i.e.,

*Received by the editors September 6, 2001; accepted for publication (in revised form) May 1, 2002; published electronically December 11, 2002.
http://www.siam.org/journals/sicon/41-5/39480.html
†Department of Electrical Engineering, Kyoto University, Yoshida, Sakyo-ku, Kyoto 606-8501, Japan (hagiwara@kuee.kyoto-u.ac.jp).

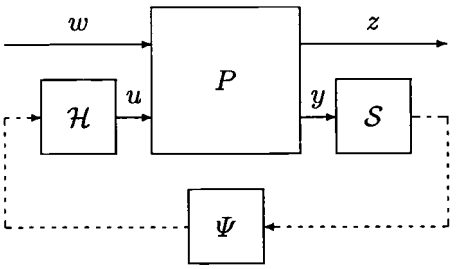


FIG. 1. *Stable sampled-data system* Σ_s.

norm) of the frequency response operator at each angular frequency, we can draw Bode diagrams of sampled-data systems with all of the effects of aliasing taken into account, which should be quite useful for grasping the frequency-domain characteristics of sampled-data systems. The computation methods for the frequency response gain of sampled-data systems have been studied quite intensively [13, 14, 17, 18, 19, 20, 21, 22], and a closed-form formula for the computation is given in [13]. Recently, a quite efficient and reliable method for the computation was derived in [22]. The feature of the latter method is that it is based on a bisection method so that it can compute the gain to any degree of accuracy without numerical problems. The main restriction of this method, however, is that it can be applied only to those sampled-data systems whose frequency response operators are compact operators. It is often the case, however, that the frequency response operator is actually noncompact. In fact, this happens to be the case if and only if there exists a nonzero direct feedthrough matrix from w to z in the continuous-time plant P [13]. Thus, when we consider, e.g., the sensitivity operators of sampled-data systems [14], they are noncompact so that the bisection method of [22] cannot be applied. Even though a closed-form formula exists for this special case (i.e., if we confine ourselves to the sensitivity operators), with which we can compute the frequency response gain exactly [19], more general cases with noncompact operators are hard to deal with in an efficient and reliable way; for example, just considering a *weighted* sensitivity operator makes the computation quite hard if the weight corresponds to a biproper multivariable system.

1.2. The purpose of this paper. This paper mainly focuses on the case in which the frequency response operator is noncompact, and the paper aims at giving a fundamental theoretical result for such a case that readily shows how to compute the frequency response gain with a bisection method. It is expected that such a theoretical result is obtained readily by a simple combination of the method for the compact case [22] and the J-unitary transformation (or the loop-shifting) [5], but, in fact, this combination turns out to lead to some critical theoretical problems. Namely, we need to guarantee that we will never encounter, roughly speaking, the following situations:

- There exists an open interval of positive real numbers such that, for each γ in that interval, a γ-dependent matrix $e^{j\varphi h}I - \mathcal{A}_\gamma$, defined appropriately, has an eigenvalue at γ.
- There exists an open interval of positive real numbers such that, for each γ in that interval, a γ-dependent operator $\mathcal{D}_\gamma$, defined appropriately, has a singular value at γ.

Even though it might look unlikely that we will encounter the above situations, the circumstances here are quite different from the seemingly related case of the H_∞ problem [2]; a different nature of the problem here prevents us from applying the small-gain theorem, which makes it nontrivial to negate the above possibility. Thus we do need to prove with some new approach that this never occurs, since otherwise the implication will be that we cannot derive a bisection method for the noncompact case. Now, the contribution of this paper is twofold: first, we describe the details of the above critical problems and show how we can circumvent them to arrive at a bisection method for the noncompact case. In the course of this study, we will also clarify some useful properties on the singular values of the frequency response operators and compression operators of sampled-data systems together with some related spectral properties; this constitutes the second contribution of this paper.

The contents of this paper are as follows. In section 2, we review the results of [13] about the lifting approach to the frequency response of sampled-data systems with a slight but crucial extension and derive a few fundamental results about some related operators. In section 3, we show that the application of the J-unitary transformation (but in a slightly different way from the H_∞ problem case) reduces the treatment of the noncompact case into that of the compact case in principle, and then we show that a finite-dimensional test can be obtained to examine if the gain at a given angular frequency is smaller than a prescribed positive number γ, *provided that some assumptions are satisfied*. In fact, such assumptions are nothing but the assumptions that neither of the two situations mentioned above occurs. To validate these assumptions, we show in section 4 that the first situation never occurs, while in section 5, we show that the second situation never occurs either. Combining these arguments, eventually we will arrive at a fundamental theoretical result that readily leads to a bisection method for the computation of the frequency response gain. In section 6, we extend some of the arguments used in the preceding sections and give a bisection method for the computation of the singular values of the frequency response operators and the associated compression operators. Finally, in section 7, we summarize the contributions of this paper.

2. Frequency response of sampled-data systems. In this section, we consider the sampled-data system Σ_s shown in Figure 1 and review the associated frequency response operator introduced by Yamamoto and Khargonekar [13] with the lifting technique [1, 2, 3, 13], with a slight but crucial extension. We next study some properties of the spectra of operators involving or related to the frequency response operator.

We assume that the state-space descriptions of P and Ψ are given, respectively, by

$$\begin{aligned} \frac{dx}{dt} &= Ax + B_1 w + B_2 u, \\ z &= C_1 x + D_{11} w + D_{12} u, \\ y &= C_2 x \end{aligned} \tag{1}$$

and

$$\begin{aligned} \xi_{k+1} &= A_\Psi \xi_k + B_\Psi y_k, \\ u_k &= C_\Psi \xi_k + D_\Psi y_k, \end{aligned} \tag{2}$$

where y_k stands for $y(kh)$, while $u(t) = u_k$ $(kh \le t < (k+1)h)$ since $\mathcal{H}$ is the zero-order hold. Throughout the paper, the sampled-data system Σ_s is assumed to be

internally stable.

With a slight abuse of notation, in the following, the Hilbert space of square (Lebesgue) integrable vector functions over the time interval $[0, h)$ with the standard inner product will be denoted by $\mathcal{K}$, whatever the dimension of the vector may be. Similarly, for notational simplicity, every finite-dimensional Euclidean space will be denoted by $\mathcal{F}$.

2.1. Frequency response operators. Now we introduce the matrices A_d, B_{d2}, and C_{d2} and the operators $\mathbf{B}_1$, $\mathbf{C}_1$, $\mathbf{D}_{11}$, and $\mathbf{D}_{12}$ as follows:

$$A_d := \exp(Ah), \quad B_{d2} := \int_0^h \exp(A\sigma)B_2 d\sigma, \quad C_{d2} := C_2, \tag{3}$$

$$\mathbf{B}_1 : \ \mathcal{K} \ni w \mapsto \int_0^h \exp\left(A(h-\sigma)\right)B_1 w(\sigma)d\sigma \in \mathcal{F}, \tag{4}$$

$$\mathbf{C}_1 : \ \mathcal{F} \ni x \mapsto z \in \mathcal{K}, \quad z(\theta) = C_1 \exp\left(A\theta\right)x, \tag{5}$$

$$\mathbf{D}_{11} : \ \mathcal{K} \ni w \mapsto z \in \mathcal{K}, \quad z(\theta) = \int_0^\theta C_1 \exp\left(A(\theta-\sigma)\right)B_1 w(\sigma)d\sigma + D_{11}w(\theta), \tag{6}$$

$$\mathbf{D}_{12} : \ \mathcal{F} \ni u \mapsto z \in \mathcal{K}, \quad z(\theta) = \int_0^\theta C_1 \exp\left(A(\theta-\sigma)\right)B_2 d\sigma\, u + D_{12}u. \tag{7}$$

Then the lifting-based transfer operator $\widehat{G}(z)$ of the sampled-data system Σ_s is defined by

$$\widehat{G}(z) := \mathcal{C}(zI - \mathcal{A})^{-1}\mathcal{B} + \mathcal{D} \tag{8}$$

with

$$\mathcal{A} := \begin{bmatrix} A_d + B_{d2}D_\Psi C_{d2} & B_{d2}C_\Psi \\ B_\Psi C_{d2} & A_\Psi \end{bmatrix} : \ \mathcal{F} \to \mathcal{F}, \tag{9}$$

$$\mathcal{B} := \begin{bmatrix} \mathbf{B}_1 \\ 0 \end{bmatrix} : \ \mathcal{K} \to \mathcal{F}, \tag{10}$$

$$\mathcal{C} := \begin{bmatrix} \mathbf{C}_1 & \mathbf{D}_{12} \end{bmatrix} \begin{bmatrix} I & 0 \\ D_\Psi C_{d2} & C_\Psi \end{bmatrix} : \ \mathcal{F} \to \mathcal{K}, \tag{11}$$

$$\mathcal{D} := \mathbf{D}_{11} : \ \mathcal{K} \to \mathcal{K}. \tag{12}$$

The operator $\mathcal{D}$ is called the compression operator.

Note that $\mathcal{A}$ is a finite-dimensional matrix, and its eigenvalues all lie inside the unit circle by the internal stability assumption of Σ_s. Hence $\widehat{G}(e^{j\varphi h})$ is well defined for each $\varphi \in [0, \omega_s)$, where $\omega_s := 2\pi/h$ denotes the sampling angular frequency. In fact, it defines a bounded operator on $\mathcal{K}$ for each φ and is called the frequency response

operator at angular frequency φ. This terminology is justified by the fact that the asymptotic output z to the "sampled-data (SD)-sinusoid" w is again an "SD-sinusoid" and that these two "SD-sinusoids" are related through $\widehat{G}(e^{j\varphi h})$ in a simple way. What this exactly means is described below in terms of the lifting technique [1, 2, 3, 13]; the discussions here are only a slight modification of that in [13] in the sense that the treatment of the initial state is considered more explicitly. This consideration, however, is crucial to giving a solid basis for the subsequent arguments.

Given a (vector) signal w over the nonnegative time interval $[0, \infty)$, the well-known lifting operation of w is defined as

$$w \mapsto \{\widehat{w}_k\}_{k=0}^{\infty}, \tag{13}$$

where $\widehat{w}_k$ is given by

$$\widehat{w}_k(\theta) = w(kh + \theta) \quad (0 \le \theta < h,\ k = 0, 1, 2, \ldots). \tag{14}$$

The signal w is called an SD-sinusoid of angular frequency φ [14] if its lifted representation satisfies

$$\widehat{w}_k(\theta) = \widehat{w}_0(\theta) e^{j\varphi kh} \quad (0 \le \theta < h) \tag{15}$$

for some $\widehat{w}_0 \in \mathcal{K}$. In this case, the "initial function" $\widehat{w}_0$ represents the "amplitude" and "phase" of the SD-sinusoid. It is a fact that the asymptotic output z of Σ_s to the SD-sinusoid w corresponding to (15) is again an SD-sinusoid of the same angular frequency, where the "initial function" $\widehat{z}_0$ corresponding to the asymptotic output is given by

$$\widehat{z}_0 = \widehat{G}(e^{j\varphi h})\widehat{w}_0. \tag{16}$$

Note carefully that $\widehat{z}_0(\theta)$ $(0 \le \theta < h)$ above is different from the actual response $z(t)$ $(0 \le t < h)$ of Σ_s for the *zero initial state* because z is not exactly an SD-sinusoid, but it just tends to an SD-sinusoid as t goes to infinity.

However, given any $\widehat{w}_0 \in \mathcal{K}$, let us take $\widehat{z}_0$ given by (16) and expand both $\widehat{w}_0$ and $\widehat{z}_0$ into the SD-sinusoids w and z, respectively, with angular frequency φ according to (15) and then (14). Then it is a fact that there exist appropriate initial states $x(0)$ of P and ξ_0 of Ψ such that the above SD-sinusoid input w *together with the initial states* yield exactly the above SD-sinusoid output z over the entire nonnegative time interval $[0, \infty)$. Conversely, if, under some initial states $x(0)$ and ξ_0, the output z to some SD-sinusoid w is exactly an SD-sinusoid over the whole nonnegative time interval $[0, \infty)$, then the "initial functions" of these SD-sinusoids satisfy (16).

These facts justify the following alternative definition for the frequency response gain defined in [13].

DEFINITION 2.1. *Suppose that $e^{j\varphi h} I - \mathcal{A}$ is invertible. Then the frequency response gain of Σ_s at angular frequency φ is given by*

$$\|\widehat{G}(e^{j\varphi h})\| = \sup \frac{\|\widehat{z}_0\|_{\mathcal{K}}}{\|\widehat{w}_0\|_{\mathcal{K}}}, \tag{17}$$

where $\widehat{w}_0$ and $\widehat{z}_0$ are, respectively, the initial functions of the input and output SD-sinusoids of angular frequency φ consistent with the sampled-data system Σ_s. Here, the consistency means that these SD-sinusoids can be the solution of Σ_s under some appropriate initial states.

Remark 2.1. Definition 2.1 is a system-theoretic definition as opposed to the more operator-theoretic one in [13]; unlike the operator-theoretic definition, we can allow nonzero initial states here and also deal with unstable sampled-data systems.[1] This, together with the notion of "consistency," plays a crucial role in the following discussions. See, e.g., the proof of Lemma 3.1 and subsection 4.1.

Here, it should be noted that $\mathcal{D}$ (and hence $\widehat{G}(e^{j\varphi h})$) is a compact operator if and only if $D_{11} = 0$ in (1) [13]. In general, noncompact operators are harder to deal with than compact operators, and we are mainly interested in the computation of the frequency response gain when $\widehat{G}(e^{j\varphi h})$ is not compact (i.e., $D_{11} \neq 0$). Such situations arise, e.g., when we deal with the sensitivity operators [14] among others. Even though there exists an explicit formula [19] for the frequency response gain of (unweighted) sensitivity operators, it is hard to compute the frequency response gain of general sampled-data systems with $D_{11} \neq 0$. The study in the following subsection is motivated by such a research direction.

2.2. Spectral analysis related to the frequency response operators. In this subsection, we derive some facts on the spectra of operators related to the frequency response operator $\widehat{G}(e^{j\varphi h})$. We also study the singular values of $\widehat{G}(e^{j\varphi h})$. Throughout the remainder of this paper, the angular frequency $\varphi \in [0, \omega_s)$ is regarded as an arbitrary fixed number.

If $D_{11} = 0$ so that $\widehat{G}(e^{j\varphi h})$ is compact, then $\widehat{G}(e^{j\varphi h})^*\widehat{G}(e^{j\varphi h})$ is compact, too, and hence it has a maximum eigenvalue. The square root of it is the maximum singular value of $\widehat{G}(e^{j\varphi h})$ and is equal to the frequency response gain $\|\widehat{G}(e^{j\varphi h})\|$ [13].

If $D_{11} \neq 0$, however, the situation becomes much more involved. Although it is known that $\|\widehat{G}(e^{j\varphi h})\|$ is no smaller than $\|D_{11}\|$ [13], the properties of the spectra of operators related to $\widehat{G}(e^{j\varphi h})$, $\widehat{G}(e^{j\varphi h})^*\widehat{G}(e^{j\varphi h})$, or $\mathcal{D}^*\mathcal{D}$ have not necessarily been clarified explicitly enough in the literature. Since we need, in the following arguments, further knowledge on the properties of the spectra of operators involving $\widehat{G}(e^{j\varphi h})$ or $\mathcal{D}$, the remainder of this subsection is devoted to such a study.

With a slight abuse of notation, given a matrix D_{11}, the operator that maps $w(\cdot) \in \mathcal{K}$ to $z(\cdot) = D_{11}w(\cdot) \in \mathcal{K}$ is also denoted by D_{11}. It will be clear from the context whether D_{11} refers to this operator or the underlying matrix. Also, consider replacing the direct feedthrough matrix D_{11} by 0 in the generalized plant (1), and denote the corresponding frequency response operator by $\widehat{G}_c(e^{j\varphi h})$. Then we have

$$\widehat{G}(e^{j\varphi h}) = D_{11} + \widehat{G}_c(e^{j\varphi h}). \tag{18}$$

Note that $\widehat{G}_c(e^{j\varphi h})$ is a compact operator. In the following, the spectrum of an operator is denoted by $\sigma(\cdot)$, and the matrix norm $\|D_{11}\|$ is denoted by d_{11} for simplicity.

LEMMA 2.2. *Suppose that* $\gamma_1^2 \in \partial\sigma(\widehat{G}(e^{j\varphi h})D_{11}^*)$, *where* $\gamma_1 > d_{11}$. *Then* γ_1^2 *is an isolated point of* $\sigma(\widehat{G}(e^{j\varphi h})D_{11}^*)$.

Proof. Since $\widehat{G}_c(e^{j\varphi h})$ is a compact operator, it follows from (18) that

$$\widehat{G}(e^{j\varphi h})D_{11}^* = D_{11}D_{11}^* + K, \tag{19}$$

[1] Unless Σ_s has an unstable mode at $e^{j\varphi h}$, every SD-sinusoid input with angular frequency φ yields an SD-sinusoid output with the same angular frequency provided that the initial state of Σ_s is set appropriately. This situation is quite similar to what we have in unstable continuous-time systems with regard to sinusoid signals.

where K is a compact operator. Hence, by Proposition XI.4.2(e) of [23], we have

$$\sigma_{le}(\widehat{G}(e^{j\varphi h})D_{11}^*) = \sigma_{le}(D_{11}D_{11}^*), \tag{20}$$

$$\sigma_{re}(\widehat{G}(e^{j\varphi h})D_{11}^*) = \sigma_{re}(D_{11}D_{11}^*), \tag{21}$$

$$\sigma_{e}(\widehat{G}(e^{j\varphi h})D_{11}^*) = \sigma_{e}(D_{11}D_{11}^*), \tag{22}$$

where $\sigma_{le}(\cdot)$ and $\sigma_{re}(\cdot)$, respectively, denote the left and the right essential spectra and $\sigma_e(\cdot)$ denotes the essential spectrum. By Proposition XI.4.6 of [23], the right-hand sides of the above three equations all coincide so that we have

$$\sigma_{le}(\widehat{G}(e^{j\varphi h})D_{11}^*) \cap \sigma_{re}(\widehat{G}(e^{j\varphi h})D_{11}^*) = \sigma_e(D_{11}D_{11}^*), \tag{23}$$

which, in turn, is equal to the set of the squared singular values of the matrix D_{11}. Thus we have

$$\gamma_1^2 \notin \sigma_{le}(\widehat{G}(e^{j\varphi h})D_{11}^*) \cap \sigma_{re}(\widehat{G}(e^{j\varphi h})D_{11}^*) \tag{24}$$

since $\gamma_1 > d_{11}$. Hence the assertion follows readily from Theorem XI.6.8 of [23]. □

LEMMA 2.3. *Suppose that $\gamma^2 \in \sigma(\widehat{G}(e^{j\varphi h})D_{11}^*)$, where $\gamma > d_{11}$. Then γ^2 is an isolated point of $\sigma(\widehat{G}(e^{j\varphi h})D_{11}^*)$.*

Proof. If $\gamma^2 \in \partial\sigma(\widehat{G}(e^{j\varphi h})D_{11}^*)$, then the assertion follows readily from Lemma 2.2. If $\gamma^2 \notin \partial\sigma(\widehat{G}(e^{j\varphi h})D_{11}^*)$, on the other hand, then, by definition, there exists some ε-neighborhood of γ^2 that fails to contain a point in the complement of $\sigma(\widehat{G}(e^{j\varphi h})D_{11}^*)$ (i.e., the ε-neighborhood is contained in $\sigma(\widehat{G}(e^{j\varphi h})D_{11}^*)$). This, together with the compactness (boundedness) of $\sigma(\widehat{G}(e^{j\varphi h})D_{11}^*)$, means that we can take some real number $\gamma_1 > \gamma$ such that $\gamma_1^2 \in \partial\sigma(\widehat{G}(e^{j\varphi h})D_{11}^*)$, and, at the same time, γ_1^2 is not an isolated point of $\sigma(\widehat{G}(e^{j\varphi h})D_{11}^*)$. Since $\gamma_1 > d_{11}$, this contradicts Lemma 2.2, and thus $\gamma^2 \notin \partial\sigma(\widehat{G}(e^{j\varphi h})D_{11}^*)$ cannot occur. This completes the proof. □

Similarly, applying Proposition XI.4.6 of [23], we can readily obtain the following result.

PROPOSITION 2.4. *Every $\gamma \in \sigma(\widehat{G}(e^{j\varphi h})^*\widehat{G}(e^{j\varphi h}))$ such that $\gamma > d_{11}^2$ is an isolated point of $\sigma(\widehat{G}(e^{j\varphi h})^*\widehat{G}(e^{j\varphi h}))$ and is in fact an eigenvalue of finite multiplicity.*

Also, by decomposing $\mathbf{D}_{11}$ given by (6) into $\mathbf{D}_{11} = D_{11} + \mathbf{D}_{11c}$ with a compact operator $\mathbf{D}_{11c}$, we can derive the following parallel result.

PROPOSITION 2.5. *Every $\gamma \in \sigma(\mathcal{D}^*\mathcal{D})$ such that $\gamma > d_{11}^2$ is an isolated point of $\sigma(\mathcal{D}^*\mathcal{D})$ and is in fact an eigenvalue of finite multiplicity.*

The square root of each of the eigenvalues of $\widehat{G}(e^{j\varphi h})^*\widehat{G}(e^{j\varphi h})$ described in Proposition 2.4 is a singular value of $\widehat{G}(e^{j\varphi h})$ larger than d_{11}, including multiplicity [24, p. 214]. Furthermore, unless there are infinitely many eigenvalues of $\widehat{G}(e^{j\varphi h})^*\widehat{G}(e^{j\varphi h})$ strictly larger than d_{11}^2, the number d_{11} is also a singular value of $\widehat{G}(e^{j\varphi h})$ (in which case, d_{11} is the smallest singular value of infinite multiplicity). In the following, the ith singular value of an operator is denoted by $s_i(\cdot)$, where $s_1(\cdot) \geq s_2(\cdot) \geq \cdots \geq 0$. Then $\lim_{i\to\infty} s_i(\widehat{G}(e^{j\varphi h})) = d_{11}$, and the maximum singular value $s_1(\widehat{G}(e^{j\varphi h}))$ (which could be equal to d_{11}) equals $\|\widehat{G}(e^{j\varphi h})\|$ [24]. The same holds true for Proposition 2.5 regarding the singular values and norm of $\mathcal{D}$. It will be helpful to keep in mind the above propositions as well as these definitions of singular values in the following arguments.

3. Bisection method and problem descriptions. For the case in which $\widehat{G}(e^{j\varphi h})$ is a compact operator, a quite efficient and reliable method for the computation of $\|\widehat{G}(e^{j\varphi h})\|$ (as well as $s_i(\widehat{G}(e^{j\varphi h}))$) was developed recently [22]. Given $\gamma > 0$, this method reduces the test of the condition $\|\widehat{G}(e^{j\varphi h})\| < \gamma$ or $s_i(\widehat{G}(e^{j\varphi h})) < \gamma$ into a finite-dimensional test, through an infinite-dimensional congruence transformation. Thus a bisection method that deals only with finite-dimensional matrices was established which allows us to compute $\|\widehat{G}(e^{j\varphi h})\|$ or $s_i(\widehat{G}(e^{j\varphi h}))$ to any degree of accuracy.

It is also suggested in [22] that this method can be extended to the noncompact case as well, in principle, to get a finite-dimensional test for $\|\widehat{G}(e^{j\varphi h})\| < \gamma$. In this section, we explicitly describe the extension to the noncompact case; the primary purpose here is to point out that there actually arise a few critical issues in that extension. As it turns out, one might argue that these issues could be serious enough to invalidate the apparent extension to the noncompact case. Indeed, this paper was motivated through an effort to address these issues in full rigor, which will be the topics of the following sections, whereas this section intends to describe the details of the issues.

3.1. Reduction to the compact case. The basic idea for the extension to the noncompact case suggested in [22] is to "remove" the matrix D_{11} so that the resulting frequency response operator becomes compact. This is achieved by the introduction of J-unitary (or unitary) transformations (or loop-shifting) [2, 5]. Note, however, that the way we apply this transformation is quite different from the seemingly related treatment in the sampled-data H_∞ problem [2] in two respects. The first is that we apply it to "remove" D_{11}, while in [2] it was applied to "remove" $\mathcal{D}$. From this difference arises the second difference that γ smaller than $\|\mathcal{D}\|$ (actually, $\|D_{11}\| < \gamma < \|\mathcal{D}\|$) should also be used as opposed to the H_∞ problem setting. Furthermore, we are interested not only in the largest singular value (i.e., the norm) but also in other singular values of the frequency response or compression operator. These differences necessitate quite different arguments in what follows from those in [2], for which basically the small-gain theorem was enough. See, e.g., the proof of Lemma 3.1 and Remark 3.1.

We begin with a routine method of loop-shifting. Suppose that $\gamma > d_{11}(= \|D_{11}\|)$. As is well known, the J-unitary transformation at level γ is

$$\begin{bmatrix} z \\ w \end{bmatrix} = S_\gamma \begin{bmatrix} z_J \\ w_J \end{bmatrix}, \tag{25}$$

where z_J and w_J are newly introduced signals and S_γ is defined by

$$S_\gamma = \begin{bmatrix} \gamma(\gamma^2 I - D_{11}D_{11}^T)^{-1/2} & \gamma(\gamma^2 I - D_{11}D_{11}^T)^{-1/2} D_{11} \\ \gamma^{-1} D_{11}^T(\gamma^2 I - D_{11}D_{11}^T)^{-1/2} & \gamma(\gamma^2 I - D_{11}^T D_{11})^{-1/2} \end{bmatrix}. \tag{26}$$

Defining J_γ as

$$J_\gamma = \begin{bmatrix} I & 0 \\ 0 & -\gamma^2 I \end{bmatrix}, \tag{27}$$

we have

$$S_\gamma^* J_\gamma S_\gamma = J_\gamma. \tag{28}$$

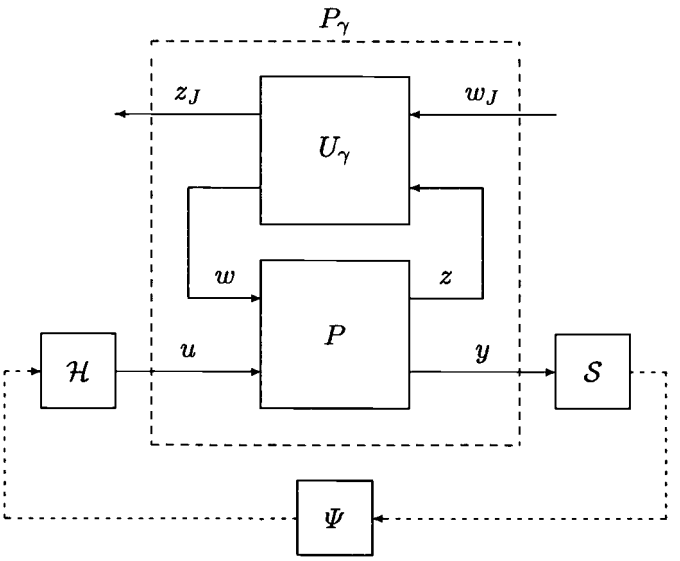


FIG. 2. *J-unitary transformed sampled-data system* Σ_J.

An alternative (equivalent) representation to (25) is

$$\begin{bmatrix} z_J \\ w \end{bmatrix} = U_\gamma \begin{bmatrix} w_J \\ z \end{bmatrix}, \tag{29}$$

where

$$U_\gamma = \begin{bmatrix} -D_{11} & \gamma^{-1}(\gamma^2 I - D_{11}D_{11}^T)^{1/2} \\ \gamma^{-1}(\gamma^2 I - D_{11}^T D_{11})^{1/2} & \gamma^{-2} D_{11}^T \end{bmatrix}. \tag{30}$$

It is easy to see that U_γ is a unitary (orthogonal) matrix when $\gamma = 1$.

Now, the J-unitary transformation at level γ applied to the sampled-data system Σ_s leads to the sampled-data system Σ_J shown in Figure 2. Here, substituting (25) into (1), we can see that P_γ shown in Figure 2 is described by

$$\begin{aligned} \frac{dx}{dt} &= A_\gamma x + B_{1\gamma} w_J + B_{2\gamma} u, \\ z_J &= C_{1\gamma} x + D_{12\gamma} u, \\ y &= C_{2\gamma} x, \end{aligned} \tag{31}$$

where

$$A_\gamma = A + B_1 D_{11}^T (\gamma^2 I - D_{11} D_{11}^T)^{-1} C_1, \tag{32}$$

$$B_{1\gamma} = \gamma B_1 (\gamma^2 I - D_{11}^T D_{11})^{-1/2}, \quad B_{2\gamma} = B_2 + B_1 D_{11}^T (\gamma^2 I - D_{11} D_{11}^T)^{-1} D_{12}, \tag{33}$$

$$C_{1\gamma} = \gamma (\gamma^2 I - D_{11} D_{11}^T)^{-1/2} C_1, \quad C_{2\gamma} = C_2, \tag{34}$$

$$D_{12\gamma} = \gamma (\gamma^2 I - D_{11} D_{11}^T)^{-1/2} D_{12}. \tag{35}$$

Since the sampled-data system Σ_J is nothing but Σ_s with P replaced by P_γ, we can readily introduce the frequency response operator $\widehat{G}_\gamma(e^{j\varphi h})$ for Σ_J. Defining $\mathcal{A}_\gamma$, $\mathcal{B}_\gamma$, $\mathcal{C}_\gamma$, and $\mathcal{D}_\gamma$ corresponding to (9), (10), (11), and (12), respectively, $\widehat{G}_\gamma(e^{j\varphi h})$ is formally given by

$$\widehat{G}_\gamma(e^{j\varphi h}) = \mathcal{C}_\gamma(e^{j\varphi h}I - \mathcal{A}_\gamma)^{-1}\mathcal{B}_\gamma + \mathcal{D}_\gamma. \tag{36}$$

Also, comparing (31) with (1), we can see that "D_{11} has been removed." Therefore, $\mathcal{D}_\gamma$ and hence $\widehat{G}_\gamma(e^{j\varphi h})$ are both compact operators. This is one of the well-known important properties of the J-unitary transformation; the following lemma gives another quite important property, which corresponds to the rigorous arguments for what has been suggested in [17] as to the treatment of nonzero D_{11} matrices. Note, however, that the strengthened assertion for the singular values for the sampled-data system is nontrivial; at least loop-shifting is usually not intended to yield such properties.

LEMMA 3.1. *For each $\gamma > d_{11}$ such that $e^{j\varphi h}I - \mathcal{A}_\gamma$ is invertible, we have that $\|\widehat{G}_\gamma(e^{j\varphi h})\| < \gamma$ if and only if $\|\widehat{G}(e^{j\varphi h})\| < \gamma$. More generally, for each $\gamma > d_{11}$ such that $e^{j\varphi h}I - \mathcal{A}_\gamma$ is invertible and for every positive integer i, we have that $s_i(\widehat{G}_\gamma(e^{j\varphi h})) < \gamma$ if and only if $s_i(\widehat{G}(e^{j\varphi h})) < \gamma$.*

Since $\|\widehat{G}(e^{j\varphi h})\|$ or $s_i(\widehat{G}(e^{j\varphi h}))$ is no less than d_{11} as mentioned in the preceding section, it is enough to use γ larger than d_{11} when we compute $\|\widehat{G}(e^{j\varphi h})\|$ or $s_i(\widehat{G}(e^{j\varphi h}))$ with a bisection method. Thus the above lemma implies that the computation of the norm or singular values of the noncompact operator $\widehat{G}(e^{j\varphi h})$ with a bisection method can be reduced essentially to the computation of those of the compact operator $\widehat{G}_\gamma(e^{j\varphi h})$.

Proof of Lemma 3.1. First note that S_γ is invertible with its inverse given by

$$S_\gamma^{-1} = \begin{bmatrix} \gamma(\gamma^2 I - D_{11}D_{11}^T)^{-1/2} & -\gamma(\gamma^2 I - D_{11}D_{11}^T)^{-1/2}D_{11} \\ -\gamma^{-1}D_{11}^T(\gamma^2 I - D_{11}D_{11}^T)^{-1/2} & \gamma(\gamma^2 I - D_{11}^T D_{11})^{-1/2} \end{bmatrix}. \tag{37}$$

Hence it is easy to see from (25) that w and z are SD-sinusoids of angular frequency φ if and only if w_J and z_J are. In other words, for each $\gamma > d_{11}$, the J-unitary transformation induces a one-to-one correspondence between the input and output SD-sinusoids of angular frequency φ consistent with the sampled-data system Σ_s and those consistent with the sampled-data system Σ_J.

Now, suppose that $s_i(\widehat{G}(e^{j\varphi h})) < \gamma$ or, equivalently (by the min-max characterization of singular values [25]),

$$\inf_{\mathcal{V}} \|\widehat{G}(e^{j\varphi h})|\mathcal{V}\| < \gamma, \tag{38}$$

where $\widehat{G}(e^{j\varphi h})|\mathcal{V}$ stands for the restriction of $\widehat{G}(e^{j\varphi h})$ to $\mathcal{V}$, and $\mathcal{V}$ ranges over the closed subspaces of $\mathcal{K}$ of codimension at most $i-1$. Then there exists some $\mathcal{V}$ of codimension at most $i-1$ such that $\|\widehat{G}(e^{j\varphi h})|\mathcal{V}\| < \gamma$. Hence we have some $\varepsilon > 0$ such that

$$\|\widehat{z}_0\|_{\mathcal{K}}^2 - (\gamma^2 - \varepsilon)\|\widehat{w}_0\|_{\mathcal{K}}^2 \le 0 \tag{39}$$

for every pair of the input and output SD-sinusoids consistent with the sampled-data system Σ_s such that $\widehat{w}_0 \in \mathcal{V}$. Here, by (27), the inequality (39) can be rearranged as

$$\int_0^h \begin{bmatrix} \widehat{z}_0(\theta) \\ \widehat{w}_0(\theta) \end{bmatrix}^* J_\gamma \begin{bmatrix} \widehat{z}_0(\theta) \\ \widehat{w}_0(\theta) \end{bmatrix} d\theta + \varepsilon\|\widehat{w}_0\|_{\mathcal{K}}^2 \le 0. \tag{40}$$

By (25) and (28), this inequality can be rearranged further as

$$\int_0^h \begin{bmatrix} \widehat{z}_{J0}(\theta) \\ \widehat{w}_{J0}(\theta) \end{bmatrix}^* J_\gamma \begin{bmatrix} \widehat{z}_{J0}(\theta) \\ \widehat{w}_{J0}(\theta) \end{bmatrix} d\theta + \varepsilon \|\widehat{w}_0\|_{\mathcal{K}}^2 \le 0 \tag{41}$$

or, equivalently,

$$\|\widehat{z}_{J0}\|_{\mathcal{K}}^2 - \gamma^2 \|\widehat{w}_{J0}\|_{\mathcal{K}}^2 + \varepsilon \|\widehat{w}_0\|_{\mathcal{K}}^2 \le 0. \tag{42}$$

Now, denoting the maximum singular value of S_γ^{-1} by σ_γ, we can see from (25) that

$$\begin{aligned} \|\widehat{w}_{J0}\|_{\mathcal{K}}^2 &\le \left\| \begin{bmatrix} \widehat{z}_{J0} \\ \widehat{w}_{J0} \end{bmatrix} \right\|_{\mathcal{K}}^2 \\ &\le \sigma_\gamma^2 \left\| \begin{bmatrix} \widehat{z}_0 \\ \widehat{w}_0 \end{bmatrix} \right\|_{\mathcal{K}}^2 \\ &= \sigma_\gamma^2 (\|\widehat{z}_0\|_{\mathcal{K}}^2 + \|\widehat{w}_0\|_{\mathcal{K}}^2) \\ &\le \sigma_\gamma^2 (1+\gamma^2) \|\widehat{w}_0\|_{\mathcal{K}}^2. \end{aligned} \tag{43}$$

Substituting the above into (42), we have

$$\|\widehat{z}_{J0}\|_{\mathcal{K}}^2 - \gamma^2 \|\widehat{w}_{J0}\|_{\mathcal{K}}^2 + \varepsilon_1 \|\widehat{w}_{J0}\|_{\mathcal{K}}^2 \le 0, \tag{44}$$

where

$$\varepsilon_1 = \frac{\varepsilon}{\sigma_\gamma^2 (1+\gamma^2)} > 0. \tag{45}$$

As $\widehat{w}_0$ ranges over $\mathcal{V}$, it is obvious that $\widehat{w}_{J0}$ ranges over some closed subspace $\mathcal{V}_J$ of $\mathcal{K}$ of codimension at most $i-1$ since

$$S_\gamma^{-1} \begin{bmatrix} \widehat{G}(e^{j\varphi h}) \\ I \end{bmatrix} \widehat{w}_0 = \begin{bmatrix} \widehat{G}_\gamma(e^{j\varphi h}) \\ I \end{bmatrix} \widehat{w}_{J0}. \tag{46}$$

Hence (44) implies $\|\widehat{G}_\gamma(e^{j\varphi h})|\mathcal{V}_J\| < \gamma$. Thus we have $s_i(\widehat{G}_\gamma(e^{j\varphi h})) < \gamma$.

Conversely, we can show that $s_i(\widehat{G}_\gamma(e^{j\varphi h})) < \gamma$ implies $s_i(\widehat{G}(e^{j\varphi h})) < \gamma$ in a similar way. This completes the proof. □

3.2. Reduction to a finite-dimensional problem. Let us defer the computation of $s_i(\widehat{G}(e^{j\varphi h}))$ to section 6 and confine ourselves to the computation of $\|\widehat{G}(e^{j\varphi h})\|$ here for simplicity. As mentioned before, Lemma 3.1 shows that computing the norm of the noncompact operator $\widehat{G}(e^{j\varphi h})$ can be reduced essentially to computing the norm of the compact operator $\widehat{G}_\gamma(e^{j\varphi h})$. The latter computation has been studied rather intensively in the literature; see, e.g., [13, 14, 18, 20, 21, 22]. Among them, the bisection method developed in [22] will be the best method to use in our context, partly because it is known to be quite efficient, but more importantly because we need only to know whether or not $\|\widehat{G}_\gamma(e^{j\varphi h})\| < \gamma$ for the given number γ; it is meaningless to compute the value itself of $\|\widehat{G}_\gamma(e^{j\varphi h})\|$ exactly. The techniques developed in [22] allow us to check if $\|\widehat{G}_\gamma(e^{j\varphi h})\| < \gamma$ quite efficiently, whereas other methods [13, 14, 18, 20, 21] can compute only the value itself of $\|\widehat{G}_\gamma(e^{j\varphi h})\|$.

Thus, in the following, we focus on the arguments developed in [22]. Also, as in [22], let us introduce the notation $\mathrm{N}(\cdot)$; for a finite-dimensional Hermitian matrix X,

the expression $\mathrm{N}(X) = (q_1, q_2)$ implies that X has q_1 repeated zero eigenvalues, while it has q_2 negative eigenvalues (counted according to multiplicities). Also, the size of the square matrix $\mathcal{A}$ is denoted by n (i.e., $n := \dim(x)+\dim(\xi)$, where x and ξ are as in (1) and (2), respectively). Then we can obtain the following theorem, which provides a method to check if $\|\widehat{G}(e^{j\varphi h})\| < \gamma$ with only finite-dimensional computations.

THEOREM 3.2. *Let $\gamma > d_{11}$. Suppose that $e^{j\varphi h}I - \mathcal{A}_\gamma$ is invertible and γ is not a singular value of $\mathcal{D}_\gamma$. Let ν be the number (counted according to multiplicities) of the singular values of $\mathcal{D}_\gamma$ larger than γ. Then the following three conditions are equivalent:*

$$\text{(i)} \quad \|\widehat{G}(e^{j\varphi h})\| < \gamma, \tag{47}$$

$$\text{(ii)} \quad \|\widehat{G}_\gamma(e^{j\varphi h})\| < \gamma, \tag{48}$$

$$\text{(iii)} \quad N(F_\gamma(e^{j\varphi h}, \gamma)) = (0, n-\nu). \tag{49}$$

Here $F_\gamma(z,\lambda)$ is a Hermitian matrix given by

$$F_\gamma(z,\lambda) = \begin{bmatrix} 0 & zI - \widetilde{E} \\ z^*I - \widetilde{E}^T & 0 \end{bmatrix} - \begin{bmatrix} \widetilde{B} & 0 \\ 0 & \widetilde{C}^T \end{bmatrix} L(\lambda) \begin{bmatrix} \widetilde{B}^T & 0 \\ 0 & \widetilde{C} \end{bmatrix}, \tag{50}$$

$$\widetilde{E} = \begin{bmatrix} 0 & 0 \\ B_\Psi C_{2\gamma} & A_\Psi \end{bmatrix}, \quad \widetilde{B} = \begin{bmatrix} I \\ 0 \end{bmatrix}, \quad \widetilde{C} = \begin{bmatrix} I & 0 \\ D_\Psi C_{2\gamma} & C_\Psi \end{bmatrix}, \tag{51}$$

$$L(\lambda) = \begin{bmatrix} \Gamma_{21} & \Gamma_{22} & \Gamma_{23} \\ I & 0 & 0 \\ \Gamma_{41} & \Gamma_{42} & \Gamma_{43} \end{bmatrix} \begin{bmatrix} \Gamma_{11} & \Gamma_{12} & \Gamma_{13} \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}^{-1}, \tag{52}$$

with Γ_{ij} defined by

$$\begin{bmatrix} \Gamma_{11} & \Gamma_{12} & \Gamma_{13} & 0 \\ \Gamma_{21} & \Gamma_{22} & \Gamma_{23} & 0 \\ 0 & 0 & I & 0 \\ \Gamma_{41} & \Gamma_{42} & \Gamma_{43} & I \end{bmatrix} = \exp\left(\begin{bmatrix} -A_\gamma^T & \frac{1}{\lambda} C_{1\gamma}^T C_{1\gamma} & \frac{1}{\lambda} C_{1\gamma}^T D_{12\gamma} & 0 \\ \frac{1}{\lambda} B_{1\gamma} B_{1\gamma}^T & A_\gamma & B_{2\gamma} & 0 \\ 0 & 0 & 0 & 0 \\ B_{2\gamma}^T & \frac{1}{\lambda} D_{12\gamma}^T C_{1\gamma} & \frac{1}{\lambda} D_{12\gamma}^T D_{12\gamma} & 0 \end{bmatrix} h \right). \tag{53}$$

In this theorem, the equivalence between (i) and (ii) has been proved in Lemma 3.1. On the other hand, the equivalence between (ii) and (iii) follows readily by applying entirely the same arguments as in [22]. Basically, what has been suggested in [22] as to the extension to the noncompact (i.e., nonzero D_{11}) case is just the qualitative fact that these two equivalence relations will still reduce the computation for the noncompact case into finite-dimensional computations.

However, in view of the definite statement of Theorem 3.2, it will be natural to ask the following questions:

- What should we do if $e^{j\varphi h}I - \mathcal{A}_\gamma$ happens to be noninvertible?
- What should we do if γ happens to be a singular value of $\mathcal{D}_\gamma$?

Intuitively, it might be expected that γ could be perturbed slightly to avoid such situations; a bisection method will still work even with such a perturbation on γ. In fact, when $D_{11} = 0$, the first situation never occurs, and the second situation can be avoided by such a perturbation, as can be seen easily. In the noncompact case, however, it is not trivial if a slight perturbation of γ would always resolve the difficulty, especially because $e^{j\varphi h}I - \mathcal{A}_\gamma$ and $\mathcal{D}_\gamma$ are dependent on γ. Even though it might look unlikely, it could possibly be the case that, over an open interval with respect to γ, $e^{j\varphi h}I - \mathcal{A}_\gamma$ is noninvertible and/or $\mathcal{D}_\gamma$ has γ as one of its singular values. If this is really the case, the situation is quite serious since it implies that we have no simple way to check if $\|\widehat{G}(e^{j\varphi h})\| < \gamma$ for such γ lying on the open interval.

Remark 3.1. The situation here is again quite different from the H_∞ problem, in which case the first type of question is always irrelevant. This is because $\gamma > \|\mathcal{D}\|$ can be assumed without loss of generality, and thus the stability of the matrix corresponding to $\mathcal{A}_\gamma$ can be ensured simply by the small-gain theorem (see [2] for details). However, here we need to consider much smaller γ, too, and thus $\mathcal{A}_\gamma$ can become unstable. One trivial example for this instability can be seen by considering the case with $\Psi = 0$, $\dim(x) = \dim(w) = \dim(z) = 1$, and $\gamma \downarrow d_{11}$. In the H_∞ problem, the second question is also irrelevant, which can be seen by Corollary 5.3.

In the first half of the remainder of this paper, we study the above issues and show that a slight perturbation of γ indeed resolves the difficulty; the first question will be dealt with in section 4, and the second will be dealt with in section 5. The second half, section 6, deals with an extension to singular value computations.

4. The case of noninvertible $e^{j\varphi h}I - \mathcal{A}_\gamma$. In this section, we study the case in which $e^{j\varphi h}I - \mathcal{A}_\gamma$ is noninvertible. We first show that such γ is always a strict lower bound of the frequency response gain $\|\widehat{G}(e^{j\varphi h})\|$. From this fact, we further show that the invertibility assumption of $e^{j\varphi h}I - \mathcal{A}_\gamma$ can actually be removed from Theorem 3.2.

4.1. Characterization of γ as a strict lower bound. Let us consider the "unforced" sampled-data system Σ_u shown in Figure 3. It is straightforward to see that the continuous-time part of this system is described by

$$\begin{aligned} \frac{dx}{dt} &= A_\gamma x + B_{2\gamma} u, \\ y &= C_{2\gamma} x, \end{aligned} \tag{54}$$

where A_γ, $B_{2\gamma}$, and $C_{2\gamma}$ are as given in (32), (33), and (34), respectively. Hence it follows that the state transition matrix of the discrete-time system (which we denote by Σ_{ud}) equivalent to Σ_u viewed at every sampling period h is equal to $\mathcal{A}_\gamma$, which has been derived through the application of the J-unitary transformation (recall subsection 3.1). In other words, the assumption that $e^{j\varphi h}I - \mathcal{A}_\gamma$ is noninvertible is nothing but the assumption that Σ_u (or Σ_{ud}) has a mode at $e^{j\varphi h}$. Thus, in this case, the unforced system Σ_{ud} has a nontrivial solution of the following form with some $[x(0)^T,\ \xi_0^T]^T$ $(\neq 0)$:

$$\begin{bmatrix} x(kh) \\ \xi_k \end{bmatrix} = \begin{bmatrix} x(0) \\ \xi_0 \end{bmatrix} e^{j\varphi kh}. \tag{55}$$

On the other hand, the lifted representation of w and z for the unforced system Σ_u can be described, by the linearity of Σ_u, as

$$\widehat{w}_k(\theta) = L_w(\theta) \begin{bmatrix} x(kh) \\ \xi_k \end{bmatrix}, \quad \widehat{z}_k(\theta) = L_z(\theta) \begin{bmatrix} x(kh) \\ \xi_k \end{bmatrix} \quad (0 \le \theta < h) \tag{56}$$

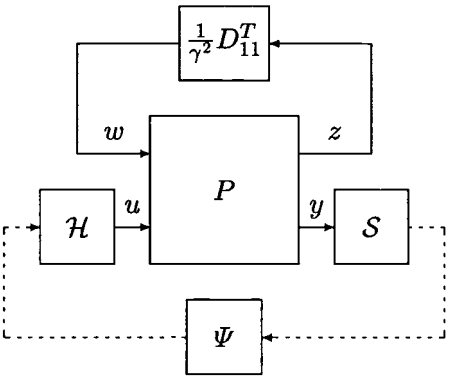


FIG. 3. *Unforced sampled-data system* Σ_u.

with some appropriate matrices $L_w(\theta)$ and $L_z(\theta)$. Substituting (55) into the above equation, we can see that Σ_u has a (nontrivial) solution of the form

$$\widehat{w}_k = \widehat{w}_0 e^{j\varphi kh}, \quad \widehat{z}_k = \widehat{z}_0 e^{j\varphi kh}, \tag{57}$$

where

$$\widehat{w}_0 = L_w \begin{bmatrix} x(0) \\ \xi_0 \end{bmatrix}, \quad \widehat{z}_0 = L_z \begin{bmatrix} x(0) \\ \xi_0 \end{bmatrix}. \tag{58}$$

Since (57) are SD-sinusoids, we have

$$\widehat{z}_0 = \widehat{G}(e^{j\varphi h})\widehat{w}_0 \tag{59}$$

by the definition of the frequency response operator (recall the discussions above Definition 2.1).

Now, from Figure 3, we readily have

$$\widehat{w}_0 = \frac{1}{\gamma^2} D_{11}^* \widehat{z}_0. \tag{60}$$

Substituting the above into (59), we have

$$\left(I - \frac{1}{\gamma^2} \widehat{G}(e^{j\varphi h}) D_{11}^* \right) \widehat{z}_0 = 0. \tag{61}$$

Here we can show that $\widehat{z}_0 \not\equiv 0$. To see this, suppose the contrary. Then we have $\widehat{w}_0 \equiv 0$ by (60), which implies that $w \equiv 0$. In this case, the unforced system Σ_u is nothing but the stable sampled-data system Σ_s so that it can never have the nontrivial solution like (55). By contradiction, we have that $\widehat{z}_0 \not\equiv 0$. Hence (61) implies that γ^2 is an eigenvalue of $\widehat{G}(e^{j\varphi h})D_{11}^*$ with $\widehat{z}_0$ being the corresponding eigenvector. It follows that $\|\widehat{G}(e^{j\varphi h})D_{11}^*\| \geq \gamma^2$, and thus we have

$$\|\widehat{G}(e^{j\varphi h})\| \cdot d_{11} \geq \gamma^2. \tag{62}$$

Since $\gamma > d_{11}$ by assumption, we have

$$\|\widehat{G}(e^{j\varphi h})\| > \gamma. \tag{63}$$

Namely, we have established that such γ that makes $e^{j\varphi h}I - \mathcal{A}_\gamma$ noninvertible is a (strict) lower bound of the frequency response gain $\|\widehat{G}(e^{j\varphi h})\|$.

Furthermore, the above arguments, in particular, lead to the following result.

LEMMA 4.1. *Let Γ be the set of $\gamma > d_{11}$ such that $e^{j\varphi h}I - \mathcal{A}_\gamma$ is noninvertible (for a fixed φ). Then every point in Γ is an isolated point of Γ.*

Proof. The above arguments imply that for every $\gamma \in \Gamma$, γ^2 is an eigenvalue of $\widehat{G}(e^{j\varphi h})D_{11}^*$. Hence the assertion follows readily from Lemma 2.3. □

4.2. Continuity study. The fact shown in the preceding subsection is in some sense enough to answer the first question raised in the preceding section; with the knowledge that γ is a lower bound, a bisection method can be continued without any problem. However, the following further study will be of more interest and value; the aim of this study is to show that we do not need to check if $e^{j\varphi h}I - \mathcal{A}_\gamma$ is invertible after all.

In the following, we assume that γ is such a value that makes $e^{j\varphi h}I - \mathcal{A}_\gamma$ noninvertible. Then, by Lemma 4.1, there exists an open interval $\mathcal{I}$ on the positive real axis containing γ such that $\mathcal{I}$ contains no point in Γ other than γ. Note that we can take $\mathcal{I}$ so that $\|\widehat{G}(e^{j\varphi h})\| > \widetilde{\gamma}$ whenever $\widetilde{\gamma} \in \mathcal{I}$ because of the strict inequality (63). Now, for each $\widetilde{\gamma} \in \mathcal{I} \setminus \{\gamma\}$, the matrix $\mathcal{A}_{\widetilde{\gamma}}$ is invertible by the construction of $\mathcal{I}$. Now let us further assume that γ is not a singular value of $\mathcal{D}$. Note that this assumption is almost always satisfied (Proposition 2.5). Furthermore, as we show in the following section (Proposition 5.1), this assumption is equivalent to the assumption that γ is not a singular value of $\mathcal{D}_\gamma$. Hence, by continuity together with compactness of $\mathcal{D}_\gamma$, we may assume, by taking the interval $\mathcal{I}$ sufficiently small, that the number of the singular values of $\mathcal{D}_{\widetilde{\gamma}}$ larger than $\widetilde{\gamma}$ is constant over $\mathcal{I}$, which we denote by ν. Note that this assumption, in particular, implies that $\widetilde{\gamma}$ is not a singular value of $\mathcal{D}_{\widetilde{\gamma}}$ (for all $\widetilde{\gamma} \in \mathcal{I}$). From all of the considerations above, we can apply Theorem 3.2 for $\widetilde{\gamma}$ to get

$$\mathrm{N}(F_{\widetilde{\gamma}}(e^{j\varphi h}, \widetilde{\gamma})) \neq (0, n-\nu) \quad \forall\, \widetilde{\gamma} \in \mathcal{I} \setminus \{\gamma\}. \tag{64}$$

Now, from the definition of the matrix $F_{\widetilde{\gamma}}(z, \widetilde{\gamma})$ shown in Theorem 3.2 and from the fact that $\widetilde{\gamma}$ is not a singular value of $\mathcal{D}_{\widetilde{\gamma}}$ (for all $\widetilde{\gamma} \in \mathcal{I}$), it follows that the matrix $F_{\widetilde{\gamma}}(e^{j\varphi h}, \widetilde{\gamma})$ has continuous entries over $\mathcal{I}$.

Here let us suppose that $F_\gamma(e^{j\varphi h}, \gamma)$ is invertible because otherwise we would obviously have that

$$\mathrm{N}(F_\gamma(e^{j\varphi h}, \gamma)) \neq (0, n-\nu). \tag{65}$$

Then, by continuity, there exists some open interval $\mathcal{I}_1 (\subset \mathcal{I})$ containing γ such that $F_{\widetilde{\gamma}}(e^{j\varphi h}, \widetilde{\gamma})$ is invertible for every $\widetilde{\gamma} \in \mathcal{I}_1$. Moreover, the number of the negative eigenvalues of $F_{\widetilde{\gamma}}(e^{j\varphi h}, \widetilde{\gamma})$ must be constant over $\mathcal{I}_1$, again by continuity. Since this constant number is not equal to $n - \nu$ by (64), we are led to the same conclusion as (65) after all. Hence we can conclude that (65) holds whenever $e^{j\varphi h}I - \mathcal{A}_\gamma$ is noninvertible as long as γ is not a singular value of $\mathcal{D}_\gamma$.

Thus we arrive at the following theorem, as claimed.

THEOREM 4.2. *Suppose that $\gamma > d_{11}$ and γ is not a singular value of $\mathcal{D}_\gamma$. Let ν be the number of the singular values of $\mathcal{D}_\gamma$ larger than γ. Then conditions* (i) *and* (iii) *of Theorem* 3.2 *are equivalent.*

Proof. Given Theorem 3.2, it is enough to consider the case in which $e^{j\varphi h}I - \mathcal{A}_\gamma$ is noninvertible. In this case, however, condition (iii) of Theorem 3.2 never holds as

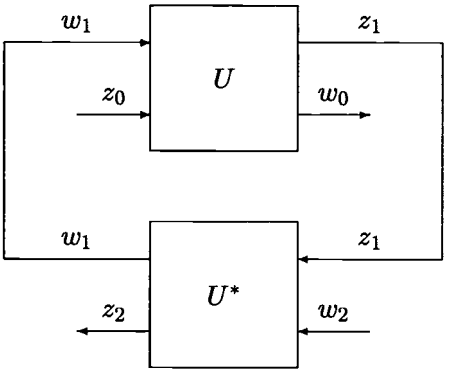


FIG. 4. *Identity system.*

we showed above. Hence it is enough to show that condition (i) never holds either. However, this follows readily from (63). □

5. The relationship between the singular values of $\mathcal{D}$ and $\mathcal{D}_\gamma$. The purpose of this section is to study the second question raised in subsection 3.2. To this end, the following proposition plays a key role, and thus most of this section is devoted to its proof. Note that, unlike in the preceding section, we have no grounds to introduce a nonzero initial state in the arguments of this section, and thus we indeed avoid doing so.

PROPOSITION 5.1. *Suppose that $\gamma > d_{11}$. Then γ is a singular value of $\mathcal{D}_\gamma$ with multiplicity m if and only if it is a singular value of $\mathcal{D}$ with the same multiplicity.*

Since each $\gamma > d_{11}$ in the set of the singular values of $\mathcal{D}$ is an isolated point of this set by Proposition 2.5, the importance of Proposition 5.1 lies in guaranteeing that the second question raised in section 3 can always be circumvented by a slight perturbation of γ. Note that this proposition also played some role in the preceding section, where the first question raised in section 3 was studied.

To prove the above proposition, the following lemma is quite useful.

LEMMA 5.2. *Let $\gamma > d_{11}$, and consider the system shown in Figure* 4, *where U is an operator on $\mathcal{K}$ corresponding to the matrix U given by*

$$U = \begin{bmatrix} -\gamma^{-1}D_{11} & (I-\gamma^{-2}D_{11}D_{11}^T)^{1/2} \\ (I-\gamma^{-2}D_{11}^TD_{11})^{1/2} & \gamma^{-1}D_{11}^T \end{bmatrix}. \tag{66}$$

Then we have $z_0 = z_2$ and $w_0 = w_2$.

Proof. The above matrix U is a unitary matrix with its upper-right block being invertible. Hence, as is well known, e.g., in the circuit theory, we have $z_0(t) = z_2(t)$ and $w_0(t) = w_2(t)$ for each t. Thus the assertion follows readily. □

Given the above lemma, Proposition 5.1 can be proved as follows.

Proof of Proposition 5.1. Comparing the above matrix U with U_γ given in (30), we can see that

$$U = \begin{bmatrix} \gamma^{-1}I & 0 \\ 0 & I \end{bmatrix} U_\gamma \begin{bmatrix} I & 0 \\ 0 & \gamma I \end{bmatrix}. \tag{67}$$

On the other hand, from Figure 4, we have

$$\begin{bmatrix} z_1 \\ w_0 \end{bmatrix} = U \begin{bmatrix} w_1 \\ z_0 \end{bmatrix}. \tag{68}$$

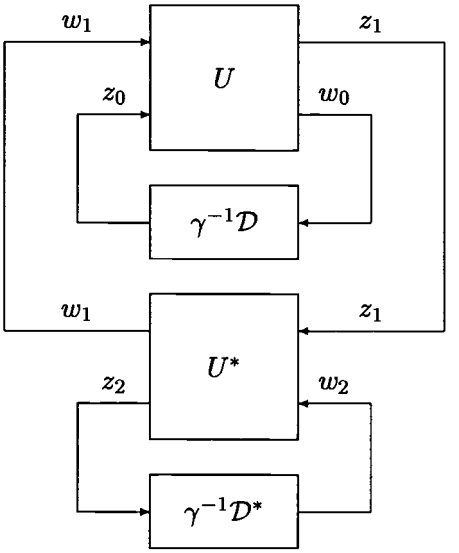


FIG. 5. *System* Σ_{c1}.

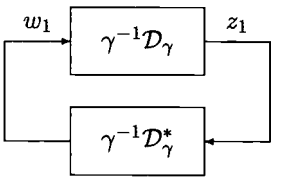


FIG. 6. *System* Σ_{c2}.

Hence, from (67), we obtain

$$\begin{bmatrix} \gamma z_1 \\ w_0 \end{bmatrix} = U_\gamma \begin{bmatrix} w_1 \\ \gamma z_0 \end{bmatrix}. \tag{69}$$

Here let us consider closing the loop between w_0 and z_0 by $z_0 = \gamma^{-1}\mathcal{D}w_0$ as in the system Σ_{c1} shown in Figure 5, and let us rewrite w_0 as w. Then we have $\gamma z_0 = z$, where $z = \mathcal{D}w$ is the output of $\mathcal{D}$ to the input w. Hence (69) can be rewritten as

$$\begin{bmatrix} \gamma z_1 \\ w \end{bmatrix} = U_\gamma \begin{bmatrix} w_1 \\ z \end{bmatrix}. \tag{70}$$

Comparing this equation with (29), we can see that w_1 may be identified with w_J, while z_1 may be identified with γ^{-1} times z_J. Since the J-unitary transformation associated with U_γ transforms $\mathcal{D}$ into $\mathcal{D}_\gamma$, we can see that the map from w_1 to z_1 in Figure 5 is nothing but $\gamma^{-1}\mathcal{D}_\gamma$. Similarly, the map from z_1 to w_1 in the same figure is nothing but $\gamma^{-1}\mathcal{D}_\gamma^*$. Thus we can see that the signals w_1 and z_1 in the system Σ_{c1} shown in Figure 5 are equivalent to those in the system Σ_{c2} shown in Figure 6. On the other hand, the signals w_0 and z_0 in the system Σ_{c1} are equivalent to those in the system Σ_{c3} shown in Figure 7 by Lemma 5.2. Now, since (68) can be rearranged as

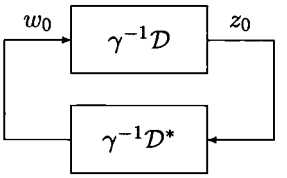


FIG. 7. *System* Σ_{c3}.

$$\begin{bmatrix} z_0 \\ w_0 \end{bmatrix} = S \begin{bmatrix} z_1 \\ w_1 \end{bmatrix}, \tag{71}$$

where S is given by

$$S = \begin{bmatrix} \gamma(\gamma^2 I - D_{11}D_{11}^T)^{-1/2} & (\gamma^2 I - D_{11}D_{11}^T)^{-1/2}D_{11} \\ D_{11}^T(\gamma^2 I - D_{11}D_{11}^T)^{-1/2} & \gamma(\gamma^2 I - D_{11}^T D_{11})^{-1/2} \end{bmatrix} \tag{72}$$

and is invertible, we can see that Σ_{c1} has m pairs of nontrivial solutions (w_0, z_0) if and only if it has m pairs of nontrivial solutions (w_1, z_1). Combining the above arguments, we are led to the conclusion that Σ_{c2} has m nontrivial solutions (w_1 and/or z_1) if and only if Σ_{c3} has the same number of nontrivial solutions (w_0 and/or z_0). Obviously, this implies that γ^2 is an eigenvalue of $\mathcal{D}_\gamma^*\mathcal{D}_\gamma$ with multiplicity m if and only if it is an eigenvalue of $\mathcal{D}^*\mathcal{D}$ with the same multiplicity, which completes the proof of Proposition 5.1. □

COROLLARY 5.3. *Let* $\gamma > d_{11}$. *Then* $\mathcal{D}_\gamma$ *has* ν *singular values larger than* γ *if and only if* $\mathcal{D}$ *has* ν *singular values larger than* γ.

Proof. The assertion follows readily from Proposition 5.1 if we note that the system $(C_{1\gamma}, A_\gamma, B_{1\gamma})$ tends to the γ-independent system (C_1, A, B_1) as $\gamma \to \infty$ and that the singular values of the compact operator $\mathcal{D}_\gamma$ are continuous with respect to γ [26, p. 57]. □

The above result is useful for computing the singular values of the noncompact compression operator $\mathcal{D}$, which will be discussed at the end of the following section. Furthermore, observe that, by Proposition 5.1 and Corollary 5.3, the statement of Theorem 4.2 can be restated further in the following "more natural" form.

THEOREM 5.4. *Suppose that* $\gamma > d_{11}$ *and* γ *is not a singular value of* $\mathcal{D}$. *Let* ν *be the number of the singular values of* $\mathcal{D}$ *larger than* γ. *Then conditions* (i) *and* (iii) *of Theorem* 3.2 *are equivalent.*

This theorem shows that we can check if condition (i) is true only by the finite-dimensional computations regarding condition (iii), where the only difficulty is the computation of the number ν. As will be discussed at the end of the following subsection, however, this can also be carried out with only finite-dimensional computations. Thus the problem of establishing a bisection method for the frequency response gain computation for the noncompact case has now been resolved completely.

6. Computations of the singular values of the frequency response and compression operators. In this section, we study a related topic on the singular values of the frequency response operator $\widehat{G}(e^{j\varphi h})$ and the compression operator $\mathcal{D}$. First, the following result is a direct consequence of Lemma 3.1.

COROLLARY 6.1. *Suppose that $\gamma > d_{11}$ and $e^{j\varphi h}I - \mathcal{A}_\gamma$ is invertible. Then $\widehat{G}_\gamma(e^{j\varphi h})$ has ν singular values larger than γ if and only if $\widehat{G}(e^{j\varphi h})$ has ν singular values larger than γ.*

This, together with Lemma 4.1 and a continuity argument as in the proof of Corollary 5.3, leads to the following result.

PROPOSITION 6.2. *Suppose that $\gamma > d_{11}$ and $e^{j\varphi h}I - \mathcal{A}_\gamma$ is invertible. Then γ is a singular value of $\widehat{G}_\gamma(e^{j\varphi h})$ with multiplicity m if and only if it is a singular value of $\widehat{G}(e^{j\varphi h})$ with the same multiplicity.*

Observe that the order of the discussions here is the opposite of that in the preceding section in that Corollary 5.3 was derived from Proposition 5.1 in the preceding section, while Proposition 6.2 was derived from Corollary 6.1 in this subsection. This is because, in the preceding section, we had to avoid the introduction of a nonzero initial state (and hence such a type of argument similar to Lemma 3.1 was not possible), and thus we had to employ a quite different approach there, even though these two problems are apparently quite similar.

Now, from the study in the case of compact operators [13, 22], it is known that, as long as γ is not a singular value of $\mathcal{D}_\gamma$ (or, equivalently, $\mathcal{D}$), the matrix $F_\gamma(e^{j\varphi h}, \gamma)$ has m repeated eigenvalues at 0 if and only if γ is a singular value of $\widehat{G}_\gamma(e^{j\varphi h})$ with multiplicity m. Hence, by Proposition 6.2 (or Corollary 6.1), we are led to the following result, which can be regarded as a generalization of the state-space formula given in [13] to the noncompact case (see also [27]). Note that the invertibility assumption has been removed in the following result, which can be validated by a continuity argument supported by Lemma 4.1.

COROLLARY 6.3. *Suppose that $\gamma > d_{11}$ and that γ is not a singular value of $\mathcal{D}$. Then γ is a singular value of the frequency response operator $\widehat{G}(e^{j\varphi h})$ with multiplicity m if and only if the finite-dimensional matrix $F_\gamma(e^{j\varphi h}, \gamma)$ has m repeated eigenvalues at 0.*

By definition, all of the singular values of $\widehat{G}(e^{j\varphi h})$ are no smaller than d_{11}, with d_{11} being the smallest singular value with infinite multiplicity if there are only a finite number of singular values strictly larger than d_{11} [24]. This, together with Corollary 6.3, implies that we can compute, in principle, all of the singular values of $\widehat{G}(e^{j\varphi h})$, including their multiplicities. A bisection method for their computation based on Corollary 6.1 is given by the following theorem, which is a generalization of Theorem 5.4.

THEOREM 6.4. *Suppose that $\gamma > d_{11}$ and γ is not a singular value of $\mathcal{D}$. Let ν be the number of the singular values of $\mathcal{D}$ larger than γ. Then the following two conditions are equivalent for any nonnegative integer i, where we define $s_0(\cdot) = +\infty$:*

$$\text{(i)} \quad s_{i+1}(\widehat{G}(e^{j\varphi h})) < \gamma < s_i(\widehat{G}(e^{j\varphi h})), \tag{73}$$

$$\text{(ii)} \quad N(F_\gamma(e^{j\varphi h}, \gamma)) = (0, n + i - \nu). \tag{74}$$

The above theorem follows readily by applying entirely the same arguments as in [22]. The only obstacle in the extension is that, when $e^{j\varphi h}I - \mathcal{A}_\gamma$ happens to be noninvertible, we do not necessarily have $s_i(\widehat{G}(e^{j\varphi h})) > \gamma$, unlike in (63). Hence it is not trivial that we can drop the invertibility assumption of $e^{j\varphi h}I - \mathcal{A}_\gamma$ as we did in subsection 4.2. However, if $s_i(\widehat{G}(e^{j\varphi h})) = \gamma$ for some i, then we are led to the conclusion that $F_\gamma(e^{j\varphi h}, \gamma)$ is noninvertible (Corollary 6.3). Namely, condition (ii) of the above theorem fails for any i. Thus the statement of the above theorem leads to the

conclusion that condition (i) fails for any i, as required (because $s_i(\widehat{G}(e^{j\varphi h})) = \gamma$), which validates the statement. If $s_{i+1}(\widehat{G}(e^{j\varphi h})) < \gamma < s_i(\widehat{G}(e^{j\varphi h}))$, on the other hand, we can still develop a continuity argument, as in subsection 4.2, supported by Lemma 4.1, and, again with the aid of Corollary 6.3, we can show that condition (i) holds if and only if condition (ii) holds, even if $e^{j\varphi h}I - \mathcal{A}_\gamma$ is noninvertible. These considerations show that we can actually drop the invertibility assumption of $e^{j\varphi h}I - \mathcal{A}_\gamma$.

If we apply the above theorem to compute the singular values of $\widehat{G}(e^{j\varphi h})$, then, for each given $\gamma > 0$, we have to know the number ν, which is the number of the singular values of the compression operator $\mathcal{D}$ larger than γ. By Corollary 5.3, the number ν is equal to the number of the singular values of the compression operator $\mathcal{D}_\gamma$ larger than γ. Here $\mathcal{D}_\gamma$ is a compact operator, and, for a compact compression operator, there exists a finite-dimensional method to compute the number of its singular values larger than given $\gamma > 0$; see [22] for details. Hence we can compute the number ν for each $\gamma > 0$, and thus we are led to a bisection method for the computation of the singular values of $\widehat{G}(e^{j\varphi h})$ with only finite-dimensional computations. It should be worthwhile emphasizing that the above-mentioned fact that, for each $\gamma > 0$, we can compute the number of the singular values of $\mathcal{D}$ larger than γ, implies that we can actually develop a bisection method for the computation of the singular values (and hence the norm, as well) of the noncompact compression operator $\mathcal{D}$.

7. Conclusion. In this paper, we studied the problem of computing the frequency response gain (to be more precise, the norm and singular values) of general sampled-data systems with a direct feedthrough matrix D_{11}. We first reviewed in section 2 the frequency response operator $\widehat{G}(e^{j\varphi h})$ of such sampled-data systems, and we made a slight but crucial extension on its definition so that nonzero initial states and unstable sampled-data systems can also be dealt with. We also gave some results on the spectra of operators involving or related to $\widehat{G}(e^{j\varphi h})$ for later use. Then, in section 3, we showed a key result, Theorem 3.2, which is a generalization, to the noncompact case, of the result derived in [22]. Although this result is expected to be a theoretical basis for the development of a bisection method for the computation of the frequency response gain $\|\widehat{G}(e^{j\varphi h})\|$ for noncompact $\widehat{G}(e^{j\varphi h})$, we also pointed out two critical issues that could be serious enough to prevent us from arriving at a bisection method (subsection 3.2). Regarding these issues, however, we clarified in section 4 that the first issue (i.e., whether or not $e^{j\varphi h}I - \mathcal{A}_\gamma$ is invertible) is actually irrelevant, and we gave a refined result (Theorem 4.2); we further showed in section 5 that the second issue (regarding the way the singular values of $\mathcal{D}_\gamma$ are dependent on γ) does not either cause a serious problem in the bisection method after all, where the spectral study in section 2 played important roles for these proofs. Hence, as a whole, we have proved up to the slightest details that the bisection method for the computation of the frequency response gain of sampled-data systems with $D_{11} = 0$ (i.e., the compact case) proposed in [22] can be extended to the case of nonzero D_{11} (i.e., the noncompact case). Finally, in section 6, the result was extended to the computation of the singular values of $\widehat{G}(e^{j\varphi h})$ (Theorem 6.4) as well as those of the compression operator $\mathcal{D}$.

To summarize the results on the singular value computations, we have shown the following: the singular values $s_i(\widehat{G}(e^{j\varphi h}))$ are no smaller than $d_{11}(= \|D_{11}\|)$, and given $\gamma > d_{11}$, the condition

$$s_{i+1}(\widehat{G}(e^{j\varphi h})) < \gamma < s_i(\widehat{G}(e^{j\varphi h})) \tag{75}$$

is equivalent to the condition (in terms of the finite-dimensional matrix given in (50))

$$\mathrm{N}(F_\gamma(e^{j\varphi h},\gamma)) = (0, n+i-\nu), \tag{76}$$

with the only exceptions being for such a countable number of γ that is equal to a singular value of $\mathcal{D}$ (for which the matrix $F_\gamma(z,\gamma)$ in (76) cannot be defined), where ν is the number of the singular values of $\mathcal{D}$ larger than γ. Here, once γ is fixed, we can obtain the number ν by computing the eigenvalues of a single finite-dimensional Hermitian matrix, as long as γ is not a singular value of $\mathcal{D}$, and we will never encounter any sort of critical problem in that process, as shown in [22]. Therefore, the condition (75) can be tested with only finite-dimensional computations, with a possible need for a slight perturbation on γ to avoid the singular values of $\mathcal{D}$. This establishes a complete bisection method for the computation of $\|\widehat{G}(e^{j\varphi h})\|$ and $s_i(\widehat{G}(e^{j\varphi h}))$ even in the case of $D_{11} \neq 0$.

Acknowledgments. The author is grateful to Y. Ito for helpful discussions on this topic. He is also grateful to M. Sawada for his numerical studies supporting the results of this paper.

REFERENCES

[1] Y. Yamamoto, *A function space approach to sampled-data systems and tracking problems*, IEEE Trans. Automat. Control, 39 (1994), pp. 703–713.

[2] B. A. Bamieh and J. B. Pearson, *A general framework for linear periodic systems with applications to H_∞ sampled-data control*, IEEE Trans. Automat. Control, 37 (1992), pp. 418–435.

[3] H. T. Toivonen, *Sampled-data control of continuous-time systems with an H_∞ optimality criterion*, Automatica J. IFAC, 28 (1992), pp. 45–54.

[4] P. T. Kabamba and S. Hara, *Worst-case analysis and design of sampled-data control systems*, IEEE Trans. Automat. Control, 38 (1993), pp. 1337–1357.

[5] Y. Hayakawa, Y. Yamamoto, and S. Hara, *H_∞ type problem for sampled-data control systems—a solution via minimum energy characterization*, IEEE Trans. Automat. Control, 39 (1994), pp. 2278–2284.

[6] T. Chen and B. Francis, *Optimal Sampled-Data Control Systems*, Springer-Verlag, New York, 1995.

[7] P. P. Khargonekar and N. Sivashankar, *H_2 optimal control for sampled-data systems*, Systems Control Lett., 17 (1991), pp. 425–436.

[8] B. Bamieh and J. B. Pearson, *The H_2 problem for sampled-data systems*, Systems Control Lett., 19 (1992), pp. 1–12.

[9] S. Hara, H. Fujioka, and P. T. Kabamba, *A hybrid state-space approach to sampled-data feedback control*, Linear Algebra Appl., 205–206 (1994), pp. 675–712.

[10] T. Hagiwara and M. Araki, *FR-operator approach to the H_2 analysis and synthesis of sampled-data systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 1411–1421.

[11] G. Dullerud and K. Glover, *Robust stabilization of sampled-data systems to structured LTI perturbations*, IEEE Trans. Automat. Control, 38 (1993), pp. 1497–1508.

[12] T. Hagiwara and M. Araki, *Robust stability of sampled-data systems under possibly unstable additive/multiplicative perturbations*, IEEE Trans. Automat. Control, 43 (1998), pp. 1340–1346.

[13] Y. Yamamoto and P. P. Khargonekar, *Frequency response of sampled-data systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 166–176.

[14] M. Araki, Y. Ito, and T. Hagiwara, *Frequency-response of sampled-data systems*, Automatica J. IFAC, 32 (1996), pp. 483–497.

[15] J. S. Freudenberg, R. H. Middleton, and J. H. Braslavsky, *Inherent design limitations for linear sampled-data feedback systems*, Internat. J. Control, 61 (1995), pp. 1387–1421.

[16] G. C. Goodwin and M. Salgado, *Frequency domain sensitivity functions for continuous time systems under sampled data control*, Automatica J. IFAC, 30 (1994), pp. 1263–1270.

[17] T. Hagiwara, Y. Ito, and M. Araki, *Computation of the frequency response gains and H_∞-norm of a sampled-data system*, Systems Control Lett., 25 (1995), pp. 281–288.

[18] S. Hara, H. Fujioka, P. P. Khargonekar, and Y. Yamamoto, *Computational aspects of gain-frequency response for sampled-data systems*, in Proceedings of the 34th IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 1995, pp. 1784–1789.

[19] J. H. Braslavsky, R. H. Middleton, and J. S. Freudenberg, *L_2-induced norms and frequency-gains of sampled-data sensitivity operators*, IEEE Trans. Automat. Control, 43 (1998), pp. 252–258.

[20] Y. Yamamoto, A. G. Madievski, and B. D. O. Anderson, *Approximation of frequency response for sampled-data control systems*, Automatica J. IFAC, 35 (1999), pp. 729–734.

[21] T. Hagiwara, M. Suyama, and M. Araki, *Upper and lower bounds of the frequency response gain of sampled-data systems*, Automatica J. IFAC, 37 (2001), pp. 1363–1370.

[22] Y. Ito, T. Hagiwara, H. Maeda, and M. Araki, *Bisection algorithm for computing the frequency response gain of sampled-data systems*, IEEE Trans. Automat. Control, 46 (2001), pp. 369–381.

[23] J. B. Conway, *A Course in Functional Analysis*, 2nd ed., Springer-Verlag, New York, 1990.

[24] I. Gohberg, S. Goldberg, and M. A. Kaashoek, *Classes of Linear Operators, Vol.* I, Birkhäuser-Verlag, Basel, 1990.

[25] N. Young, *An Introduction to Hilbert Space*, Cambridge University Press, Cambridge, UK, 1988.

[26] P. R. Halmos, *A Hilbert Space Problem Book*, 2nd ed., Springer-Verlag, New York, 1982.

[27] K. Sugimoto and M. Suzuki, *On γ-positive real sampled-data control systems and their phase property*, Transactions of the Society of Instrument and Control Engineers, 35 (1999), pp. 71–76 (in Japanese).

SIAM J. CONTROL OPTIM.
Vol. 41, No. 5, pp. 1372–1388

ON THE λ-EQUATIONS FOR MATCHING CONTROL LAWS*

DAVID AUCKLY† AND LEV KAPITANSKI†

Abstract. We discuss matching control laws for underactuated systems. We previously showed that this class of matching control laws is completely characterized by a linear system of first order partial differential equations for one set of variables (λ) followed by a linear system of first order partial differential equations for the second set of variables ($\widehat{g}$, $\widehat{V}$). Here we derive a new first order system of partial differential equations that encodes all compatibility conditions for the λ-equations. We give four examples illustrating different features of matching control laws. The last example is a system with two unactuated degrees of freedom that admits only basic solutions to the matching equations. There are systems with many matching control laws where only basic solutions are potentially useful. We introduce a rank condition indicating when this is likely to be the case.

Key words. nonlinear control, matching control laws, λ-equations, stabilization

AMS subject classifications. 93C10, 93D15

PII. S0363012901393304

1. Introduction.

Effective procedures for designing control laws are very important in nonlinear control theory. Explicit analytic formulae for control laws play a role similar to that played by explicit solutions to differential equations. Such formulae exist in only a few special cases, but those that exist serve as simple models to develop and test more general techniques.

In this paper, we discuss a class of full state feedback control laws for underactuated systems. In [5], we showed that this class of matching control laws is completely characterized by a linear system of first order partial differential equations for one set of variables (λ) followed by a linear system of first order partial differential equations for the second set of variables ($\widehat{g}$, $\widehat{V}$). These equations always have a simple family of solutions which we call basic solutions. The system of equations for the first set of variables (λ-equations) is overdetermined. Here we derive a new first order system of partial differential equations that encodes all compatibility conditions for the λ-equations. (We call these the ν-equations.) If only one degree of freedom is unactuated, the solutions to all of these systems of partial differential equations can be completely analyzed. It is often possible to get explicit formulae for the solutions to these equations. We also provide an example of a system with two unactuated degrees of freedom that has only basic solutions. There are systems with many matching control laws where only basic solutions are potentially useful. We write down a rank condition indicating when this is likely to be the case (Remark 5.1).

During the last few years, several researchers have investigated control laws in which the closed loop system assumes a certain structure. Numerous papers have been written on this subject; see [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13] and the references therein. The control laws that form the subject of this present paper are described by (2.4) and (2.6). These equations were independently derived in [10] and [5]. The λ-equations were first introduced in [5]. Even though the initial

*Received by the editors August 3, 2001; accepted for publication (in revised form) April 23, 2002; published electronically December 11, 2002. This work was partially supported by grants CMS-9813182 and DMS-9970638 from the National Science Foundation.
http://www.siam.org/journals/sicon/41-5/39330.html
†Department of Mathematics, Kansas State University, Manhattan, KS 66506 (dav@math.ksu.edu, levkapit@math.ksu.edu).

matching equations of [10] and [5] form a highly nonlinear system of partial differential equations, introduction of the λ variables triangulates the system. The system is triangulated in the sense that all solutions are obtained by first solving first order linear equations for λ and then solving first order linear equations for the remaining variables.

This paper is organized as follows. Section 2 reviews matching control laws and the λ-equations and introduces the ν-equations. Section 3 specializes to systems with one unactuated degree of freedom. Sections 4, 5, and 6 contain examples illustrating three different features of matching control laws. The rank condition appears at the end of section 5. Later we apply it in section 6. In section 7, we describe the final example of a system with two unactuated degrees of freedom. We show that this system has only basic matching control laws.

2. Matching equations. We use the following notation.

- n is the number of the degrees of freedom of the mechanical system.
- $x = (x^1, \dots, x^n)$ are configuration variables denoting the position of the system, and $\dot{x} = (\dot{x}^1, \dots, \dot{x}^n)$ are the corresponding velocities.
- $g_{ij}(x)$ is the mass matrix.
- $V(x)$ is the potential energy.
- $C_i(x, \dot{x})$ are the dissipation terms.
- $u_i(x, \dot{x})$ are the control inputs.

Let $m \le n$ be the number of unactuated degrees of freedom. We will assume that degrees of freedom numbered 1 through m are unactuated and use indices $a, b, \dots$ to indicate unactuated degrees of freedom. The indices $i, j, \dots$ will run from 1 to n. We adopt the convention of summation over the repeated indices.

Given this, the equations of motion of the system are

$$g_{rj}\ddot{x}^j + [j\,k,\,r]\,\dot{x}^j\,\dot{x}^k + C_r + \frac{\partial V}{\partial x^r} = u_r\,, \quad r = 1, \dots, n, \tag{2.1}$$

where $[i\,j,\,k]$ is the Christoffel symbol of the first kind,

$$[i\,j,\,k] = \frac{1}{2}\left(\frac{\partial g_{jk}}{\partial x^i} + \frac{\partial g_{ik}}{\partial x^j} - \frac{\partial g_{ij}}{\partial x^k}\right). \tag{2.2}$$

Our assumption that the first m degrees of freedom are not actuated means that

$$u_1 = \cdots = u_m = 0. \tag{2.3}$$

We are looking for control laws u_i such that the closed loop system can be written in the form

$$\widehat{g}_{rj}\ddot{x}^j + \widehat{[j\,k,\,r]}\,\dot{x}^j\,\dot{x}^k + \widehat{C}_r + \frac{\partial \widehat{V}}{\partial x^r} = 0, \quad r = 1, \dots, n,$$

where $\widehat{[i\,j,\,k]}$ is defined as in (2.2) with $\widehat{g}$ in place of g. Such a control law will be given by

$$\begin{aligned} u_\ell = {} & \left([j\,k,\,\ell] - g_{\ell i}\,\widehat{g}^{ir}\,\widehat{[j\,k,\,r]}\right)\dot{x}^j\,\dot{x}^k + \left(C_\ell - g_{\ell i}\,\widehat{g}^{ij}\,\widehat{C}_j\right) \\ & + \left(\frac{\partial V}{\partial x^\ell} - g_{\ell i}\,\widehat{g}^{ij}\,\frac{\partial \widehat{V}}{\partial x^j}\right), \qquad \ell = 1, \dots, n. \end{aligned} \tag{2.4}$$

Condition (2.3) translates into

$$\big([j\,k,\,a] \,-\, g_{ai}\,\widehat{g}^{ir}\,\widehat{[j\,k,\,r]}\,\big)\,\dot{x}^j\,\dot{x}^k \,+\, \big(\,C_a \,-\, g_{ai}\,\widehat{g}^{ij}\,\widehat{C}_j\,\big)$$

$$+\left(\frac{\partial V}{\partial x^a} \,-\, g_{ai}\,\widehat{g}^{ij}\,\frac{\partial \widehat{V}}{\partial x^j}\right) = 0, \qquad a = 1,\dots,m. \tag{2.5}$$

In order to satisfy these equations, it is sufficient to have

$$\begin{aligned} g_{ai}\,\widehat{g}^{ir}\,\widehat{[j\,k,\,r]} \;&=\; [j\,k,\,a], \\ g_{ai}\,\widehat{g}^{ir}\,\widehat{C}_r \;&=\; C_a, \\ g_{ai}\,\widehat{g}^{ir}\,\frac{\partial \widehat{V}}{\partial x^r} \;&=\; \frac{\partial V}{\partial x^a}. \end{aligned} \tag{2.6}$$

These are the *matching equations*; see [5, 10]. Following [5], introduce variables λ_a^j relating the unknown mass matrix $\widehat{g}$ to the original mass matrix g,

$$\lambda_a^r \;=\; g_{ai}\,\widehat{g}^{ir}. \tag{2.7}$$

Using λ_a^j, the matching equations take the form

$$\begin{aligned} \lambda_a^r\,\widehat{[j\,k,\,r]} \;&=\; [j\,k,\,a], \\ \lambda_a^j\,\widehat{C}_j \;&=\; C_a, \\ \lambda_a^j\,\frac{\partial \widehat{V}}{\partial x^j} \;&=\; \frac{\partial V}{\partial x^a}. \end{aligned} \tag{2.8}$$

THEOREM 2.1. *The following equations are equivalent to the matching equations* (2.7), (2.8) *in a neighborhood of a point* x_0.

λ-equations.

$$\frac{\partial}{\partial x^k}\,(\,g_{ai}\,\lambda_b^i) \,-\, [k\,a,\,i]\,\lambda_b^i \,-\, [k\,b,\,i]\,\lambda_a^i \,=\, 0, \qquad \begin{matrix} k = 1,\dots,n, \\ a,b \,=\, 1,\dots,m. \end{matrix} \tag{2.9}$$

$\widehat{g}$-equations.

$$\lambda_a^\ell\,\frac{\partial \widehat{g}_{ij}}{\partial x^\ell} \,+\, \frac{\partial \lambda_a^\ell}{\partial x^i}\cdot\widehat{g}_{\ell j} \,+\, \frac{\partial \lambda_a^\ell}{\partial x^j}\cdot\widehat{g}_{\ell i} \,=\, \frac{\partial g_{ij}}{\partial x^a}\,, \qquad \begin{matrix} a = 1,\dots,m, \\ i,j = 1,\dots,n. \end{matrix} \tag{2.10}$$

$\widehat{V}$-equations.

$$\lambda_a^j\,\frac{\partial \widehat{V}}{\partial x^j} \;=\; \frac{\partial V}{\partial x^a}. \tag{2.11}$$

$\widehat{C}$-equations.

$$\lambda_a^j\,\widehat{C}_j \;=\; C_a. \tag{2.12}$$

Initial compatibility. There exists a hypersurface S containing x_0, transverse to each of the vector-fields $\lambda_a^\ell\frac{\partial}{\partial x^\ell}$, $a = 1,\dots,m$, and on which $\widehat{g}_{ij}$ is invertible and symmetric ($\widehat{g}_{ij} = \widehat{g}_{ji}$), and $g_{ai} = \lambda_a^j\widehat{g}_{ji}$.

That (2.7) and (2.8) imply (2.9), (2.10), (2.11), and (2.12) was originally shown in [5] (for indicial notation, see [1]). In the opposite direction, we show in [4, p. 33] that (2.9), (2.10), and the condition $g_{ai} = \lambda_a^j \widehat{g}_{ji}$ imply the first equation in (2.8). Note that the other two equations in (2.8) are (2.11) and (2.12), and, if $\widehat{g}_{ij}$ is invertible, then (2.7) follows from $g_{ai} = \lambda_a^j \widehat{g}_{ji}$. To complete the proof, we will show now that the *initial compatibility* conditions of the theorem are preserved in some neighborhood of the point x_0 by virtue of (2.9) and (2.10). Indeed, in view of (2.10), each difference $w_{ij} = \widehat{g}_{ij} - \widehat{g}_{ji}$ satisfies m differential equations $\lambda_a^\ell \frac{\partial}{\partial x^\ell} w_{ij} = 0$. Each equation guarantees that $w_{ij}(x) = 0$ in some neighborhood of x_0 provided $w_{ij}(x) = 0$ on the initial hypersurface S. The nondegeneracy of $\widehat{g}$ follows by continuity. Now, denote $\Xi_{ai} = \lambda_a^j \widehat{g}_{ji} - g_{ai}$. Multiplying (2.10) by λ_b^j, summing over j, and rearranging the terms, we obtain

$$\lambda_a^\ell \frac{\partial \Xi_{bi}}{\partial x^\ell} + \frac{\partial \lambda_a^\ell}{\partial x^i} \Xi_{b\ell} + \left[\frac{\partial}{\partial x^i} (\lambda_a^\ell g_{b\ell}) - [ia, \ell]\, \lambda_b^\ell - [ib, \ell]\, \lambda_a^\ell \right] = \left(\lambda_a^\ell \frac{\partial \lambda_b^j}{\partial x^\ell} - \lambda_b^\ell \frac{\partial \lambda_a^j}{\partial x^\ell} \right) \widehat{g}_{ji} - \frac{1}{2} \left\{ \lambda_a^\ell \frac{\partial g_{bi}}{\partial x^\ell} - \lambda_b^\ell \frac{\partial g_{ai}}{\partial x^\ell} - \lambda_a^\ell \frac{\partial g_{b\ell}}{\partial x^i} + \lambda_b^\ell \frac{\partial g_{a\ell}}{\partial x^i} + \lambda_a^\ell \frac{\partial g_{i\ell}}{\partial x^b} - \lambda_b^\ell \frac{\partial g_{i\ell}}{\partial x^a} \right\}.$$

Using (2.9), we simplify this to

$$\lambda_a^\ell \frac{\partial \Xi_{bi}}{\partial x^\ell} + \frac{\partial \lambda_a^\ell}{\partial x^i} \Xi_{b\ell} = \mathcal{R}_{ab},$$

with $\mathcal{R}_{ab}$ antisymmetric in a, b. If $b = a$, this equation reduces to

$$\lambda_a^\ell \frac{\partial \Xi_{ai}}{\partial x^\ell} + \frac{\partial \lambda_a^\ell}{\partial x^i} \Xi_{a\ell} = 0,$$

where *there is no summation over a*. By assumption, $\Xi_{aj}(x) = 0$, $j = 1, \ldots, n$, on the noncharacteristic surface S. Hence $\Xi_{aj}(x) = 0$ in some neighborhood of S. This completes the proof of the theorem.

Remark 2.2. Note that (2.9), (2.10), (2.11), and (2.12) always have a set of solutions of the form

$$\lambda_a^k = \varkappa \delta_a^k, \qquad \widehat{g} = \frac{1}{\varkappa} g + g^o, \qquad \widehat{V} = \frac{1}{\varkappa} V + V^o, \qquad \widehat{C}_j = \frac{1}{\varkappa} C_j$$

with $\varkappa \neq 0$ any constant, V^o an arbitrary function of the variables x^ℓ, $\ell = m+1, \ldots, n$, and g^o any symmetric matrix valued function of the variables x^ℓ such that $g^o_{ia} = 0$. We will call these solutions *basic*.

The λ-equations are a system of $\frac{1}{2} m(m+1) \cdot n$ equations for $n \cdot m$ unknowns. It is not surprising that there are extra compatibility conditions. By viewing system (2.9) in the correct way, we are able to write down the compatibility conditions. Denote

$$\nu_{ab} = g_{ai}\, \lambda_b^i. \tag{2.13}$$

Because the matrix g_{ij} is assumed to be nondegenerate, the matrix comprised of its m first rows has rank m. This implies that m^2 out of $m \cdot n$ λ's can be expressed as linear combinations of ν's; i.e.,

$$\lambda_b^\beta = h^{\beta a}\, \nu_{ab}, \quad \beta \in \mathcal{I}_m, \tag{2.14}$$

where $\mathcal{I}_m$ is some m element subset of $\{1,\dots,n\}$. The indices $\mathcal{I}_m$ are chosen so that the matrix $(g_{a\sigma})$ with $a=1,\dots,m$ and $\sigma\in\mathcal{I}_m$ is nondegenerate; and we denote by $(h^{\beta a})$ the inverse of $(g_{a\sigma})$. Substituting (2.14) in the λ-equations, we obtain

$$\partial_k\,\nu_{ab} \;-\; [a\,k,\,\beta]\,h^{\beta d}\,\nu_{db} \;-\; [b\,k,\,\beta]\,h^{\beta d}\,\nu_{da} \;=\; [a\,k,\,\rho]\,\lambda^{\rho}_{b} \;+\; [b\,k,\,\rho]\,\lambda^{\rho}_{a}, \tag{2.15}$$

where index ρ varies over the remaining $(n-m)$ indices $\{1,\dots,n\}\setminus\mathcal{I}_m$. We will view system (2.15) of $\frac{1}{2}\,m(m+1)\cdot n$ equations as a linear algebraic system for the $m\,(n-m)$ variables λ^{ρ}_{a},

$$\bigl[A_{(k,a,b)}\bigr]^{c}_{\rho}\;\lambda^{\rho}_{c} \;=\; F_{(k,a,b)}. \tag{2.16}$$

Here

$$\bigl[A_{(k,a,b)}\bigr]^{c}_{\rho} \;=\; \delta^{cb}\,[ak,\rho]+\delta^{ca}\,[bk,\rho],$$

and

$$F_{(k,a,b)} \;=\; \partial_k\,\nu_{ab} \;-\; [a\,k,\,\beta]\,h^{\beta d}\,\nu_{db} \;-\; [b\,k,\,\beta]\,h^{\beta d}\,\nu_{da}$$

with δ^{cb} being the Kronecker delta, as usual. Define a linear map A from the vector space of $m\times(n-m)$ two-dimensional arrays into the vector space of $n\times m\times m$ three-dimensional arrays that are symmetric in the last two indices by the coordinate expression

$$[\,A\,\eta\,]_{(k,a,b)} \;=\; \bigl[A_{(k,a,b)}\bigr]^{c}_{\rho}\;\eta^{\rho}_{c}.$$

We know that system (2.16) has at least one solution by Remark 2.2. Thus the rank of the linear map A is at most $m\cdot n$. In order for system (2.16) to have a solution, the vector (three-dimensional array) $F_{(k,a,b)}$ must be perpendicular to the kernel of the adjoint map $A^{\star}$,

$$F \;\perp\; \ker\,A^{\star}. \tag{2.17}$$

Let the kernel of the map $A^{\star}$ be generated by the vectors (three-dimensional arrays) ξ_r. The orthogonality condition (2.17) then takes the form of the following system of linear first order partial differential equations for ν_{ab}:

$$\xi_r^{(k,a,b)}(x)\;\bigl[\,\partial_k\,\nu_{ab} \;-\; [a\,k,\,\beta]\,h^{\beta d}\,\nu_{db} \;-\; [b\,k,\,\beta]\,h^{\beta d}\,\nu_{da}\,\bigr] \;=\; 0. \tag{2.18}$$

THEOREM 2.3. *The general solution to the λ-equations is given by any set of λ^{ρ}_{a} solving the algebraic system* (2.16), *and $\lambda^{\beta}_{b} = h^{\beta a}\,\nu_{ab}$, where ν_{ab} is any solution to* (2.18).

In general, if $m>1$, system (2.17) may be quite complicated, and we do not have a satisfactory description of its solutions.

3. Systems with one unactuated degree of freedom. If only one degree of freedom is unactuated, we do have a reasonable description of all solutions to system (2.18). Assume, for simplicity, that $g_{11}(x)>0$. Then, after rescaling x^1 if necessary, we will have $g_{11}(x)=1$. More precisely, from the very beginning, we could use, instead of $(x^1,x^2,\dots,x^n)$, the coordinates $(z^1,z^2,\dots,z^n)$ which are related to x as follows:

$$\frac{\partial z^1}{\partial x^1} \;=\; \sqrt{g_{11}(x)}, \qquad z^2 \;=\; x^2,\dots,\; z^n \;=\; x^n.$$

In z coordinates, the mass matrix is $\tilde{g}_{ij}(z) = g_{k\ell}(x)\,\frac{\partial x^k}{\partial z^i}\,\frac{\partial x^\ell}{\partial z^j}$, and hence $\tilde{g}_{11}(z) = 1$. On the other hand, the structure of the equations of motion (2.1) does not change because of their tensorial form, and the condition $u_1 = 0$ remains the same again because $\tilde{u}_1 = u_k\,\frac{\partial x^k}{\partial z^1} = u_1\,\sqrt{g_{11}(x)}$. Thus we assume that the coordinates are chosen appropriately, and $g_{11}(x) = 1$.

In the case of one unactuated degree of freedom, one is solving for λ_1^i. The λ-equation reads

$$\frac{\partial \nu}{\partial x^k} = 2\,[k\,1,\,i]\,\lambda_1^i, \tag{3.1}$$

where $\nu = g_{1i}\,\lambda_1^i$. Notice that $[k\,1,\,1] = 0$. View (3.1) as a system of linear algebraic equations for the variables λ_1^ρ, $\rho = 2, \dots, n$. In order for this system of n equations in $(n-1)$ unknowns to have a solution, the vector

$$v = \begin{pmatrix} \partial_1\,\nu \\ \dots \\ \partial_n \nu \end{pmatrix}$$

must be perpendicular to the kernel of the matrix

$$A^\star = \begin{pmatrix} [1\,1,\,2] & \dots & [n\,1,\,2] \\ \dots & \dots & \dots \\ [1\,1,\,n] & \dots & [n\,1,\,n] \end{pmatrix}.$$

Let the kernel of $A^\star$ be generated by the vectors $\xi_r = (\xi_r^1, \dots, \xi_r^n)$. The orthogonality condition for v translates into the system of equations

$$X_r(\nu) \equiv \xi_r^1(x)\,\frac{\partial \nu}{\partial x^1} + \dots + \xi_r^n(x)\,\frac{\partial \nu}{\partial x^n} = 0. \tag{3.2}$$

The standard procedure to solve such a system of equations is to complete the system into an involutive system by adding equations $[X_r,\,X_s](\nu) = 0$, $[[X_r,\,X_s],\,X_t](\nu) = 0$, $\dots$, where $[\eta^i\,\partial_i,\,\zeta^j\,\partial_j] = \big(\eta^i \partial_i(\zeta^k) - \zeta^i\,\partial_i(\eta^k)\big)\,\partial_k$ is the commutator of vector-fields. Recall that a system of equations

$$Y_1(\nu) = 0, \dots, Y_K(\nu) = 0$$

is involutive if $[Y_p,\,Y_q] = f_{pq}^r(x)\,Y_r$.

Thus we have proved the following result.

THEOREM 3.1. *With one unactuated degree of freedom there is a coordinate system such that the ν-equations,* (2.18), *become a homogeneous linear system of equations for one unknown function. This system,* (3.2), *can be completed into an involutive system.*

This theorem extends the general analysis in [4] of systems with two degrees of freedom, one of which is unactuated. Equation (11) of [4] is the analogue of (3.2).

Given any nonzero solution of the λ-equations, there is a local coordinate system $y^1, \dots, y^n$ such that

$$\lambda_1^i(x)\,\frac{\partial y^j}{\partial x^i} = \delta_1^j. \tag{3.3}$$

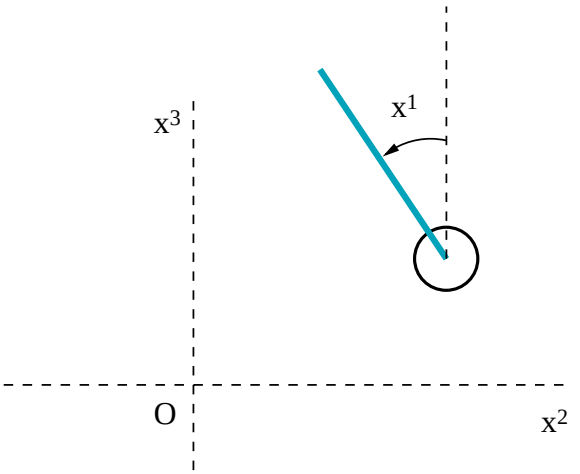


FIG. 4.1.

Let G and $\widehat{G}$ represent g and $\widehat{g}$ in the y-coordinates; i.e., $G_{ij} = g_{kl}\,\frac{\partial x^k}{\partial y^i}\,\frac{\partial x^l}{\partial y^j}$ and similarly for $\widehat{G}$. From (2.7), one gets $g_{1r} = \lambda_1^k\,\widehat{g}_{kr}$. In y-coordinates, we have $g_{1r} = G_{ij}(y)\,\frac{\partial y^i}{\partial x^1}\,\frac{\partial y^j}{\partial x^r}$ and $\lambda_1^k\,\widehat{g}_{kr} = \lambda_1^k\,\widehat{G}_{ij}\,\frac{\partial y^i}{\partial x^k}\,\frac{\partial y^j}{\partial x^r} = \widehat{G}_{1j}\,\frac{\partial y^j}{\partial x^r}$. Thus we obtain

$$G_{ij}(y)\,\frac{\partial y^i}{\partial x^1}\,\frac{\partial y^j}{\partial x^r} = \widehat{G}_{1j}\,\frac{\partial y^j}{\partial x^r}. \tag{3.4}$$

In a similar fashion, one sees that the $\widehat{g}$-equations in y-coordinates read

$$\frac{\partial \widehat{G}_{ij}}{\partial y^1} = \frac{\partial g_{k\ell}}{\partial x^1}\,\frac{\partial x^k}{\partial y^i}\,\frac{\partial x^\ell}{\partial y^j}\,, \tag{3.5}$$

and the $\widehat{V}$-equations become

$$\frac{\partial \widehat{V}}{\partial y^1} = \frac{\partial V}{\partial x^1}\,.$$

It is easy to see now that the following result holds.

THEOREM 3.2. *Given any nonzero solution to the λ-equation, there is a unique solution to the $\widehat{g}$- and $\widehat{V}$-equations with initial data prescribed at $y^1 = 0$.*

Remark 3.3. Note that (3.4) directly gives

$$\widehat{G}_{k1} = G_{ki}\,\frac{\partial y^i}{\partial x^1}\,,$$

and so one needs only to solve (3.5) for $n\,(n-1)/2$ quantities $\widehat{G}_{ij}$, $2 \le i,\, j \le n$.

4. Example 1: Inverted pendulum in a vertical plane. As the first example, we consider the inverted pendulum restricted to a vertical plane with horizontal and vertical actuation of the base; see Figure 4.1.

After rescaling units, the mass matrix and potential energy are given by

$$g = \begin{pmatrix} 1 & -a\,\cos(x^1) & -a\,\sin(x^1) \\ -a\,\cos(x^1) & 1 & 0 \\ -a\,\sin(x^1) & 0 & 1 \end{pmatrix},$$

$$V = b\,x^3 + \cos(x^1).$$

Since only x^1 is unactuated, we will simplify notation and use λ^i to denote λ^i_1. The λ-equations (2.9) are

$$\partial_1 \nu = 2a \sin(x^1)\,\lambda^2 - 2a \cos(x^1)\,\lambda^3, \quad \partial_2 \nu = 0, \quad \partial_3 \nu = 0,$$

with $\nu = \lambda^1 - a\cos(x^1)\,\lambda^2 - a\sin(x^1)\,\lambda^3$. (Note that $\partial_2 \nu = \partial_3 \nu = 0$ are the ν-equations.) It is not difficult to see that the general solution to these equations is

$$\begin{aligned}
\lambda^1 &= \nu(x^1) + \frac{1}{2}\cot(x^1)\,\partial_1\nu(x^1) + a\frac{\lambda^3(x^1,x^2,x^3)}{\sin(x^1)},\\
\lambda^2 &= \frac{1}{2a\sin(x^1)}\,\partial_1\nu(x^1) + \cot(x^1)\,\lambda^3(x^1,x^2,x^3),\\
\lambda^3 &= \lambda^3(x^1,x^2,x^3),
\end{aligned}$$

where $\nu(x^1)$, $\lambda^3(x^1,x^2,x^3)$ are arbitrary. In order to obtain a manageable explicit solution to the matching equations, we will choose

$$\nu(x^1) = a\,\mu_0 \sin^2(x^1) + \sigma_0 - a\,\mu_0, \qquad \lambda^3 = 0,$$

with free parameters σ_0 and μ_0. Then

$$\lambda^1 = \sigma_0, \qquad \lambda^2 = \mu_0 \cos(x^1).$$

The coordinates

$$y^1 = \frac{1}{\sigma_0}\,x^1, \quad y^2 = x^2 - \mu_0 \sin(x^1), \quad y^3 = x^3$$

satisfy (3.3). Following Remark 3.3, we need to solve the $\widehat{g}$-equations only for $\widehat{g}_{22}$, $\widehat{g}_{23}$, and $\widehat{g}_{33}$. These equations are

$$\frac{\partial}{\partial y^1}\,\widehat{g}_{22} = \frac{\partial}{\partial y^1}\,\widehat{g}_{23} = \frac{\partial}{\partial y^1}\,\widehat{g}_{33} = 0.$$

Clearly,

$$\begin{aligned}
\widehat{g}_{22} &= \widehat{g}_{22}(y^2,y^3) = \widehat{g}_{22}(x^2 - \mu_0 \sin(x^1),\ x^3),\\
\widehat{g}_{23} &= \widehat{g}_{23}(x^2 - \mu_0 \sin(x^1),\ x^3),\\
\widehat{g}_{33} &= \widehat{g}_{33}(x^2 - \mu_0 \sin(x^1),\ x^3).
\end{aligned}$$

From $g_{ai} = \lambda^j_a\,\widehat{g}_{ji}$, we obtain the rest of $\widehat{g}_{ij}$:

$$\begin{aligned}
\widehat{g}_{11} &= \frac{1}{\sigma_0} + \frac{a\mu_0}{\sigma_0^2}\cos^2(x^1) + \frac{\mu_0^2}{\sigma_0^2}\cos^2(x^1)\,\widehat{g}_{22},\\
\widehat{g}_{12} &= -\frac{a}{\sigma_0} - \frac{\mu_0}{\sigma_0}\cos(x^1)\,\widehat{g}_{22},\\
\widehat{g}_{13} &= -\frac{a}{\sigma_0}\sin(x^1) - \frac{\mu_0}{\sigma_0}\cos(x^1)\,\widehat{g}_{23}.
\end{aligned}$$

The $\widehat{V}$-equation yields

$$\widehat{V} = \frac{1}{\sigma_0}\cos(x^1) + w(y^2,y^3).$$

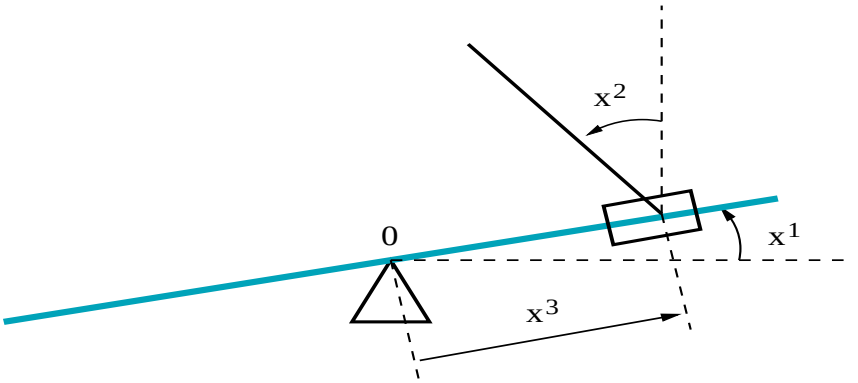


FIG. 5.1.

The $\widehat{C}$-equation reads $\lambda^j \widehat{C}_j = 0$. One solution is

$$\widehat{C} = -\sigma_0 R(x) \begin{pmatrix} \frac{\mu_0^2}{\sigma_0^2}\cos^2(x^1) & -\frac{\mu_0}{\sigma_0}\cos(x^1) & -\frac{\mu_0}{\sigma_0}\cos(x^1) \\ -\frac{\mu_0}{\sigma_0}\cos(x^1) & 1 & 1 \\ -\frac{\mu_0}{\sigma_0}\cos(x^1) & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} \dot{x}^1 \\ \dot{x}^2 \\ \dot{x}^3 \end{pmatrix}.$$

The resulting control law can be obtained explicitly from (2.4). The expression is too long to be included in this paper.

PROPOSITION 4.1. *If the functions $\widehat{g}_{22}(y^2, y^3)$, $\widehat{g}_{23}(y^2, y^3)$, $\widehat{g}_{33}(y^2, y^3)$, $w(y^2, y^3)$, and $R(x)$ and the parameters μ_0 and σ_0 are chosen so that*

$$\begin{aligned} &\widehat{g}_{22}(0) > 0, \quad \widehat{g}_{23}(0) = 0, \quad \widehat{g}_{33}(0) = 1, \\ &\partial^2_{y^2} w(0) > 0, \quad \partial_{y^2}\partial_{y^3} w(0) = 0, \quad \partial^2_{y^3} w(0) > 0, \qquad R(0) > 0, \\ &\sigma_0 < 0, \quad \widehat{g}_{22}(0)\,\mu_0^2 + a\,\mu_0 + \sigma_0 > 0, \quad \widehat{g}_{22}(0)\,(a\,\mu_0 - \sigma_0) + a^2 < 0, \end{aligned}$$

then $x = \dot{x} = 0$ is a locally asymptotically stable equilibrium of the closed loop system.

It follows from the above inequalities that the function $\widehat{H}(x, \dot{x}) = \frac{1}{2}\widehat{g}_{ij}(x)\,\dot{x}^i\dot{x}^j + \widehat{V}(x)$ has a strict local minimum at $x = \dot{x} = 0$. Also, the matching procedure automatically ensures that $\frac{d}{dt}\widehat{H} = -\widehat{g}(\widehat{C}(x, \dot{x}), \dot{x})$. The above inequalities guarantee that $-\widehat{g}(\widehat{C}(x, \dot{x}), \dot{x}) \leq 0$ in the neighborhood of $x = \dot{x} = 0$. In addition, the set where $\widehat{g}(\widehat{C}(x, \dot{x}), \dot{x}) = 0$ contains no local solutions of the closed loop system except for $x = \dot{x} = 0$. This proves the proposition.

5. Example 2: Inverted pendulum cart on a seesaw. In the previous example, the kernel of the matrix $A^\star$, (2.16), was two-dimensional. Generically, for systems with one unactuated degree of freedom, the dimension of the kernel will be 1. The following example illustrates this situation. The inverted pendulum cart on a seesaw is shown in Figure 5.1. There are several interesting ways to actuate this system. We will consider the case with actuated cart and pendulum, and unactuated seesaw.

The rescaled mass matrix and potential energy of the system are given by

$$g = \begin{pmatrix} b + (x^3)^2 & a\,x^3\sin(x^1 - x^2) & 0 \\ a\,x^3\sin(x^1 - x^2) & 1 & -a\cos(x^1 - x^2) \\ 0 & -a\cos(x^1 - x^2) & 1 \end{pmatrix}$$

and

$$V = x^3 \sin(x^2) + a \cos(x^1).$$

The theory in section 3 was presented with special coordinates so that $g_{11} = 1$. However, in practice, this is not necessary.

As before, we write λ^i for λ_1^i and ν for $g_{1j}\,\lambda^j$. The λ-equations are

$$\begin{aligned}
\partial_1 \nu &= 2a\,x^3 \cos(x^1 - x^2)\,\lambda^2 - 2\,x^3\,\lambda^3,\\
\partial_2 \nu &= 0,\\
\partial_3 \nu &= 2a \sin(x^1 - x^2)\,\lambda^2 + 2\,x^3\,\lambda^1.
\end{aligned}$$

Note that, in this case, the ν-equations (2.17) are simply $\partial_2 \nu = 0$. Hence $\nu = \nu(x^1, x^3)$. Plug in $\lambda^1 = (\nu - g_{12}\,\lambda^2 - g_{13}\,\lambda^3)/g_{11}$, and solve for λ^2 and λ^3:

$$\begin{aligned}
\lambda^1 &= \frac{1}{2b}\,(\,2\nu - x^3 \partial_3\,\nu\,),\\
\lambda^2 &= \frac{1}{2ab\,\sin(x^1 - x^2)}\,(-\,2\,x^3\,\nu + (b + (x^3)^2)\,\partial_3\,\nu),\\
\lambda^3 &= \frac{1}{2b\,x^3\,\sin(x^1 - x^2)}\,(-\,2(x^3)^2 \cos(x^1 - x^2)\,\nu\\
&\qquad + x^3\,(b + (x^3)^2) \cos(x^1 - x^2)\,\partial_3\nu - b \sin(x^1 - x^2)\,\partial_2\nu).
\end{aligned}$$

Notice that λ^2 and λ^3 blow up as x approaches 0 unless $\nu = \kappa\,(b + (x^3)^2)$. Since $g = \widehat{g}\,\lambda$, one must have $\det \widehat{g} \to 0$ as $x \to 0$; i.e., $\widehat{g}$ degenerates at $x = 0$. This means that $\widehat{H}(x, \dot{x}) = \frac{1}{2}\,\widehat{g}_{ij}\,\dot{x}^i\,\dot{x}^j + \widehat{V}$ cannot serve as a Lyapunov function unless $\nu = \kappa\,(b + (x^3)^2)$. This ν corresponds exactly to the basic solutions of the matching equations from Remark 2.2. This illustrates the following general principle.

Remark 5.1. If $(x_0, 0)$ is the desired equilibrium of a system and

$$\operatorname{rank} A^\star(x_0) < \limsup_{x \to x_0} \operatorname{rank} A^\star(x), \tag{5.1}$$

then only basic solutions of the matching equations should be tested to produce a stabilizing control law from (2.4).

6. Example 3: Inverted pendulum cart on a roller coaster. Consider a cart with an inverted pendulum on a roller coaster. Special cases of this mechanical system include the inverted pendulum on a rotor arm, the inverted pendulum on a vertical disk, and the inverted pendulum cart on an incline. By assuming that the size of the base of the cart is relatively small, we may neglect the inertia of the base of the cart. It is therefore sufficient to model the cart with one point mass for the base and one point mass a fixed distance away for the pendulum. The pendulum joint will be unactuated.

The configuration of the system may be described by a position and an angle. Assume that the shape of the roller coaster is given as a curve $x(s)$ in $\mathbb{R}^3$ parametrized by arc length, s, from a fixed point. Assume that the pendulum is always in the plane containing the tangent vector, $\tau(s)$, and the vertical direction, e_3. Let ϕ be the angle between the pendulum and e_3. By rescaling mass, length, and time, we will

write

$$g = \begin{pmatrix} 1 & b\sin(\alpha-\phi) \\ b\sin(\alpha-\phi) & \left(1+k(s)^2\sin^2\phi\left(\frac{\sin^2\alpha-n_3^2}{\sin^4\alpha}\right)\right) \end{pmatrix},$$

$$V = a\,x^3 + \cos\phi,$$

where a and b are positive parameters, $0 < b < 1$, and x^3 is the vertical component of $x(s)$. The (unit) tangent vector to the curve is $\tau(s) = \frac{x'(s)}{|x'(s)|}$, where $'$ stands for the derivative with respect to s. The curvature of the curve is $k(s) = |\tau'(s)|$. Denote by $n(s)$ the principal normal to the curve. Recall that $\tau'(s) = k(s)\,n(s)$. In the formula above, n_3 is the vertical component of the principal normal, and α is the angle between τ and the vertical direction. Index 1 corresponds to ϕ, and index 2 corresponds to s. The unactuated degree of freedom corresponds to the ϕ variable. The λ-equations (3.1) then read as follows:

$$\begin{aligned} \partial_1\nu &= -2\,b\,\cos(\alpha-\phi)\,\lambda_1^2, \\ \partial_2\nu &= k(s)^2\sin(2\,\phi)\left(\frac{\sin^2\alpha-n_3^2}{\sin^4\alpha}\right)\lambda_1^2\,. \end{aligned} \tag{6.1}$$

The orthogonality equation (3.2) then, obviously, is

$$k(s)^2\sin(2\,\phi)\left(\frac{\sin^2\alpha-n_3^2}{\sin^4\alpha}\right)\frac{\partial\nu}{\partial\phi} + 2\,b\,\cos(\alpha-\phi)\,\frac{\partial\nu}{\partial s} = 0. \tag{6.2}$$

It is not clear if all solutions to this equation can be written explicitly for a general curve. We consider here two particular cases when this is possible. The first case is when $\sin^2\alpha = n_3^2$. This occurs exactly when the roller coaster lies in one vertical plane. The second case is when $\alpha(s)$ is constant. This occurs when the track is constantly inclined.

6.1. Case 1: $\sin^2\alpha = n_3^2$. Note that this case includes the interesting examples of an inverted pendulum on a vertical disk and an inverted pendulum cart on an incline.

As is readily seen from (6.2), the general solution of (6.2) in this case is $\nu = \nu(\phi)$, an arbitrary function. Then

$$\lambda_1^2 = -\frac{1}{2\,b\,\cos(\alpha-\phi)}\,\frac{\partial\nu}{\partial\phi},$$

and

$$\lambda_1^1 = \nu(\phi) + \frac{1}{2}\tan(\alpha-\phi)\,\frac{\partial\nu}{\partial\phi}.$$

This is a general solution of the λ-equation. From here, one must solve the $\widehat{g}$- and $\widehat{V}$-equations. For special choices of $\alpha(s)$ and/or $\nu(\phi)$, these equations have explicit closed form solutions.

6.2. Case 2: $\alpha(s) = \alpha_0$. Examples with $\alpha(s)$ constant include an inverted pendulum cart traveling on any path in a horizontal plane and a cart on a vertically oriented circular helix or any constantly inclined track.

Since $\frac{d\alpha}{ds} = -k(s)\frac{n_3(s)}{\sin(\alpha)}$, if $\alpha(s) = \alpha_0$, then we have $k(s)n_3(s) = 0$. To solve (6.2), we introduce new coordinates

$$\begin{aligned} z^1 &= \beta(s), \\ z^2 &= \beta(s) + \cos(\alpha_0)\ln|\csc\phi + \cot\phi| - \sin(\alpha_0)\ln|\sec\phi + \tan\phi|, \end{aligned}$$

where

$$\beta(s) = \int_0^s \frac{k^2(p)}{b\sin^2(\alpha_0)}\,dp.$$

Equation (6.2) is equivalent to the fact that ν is an arbitrary function of z^2. Thus, from (6.1) and the definition of ν, we find

$$\lambda_1^1 = \nu(z^2) - \frac{\sin(\alpha_0 - \phi)}{\sin(2\phi)}\frac{d\,\nu}{d\,z^2}; \qquad \lambda_1^2 = \frac{1}{b\,\sin(2\phi)}\frac{d\,\nu}{d\,z^2}.$$

6.3. End of the roller coaster example. From the computations in case 1 and case 2, one can see that the general solution to the matching equations for the cart on a roller coaster will be fairly complicated. However, we can show that any *linear* control law is the first order germ (linearization) of some matching control law. In fact, the only requirement for this is that there is no rank drop at the equilibrium; i.e.,

$$\operatorname{rank} A^\star(x_0) = \limsup_{x\to x_0} \operatorname{rank} A^\star(x). \tag{6.3}$$

We assume that the dissipative term at the equilibrium satisfies the following natural assumptions:

$$C_\ell(x_0, 0) = 0, \qquad \left.\frac{\partial}{\partial x^i}\right|_{(x_0,0)} C_\ell = 0.$$

LEMMA 6.1. *If condition* (6.3) *is satisfied for a two degree of freedom system, then the first order germs of matching control laws at* $(x_0, 0)$ *exhaust all linear control laws for which the closed loop system has an equilibrium at* $(x_0, 0)$.

Proof. Given a linear control input

$$u_j^{\text{lin}} = v_j + a_{ij}\,(x^i - x_0^i) + b_{ij}\,\dot{x}^i$$

with $u_1^{\text{lin}} = 0$, we will find a matching control law with the same germ. From the general expression (2.4) for the matching control law, we see that the first order germ is

$$u_j^{\text{germ}} = (V_j - g_{j\ell}\,\widehat{g}^{\ell r}\,\widehat{V}_r) + \left(V_{jr} - g_{j\ell}\,\widehat{g}^{\ell i}\,\widehat{V}_{ir}\right)(x^r - x_0^r) + \left(C_{ji} - g_{j\ell}\,\widehat{g}^{\ell r}\,\widehat{C}_{ri}\right)\dot{x}^i,$$

where

$$V_j = \left.\frac{\partial V}{\partial x^j}\right|_{x_0}, \quad V_{jr} = \left.\frac{\partial^2 V}{\partial x^j \partial x^r}\right|_{x_0}, \quad C_{ji} = \left.\frac{\partial C_j}{\partial \dot{x}^i}\right|_{(x_0,0)},$$

and $\widehat{V}_j$, $\widehat{V}_{jr}$, and $\widehat{C}_{ji}$ are defined similarly. Equating like terms gives

$$\widehat{V}_{\ell j} = \widehat{g}_{\ell i}\, g^{ir}\,(V_{rj} - a_{rj}), \quad \widehat{C}_{\ell j} = \widehat{g}_{\ell i}\, g^{ir}\,(C_{rj} - b_{rj}), \quad \widehat{V}_{\ell} = \widehat{g}_{\ell i}\, g^{ir}\,(V_r - v_r) = 0. \tag{6.4}$$

One can see that $\widehat{V}_j$, $\widehat{V}_{jr}$, and $\widehat{C}_{ji}$ are specified once $\widehat{g}_{ij}(x_0)$ is known. Moreover, there exists a nondegenerate symmetric $\widehat{g}_{ij}(x_0)$ such that the resulting $\widehat{V}_{\ell j}$ will be symmetric; see [1, Lemma 1]. To conclude the argument, we now show that any nondegenerate symmetric $\widehat{g}_{ij}(x_0)$ arises as a zero order germ of a solution to the $\widehat{g}$-equation. Also, any $\widehat{V}_\ell$, $\widehat{V}_{\ell j}$ satisfying $\widehat{V}_\ell = \widehat{g}_{\ell i}\, g^{ir}\,(V_r - v_r)$ and $\widehat{V}_{\ell j} = \widehat{g}_{\ell i}\, g^{ir}\,(V_{rj} - a_{rj})$ arises as a solution to the $\widehat{V}$-equation.

Given a nondegenerate $\widehat{g}_{ij}(x_0)$, define the nondegenerate $\lambda_i^j(x_0) = g_{ik}(x_0)\,\widehat{g}^{kj}(x_0)$. Set $\nu_0 = g_{11}(x_0)\,\lambda_1^1(x_0) + g_{12}(x_0)\,\lambda_1^2(x_0)$. The λ-equations in this case are

$$\begin{aligned} \partial_1\nu - 2\,[1\,1,\,2]\,\frac{1}{g_{11}}\,\nu &= 2\,[1\,1,\,2]\,\lambda_1^2, \\ \partial_2\nu - 2\,[1\,2,\,1]\,\frac{1}{g_{11}}\,\nu &= 2\,[2\,1,\,2]\,\lambda_1^2. \end{aligned} \tag{6.5}$$

By the rank condition, we know that the rank of $A^\star$ in the neighborhood of x_0 is either identically 0 or identically 1. If this rank is 0, then ν can be any constant times g_{11}. We simply choose $\nu(x) = (\nu_0/g_{11}(x_0))\,g_{11}(x)$. Any solution to the algebraic equation $\nu(x) = g_{11}(x)\,\lambda_1^1(x) + g_{12}(x)\,\lambda_1^2(x)$ is a solution to the λ-equation. If the rank of $A^\star$ is 1, then the orthogonality condition, (2.17), is

$$[2\,1,\,2]\,\partial_1\nu - [1\,1,\,2]\,\partial_2\nu + 2\,\left([1\,1,\,2]\,[1\,2,\,1] - [2\,1,\,2]\,[1\,1,\,2]\right)\,\frac{1}{g_{11}}\,\nu = 0. \tag{6.6}$$

At $x = x_0$, either ∂_1 or ∂_2 is not parallel to the vector $[2\,1,\,2]\,\partial_1 - [1\,1,\,2]\,\partial_2$. Assume it is ∂_1. Then initial conditions to (6.6) can be specified along the line $x^2 = x_0^2$. In particular, we can choose the initial values so that

$$\left(\partial_1\nu - 2\,[1\,1,\,2]\,\frac{1}{g_{11}}\,\nu\right)\Big|_{x=x_0} = 2\,[1\,1,\,2]\Big|_{x=x_0}\lambda_1^2(x_0), \qquad \nu(x_0) = \nu_0. \tag{6.7}$$

The second equation in (6.5),

$$\left(\partial_2\nu - 2\,[1\,2,\,1]\,\frac{1}{g_{11}}\,\nu\right)\Big|_{x=x_0} = 2\,[2\,1,\,2]\Big|_{x=x_0}\lambda_1^2(x_0),$$

will be satisfied automatically since the rank of $A^\star$ is 1. Let $\nu(x)$ be a solution of (6.6) with initial condition, $\nu(x^1, 0)$, satisfying (6.7). Now one solves (6.5) for $\lambda_1^2(x)$ and then $\nu(x) = g_{11}(x)\,\lambda_1^1(x) + g_{12}(x)\,\lambda_1^2(x)$ for $\lambda_1^1(x)$.

Now that the λ-equations are solved, we turn to the $\widehat{g}$-equations. These equations take the form

$$\frac{\partial}{\partial y^1}\,\widehat{g} + R\,\widehat{g} = S,$$

where $\frac{\partial}{\partial y^1} = \lambda_1^1\,\partial_1 + \lambda_1^2\,\partial_2$. The initial conditions can be set on any line transverse to $\frac{\partial}{\partial y^1}$, in particular, along the line $\lambda_1^1(x_0)\,(x^1 - x_0^1) + \lambda_1^2(x_0)\,(x^2 - x_0^2) = 0$. On this

line, set $\lambda_2^i(x) = \lambda_2^i(x_0)$ and $\widehat{g}(x) = g(x) \cdot (\lambda(x_0))^{-1}$. The solution to (2.10) with this initial data then has the desired value at $x = x_0$.

It remains to show that the $\widehat{V}$-equation has a solution such that $\widehat{V}_\ell = 0$ and

$$\lambda_r^\ell(x_0)\, \widehat{V}_{\ell j} = V_{rj} - a_{rj}, \tag{6.8}$$

where $\lambda_r^\ell(x_0) = g_{rk}(x_0)\, \widehat{g}^{k\ell}(x_0)$. Since $\lambda_r^\ell(x_0)$ is nondegenerate, either $\lambda_1^1(x_0) \neq 0$ or $\lambda_1^2(x_0) \neq 0$. Consider the case with $\lambda_1^2(x_0) \neq 0$. In this case, the line $x^2 = x_0^2$ is noncharacteristic for the $\widehat{V}$-equation

$$\lambda_1^1\, \partial_1 \widehat{V} + \lambda_1^2\, \partial_2 \widehat{V} = \partial_1 V. \tag{6.9}$$

Pick the initial value, $\widehat{V}\big|_{x^2 = x_0^2}$, so that

$$\widehat{V}_1 = 0, \quad \widehat{V}_{11} = \frac{\lambda_2^2\, V_{11} - \lambda_1^2\, (V_{12} - a_{12})}{\lambda_1^1\, \lambda_2^2 - \lambda_2^1\, \lambda_1^2}\bigg|_{x = x_0},$$

and solve (6.9). Since $x = x_0$ is an equilibrium, $V_1 = 0$ and $(V_2 - v_2) = 0$. From the differential equation (6.9), we see that $\widehat{V}_2 = 0$. Differentiating equation (6.9) with respect to x^1 and x^2, we see that $W_{ij} = \widehat{V}_{ij}$ satisfies

$$\begin{aligned} \lambda_1^1(x_0)\, W_{11} + \lambda_1^2(x_0)\, W_{12} &= V_{11}, \\ \lambda_1^1(x_0)\, W_{12} + \lambda_1^2(x_0)\, W_{22} &= V_{12}. \end{aligned} \tag{6.10}$$

By construction,

$$\lambda_2^1(x_0)\, W_{11} + \lambda_2^2(x_0)\, W_{12} = V_{12} - a_{12}. \tag{6.11}$$

Notice that (6.4) implies that $W_{\ell j} = \widehat{g}_{\ell i}\, g^{ir}\, (V_{rj} - a_{rj})$ also satisfy (6.10) and (6.11). Since the solution to the algebraic system (6.10), (6.11) is unique, we conclude that (6.8) is valid, as required.

7. Example 4: A double pendulum on a wheel. Our next example is the system with two unactuated degrees of freedom depicted in Figure 7.1. Only joint **A** is actuated.

After rescaling, the entries g_{ij} of the mass matrix are

$$g_{ij} = m_{ij} \cos(x^i - x^j),$$

and the potential energy is

$$V = a_1 \cos(x^1) + a_2 \cos(x^2) + a_3 \cos(x^3).$$

The parameters $m_{ij} = m_{ji}$ and a_j are positive.

There are six unknown λ_a^i. Define $\nu_{ab} = g_{ai}\, \lambda_b^i$. Note that we must have $\nu_{12} = \nu_{21}$. The computations in this section were performed using Maple. The λ-equations

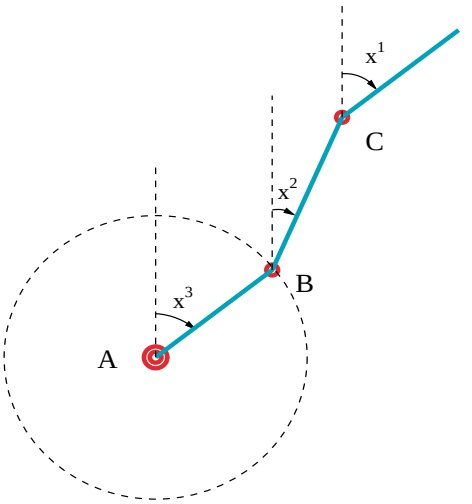


FIG. 7.1.

(2.9) are

$$\begin{aligned}
\partial_1 \nu_{11} &= -2\, m_{12} \sin(x_1 - x_2)\, \lambda_1^2 - 2\, m_{13} \sin(x_1 - x_2)\, \lambda_1^3,\\
\partial_2 \nu_{11} &= 0,\\
\partial_3 \nu_{11} &= 0,
\end{aligned}$$

$$\begin{aligned}
\partial_1 \nu_{22} &= 0,\\
\partial_2 \nu_{22} &= +2\, m_{12} \sin(x_1 - x_2)\, \lambda_2^1 - 2\, m_{23} \sin(x_2 - x_3)\, \lambda_2^3,\\
\partial_3 \nu_{22} &= 0,
\end{aligned}$$

$$\begin{aligned}
\partial_1 \nu_{12} &= -m_{12} \sin(x_1 - x_2)\, \lambda_2^2 - m_{13} \sin(x_1 - x_3)\, \lambda_2^3,\\
\partial_2 \nu_{12} &= +m_{12} \sin(x_1 - x_2)\, \lambda_1^1 - m_{23} \sin(x_2 - x_3)\, \lambda_1^3,\\
\partial_3 \nu_{12} &= 0.
\end{aligned}$$

The second step is to express λ_1^1, λ_2^1, λ_1^2, and λ_2^2 in terms of ν_{11}, ν_{12}, and ν_{22}.

After substitution into the above equations, we obtain

$$\begin{aligned}
\partial_1 \nu_{11} &= D^1_{1,1}\, \nu_{11} + D^2_{1,1}\, \nu_{12} + B^1_{1,1}\, \lambda_1^3,\\
\partial_2 \nu_{11} &= 0,\\
\partial_3 \nu_{11} &= 0,\\
\partial_1 \nu_{22} &= 0,\\
\partial_2 \nu_{22} &= D^2_{3,2}\, \nu_{12} + D^3_{3,2}\, \nu_{22} + B^2_{3,2}\, \lambda_2^3,\\
\partial_3 \nu_{22} &= 0,\\
\partial_1 \nu_{12} &= D^2_{2,1}\, \nu_{12} + D^3_{2,1}\, \nu_{22} + B^2_{2,1}\, \lambda_2^3,\\
\partial_2 \nu_{12} &= D^1_{2,2}\, \nu_{11} + D^2_{2,2}\, \nu_{12} + B^1_{2,2}\, \lambda_1^3,\\
\partial_3 \nu_{12} &= 0.
\end{aligned} \tag{7.1}$$

Here the $D^k_{i,j}$ and $B^k_{i,j}$ are explicit expressions involving x. In section 2, we described a general procedure to obtain compatibility conditions for this system. In this particular case, however, we use a different tactic: we compute and compare the mixed derivatives of ν_{ab}. The first set of equations we obtain is

$$\begin{aligned}\partial_3\partial_1\,\nu_{12} &= K_{11}\,\partial_3\lambda^3_2 + K_{12}\,\lambda^3_2 = 0,\\ \partial_3\partial_2\,\nu_{22} &= K_{21}\,\partial_3\lambda^3_2 + K_{22}\,\lambda^3_2 = 0.\end{aligned}$$

Direct computation shows that $\det(K_{ij}) \neq 0$. Hence $\lambda^3_2 = 0$. Similarly,

$$\begin{aligned}\partial_3\partial_1\,\nu_{11} &= L_{11}\,\partial_3\lambda^3_1 + L_{12}\,\lambda^3_1 = 0,\\ \partial_3\partial_2\,\nu_{12} &= L_{21}\,\partial_3\lambda^3_1 + L_{22}\,\lambda^3_1 = 0,\end{aligned}$$

and $\det(L_{ij}) \neq 0$. Hence $\lambda^3_1 = 0$.

Next we substitute $\lambda^3_1 = \lambda^3_2 = 0$ into (7.1) and solve for ν_{11}, ν_{12}, ν_{22}. This gives

$$\nu_{11} = \text{const}, \qquad \nu_{22} = \frac{m_{22}}{m_{11}}\,\nu_{11}, \qquad \nu_{12} = \frac{m_{12}}{m_{11}}\cos(x^2 - x^1)\,\nu_{11}.$$

Returning to λ-equations, we see that

$$\lambda^1_1 = \lambda^2_2 = \frac{1}{m_{11}}\,\nu_{11}, \qquad \lambda^1_2 = \lambda^2_1 = 0.$$

Our computation shows that the only solutions of the matching equations are the basic solutions defined in Remark 2.2.

8. Conclusion. This paper continues the development of the λ-method for matching control laws for underactuated systems. The matching control laws are those for which the closed loop system retains the structure of the open loop system. The general conditions for such control inputs can be expressed as a nonlinear system of partial differential equations. The λ-method was introduced to replace these nonlinear equations by a sequence of linear systems of partial differential equations, λ-equations, $\widehat{g}$-equations, $\widehat{V}$-equations, and a linear algebraic system (the $\widehat{C}$-equations). The coefficients for the $\widehat{g}$- and $\widehat{V}$-equations are obtained from the solutions of the λ-equations. The λ-equations form an overdetermined system. This paper introduces a new system of equations (the ν-equations) encoding the compatibility conditions of the λ-equations. Four examples are presented to illustrate different aspects of the theory with one or more unactuated degrees of freedom.

REFERENCES

[1] F. Andreev, D. Auckly, S. Gosavi, L. Kapitanski, A. Kelkar, and W. White, *Matching linear system, and ball and beam*, Automatica J. IFAC, 2002, to appear; archived online at http://xxx.lanl.gov/abs/math.OC/0003177.

[2] F. Andreev, D. Auckly, L. Kapitanski, A. Kelkar, and W. White, *Matching control laws for a ball and beam system*, in Proceedings of the IFAC Workshop on Lagrangian and Hamiltonian Methods for Nonlinear Control, Princeton, NJ, 2000, pp. 161–162.

[3] F. Andreev, D. Auckly, L. Kapitanski, A. Kelkar, and W. White, *Matching and digital control implementation for underactuated systems*, in Proceedings of the American Control Conference, Chicago, IL, 2000, pp. 3934–3938.

[4] D. Auckly and L. Kapitanski, *Mathematical problems in the control of underactuated systems*, in Nonlinear Dynamics and Renormalization Group, I. M. Sigal and C. Sulem, eds., CRM Proc. Lecture Notes 27, AMS, Providence, RI, 2001, pp. 29–40.

[5] D. Auckly, L. Kapitanski, and W. White, *Control of nonlinear underactuated systems*, Comm. Pure Appl. Math., 53 (2000), pp. 354–369.

[6] G. Blankenstein, R. Ortega, and A. J. van der Schaft, *The matching conditions of controlled Lagrangians and IDA-passivity based control*, Internat. J. Control, 75 (2002), pp. 645–665.

[7] A. M. Bloch, D. Chang, N. Leonard, and J. E. Marsden, *Controlled Lagrangians and the stabilization of mechanical systems* II: *Potential shaping*, IEEE Trans. Automat. Control, 46 (2001), pp. 1556–1571.

[8] A. Bloch, N. Leonard, and J. Marsden, *Stabilization of mechanical systems using controlled Lagrangians*, in Proceedings of the 36th IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 1997, pp. 2356–2361.

[9] A. M. Bloch, N. Leonard, and J. E. Marsden, *Controlled Lagrangians and the stabilization of mechanical systems* I: *The first matching theorem*, IEEE Trans. Automat. Control, 45 (2000), pp. 2253–2270.

[10] J. Hamberg, *General matching conditions in the theory of controlled Lagrangians*, in Proceedings of the IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 1999, pp. 2519–2523.

[11] J. Hamberg, *Controlled Lagrangians, symmetries and conditions for strong matching*, in Proceedings of the IFAC Workshop on Lagrangian and Hamiltonian Methods for Nonlinear Control, Princeton, NJ, 2000, pp. 62–67.

[12] J. Hamberg, *Simplified conditions for matching and for generalized matching in the theory of controlled Lagrangians*, in Proceedings of the American Control Conference, Chicago, IL, 2000, pp. 3918–3923.

[13] D. V. Zenkov, A. M. Bloch, N. E. Leonard, and J. E. Marsden, *Matching and stabilization of low-dimensional nonholonomic systems*, in Proceedings of the 39th IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 2000, pp. 1289–1295.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 5, pp. 1389–1412

GEOMETRIC DESCRIPTION OF VAKONOMIC AND NONHOLONOMIC DYNAMICS. COMPARISON OF SOLUTIONS*

J. CORTÉS†, M. DE LEÓN‡, D. MARTÍN DE DIEGO†, AND S. MARTÍNEZ§

Abstract. We treat the vakonomic dynamics with general constraints within a new geometric framework, which can be useful in the study of optimal control problems. We compare our formulation with the one of Vershik and Gershkovich in the case of linear constraints. We show how nonholonomic mechanics also admits a new geometrical description which allows us to develop an algorithm of comparison between the solutions of both dynamics. Examples illustrating the theory are treated.

Key words. vakonomic dynamics, nonholonomic dynamics, optimal control, symplectic geometry

AMS subject classifications. 34A26, 49K15, 70F25

PII. S036301290036817X

1. Introduction.

As is well known, the application of tools from modern differential geometry in the fields of mechanics and control theory has caused an important progress in these research areas. For example, the study of the geometrical formulation of the nonholonomic equations of motion has led to a better comprehension of locomotion generation, controllability, motion planning, and trajectory tracking, raising new interesting questions in these subjects (see [4, 5, 26, 28, 35, 47, 49, 50] and references therein). On the other hand, there are by now many papers in which optimal control problems are addressed using geometric techniques (references [8, 23, 24, 59, 60] are good examples).

In this context, we present a unified geometrical formulation of the dynamics of nonholonomic and vakonomic systems. Both kinds of systems have the same mathematical "ingredients": a Lagrangian function and a set of nonintegrable constraints. But the way in which the equations of motion are derived differs. In the case of vakonomic systems, the dynamics is obtained through the application of a constrained variational principle [1]. In particular, an optimal control problem can be seen as a vakonomic one. The term "vakonomic" ("variational axiomatic kind") is inherited from Kozlov [29], who proposed this mechanics as an alternative set of equations of motion for a physical system in the presence of nonholonomic constraints. Nonholonomic equations of motion are deduced using d'Alembert's principle when the constraints are linear or affine.

The two approaches have received a lot of attention in recent years (see [1, 2,

*Received by the editors February 24, 2000; accepted for publication (in revised form) July 11, 2002; published electronically January 3, 2003. This research was partially supported by Spanish DGICYT grants PB97-1257 and PGC2000-2191-E. The research of the first and fourth authors was partially supported by FPU and FPI grants from the Spanish Ministerio de Educación y Cultura and Ministerio de Ciencia y Tecnología, respectively.

http://www.siam.org/journals/sicon/41-5/36817.html

†Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, 1308 West Main Street, Urbana, IL 61801 (jorge@motion.csl.uiuc.edu).

‡Laboratory of Dynamical Systems, Mechanics and Control, Instituto de Matemáticas y Física Fundamental, CSIC, Serrano 123, Madrid 28006, Spain (mdeleon@imaff.cfmac.csic.es, d.martin@imaff.cfmac.csic.es).

§Escola Universitària Politècnica de Vilanova i la Geltrú, Universidad Politécnica de Cataluña, Av. V. Balaguer s/n, Vilanova i la Geltrú 08800, Spain (soniam@mat.upc.es).

10, 14, 33, 31, 36, 41, 64, 68] and references therein). Vakonomic mechanics (also called dynamical optimization subject to nonholonomic constraints) is used in mathematical economics (growth economic theory), sub-Riemannian geometry, motion of microorganisms, etc., while nonholonomic mechanics provides the evolution equations for wheeled and autonomous vehicles, robotic systems, etc.

Several authors have discussed the domains of validity of both approaches [1, 29, 36, 68]. The solutions of the resulting dynamical systems do not coincide, in general, though there are examples in which the nonholonomic solutions can be seen as solutions of the constrained variational problem. In recent papers [20, 36] the characterization of this situation has been studied. In [36] Lewis and Murray considered the example of a ball on a rotating table and showed that the subset of solutions of the nonholonomic problem is not included in the set of vakonomic ones. In [20] Favretti obtains conditions in some particular cases for the equivalence between both formulations.

Our project of unifying the comparative studies of both types of dynamics from a geometrical point of view has brought us to develop a new geometric setting for vakonomic and nonholonomic mechanics, strongly inspired on the Skinner and Rusk formulation for singular Lagrangian systems [58]. Herewith, we are able to compare them using an algorithm which gives rise, under appropriate conditions, to a final constraint submanifold containing all the nonholonomic solutions which are also vakonomic. As an application of the proposed algorithm, we extend several known results [4, 20, 36]. In particular, we prove that any solution of the unconstrained problem which verifies the constraints is simultaneously a solution of the nonholonomic and the vakonomic problems. This allows us to generalize to arbitrary metrics a result proven in [20] for bundle-like metrics and kinetic energy Lagrangians, $L = \frac{1}{2}g$.

The paper is structured as follows. In section 2, we obtain the equations of motion for vakonomic mechanics, assuming an admissibility condition, which permits us to present them in terms of the restriction of the Lagrangian to the constraint submanifold M. Let us recall that from a geometrical point of view, the Lagrangian L is defined on the tangent bundle TQ of the configuration manifold Q, and M represents the submanifold of TQ determined by the vanishing of the nonholonomic constraint functions. We will deal here with arbitrary submanifolds, that is, the constraints may be nonlinear. It should also be pointed out that we do not consider abnormal solutions. It is interesting to note that our derivation of the equations of motion shows that the information provided by L outside M is completely irrelevant for the vakonomic problem. This fact is not clearly seen in the classical way of writing the equations for vakonomic systems [1, 29].

Section 3 is devoted to a reformulation of vakonomic mechanics in geometric terms. In this section we will use as ambient space the fibered manifold $W_0 = T^*Q \times_Q M$, which is in fact a subbundle of the Whitney sum $T^*Q \oplus TQ$ (the phase space in the Skinner and Rusk approach). Since T^*Q is equipped with a canonical symplectic form we can induce a presymplectic structure ω on $T^*Q \times_Q M$. Moreover, we can consider the Hamiltonian function $H_{W_0} = \langle \pi_1, \pi_2 \rangle - \pi_2^* \tilde{L}$, where π_1 and π_2 are the canonical projections, $\langle \cdot, \cdot \rangle$ denotes the natural pairing between vectors and covectors on Q, and $\tilde{L}$ is the restriction of L to M. Then, we prove that the equations of motion of vakonomic mechanics are intrinsically represented by the presymplectic Hamiltonian equation $i_X \omega = dH_{W_0}$. Since the 2-form ω is presymplectic, a constraint algorithm must be applied in order to obtain well-defined solutions of the dynamics. If the algorithm stabilizes, we obtain a family of explicit solutions on the final constraint

submanifold. In addition, a compatibility condition is found which determines when the first constraint submanifold W_1 is symplectic (and therefore the algorithm stabilizes at the first step). We illustrate in subsections 3.1 and 3.2 how this framework can be of use in the analysis of optimal control problems.

In section 4, we compare our approach with the one of Vershik and Gershkovich [64] for vakonomic systems with linear constraints. We prove that both are related by a convenient presymplectomorphism, so that our approach could be considered as a generalization to the case of nonlinear constraints.

Since we want to compare vakonomic and nonholonomic dynamics, it is necessary to construct a geometrical framework for nonholonomic mechanics using a closed phase space. Indeed, in section 5 it is proved that the nonholonomic dynamics lives on a submanifold $\tilde{M}$ of W_0. In general, we have again a presymplectic system and a constraint algorithm is needed to obtain the dynamics on the final constraint submanifold.

In section 6, assuming that the vakonomic and the nonholonomic dynamics live on W_1 and $\tilde{M}$, respectively, we can compare their solutions by means of the map $\Upsilon : W_1 \to \tilde{M}$, $(\alpha, v) \mapsto (Leg_L(v), v)$. We present here an algorithm that selects those solutions of the nonholonomic problem that can be seen as solutions of the constrained variational one. Several illustrative examples are worked out in order to illustrate the different behaviors, showing that our framework provides a generalization and common context for the equivalence results in [4, 20, 36]. In particular, in the example of the planar mobile robot, we prove that, under an appropriate design of the system, every solution of the nonholonomic problem can be seen as a solution of the vakonomic one.

2. Variational approach to constrained mechanics. Let Q be the configuration manifold with dimension n and $L : TQ \longrightarrow \mathbb{R}$ an autonomous Lagrangian function. If (q^A), $1 \leq A \leq n$, are coordinates on Q, we denote by $(q^A, \dot{q}^A)$ the natural bundle coordinates on TQ in terms of which the tangent bundle projection $\tau_Q : TQ \longrightarrow Q$ reads as $\tau_Q(q^A, \dot{q}^A) = (q^A)$.

Let us suppose that the system is subject to some constraints given by a $(2n-m)$-dimensional submanifold M of TQ, locally defined by $\Phi^\alpha = 0$, $1 \leq \alpha \leq m$, where $\Phi^\alpha : TQ \longrightarrow \mathbb{R}$. Throughout the paper, we will assume the following admissibility condition for the submanifold $M \subseteq TQ$: for all $x \in M$, $\dim T_x M^o = \dim S^* T_x M^o$, where $S = dq^A \otimes \frac{\partial}{\partial \dot{q}^A}$ is the canonical vertical endomorphism (see [34]). This is equivalent to saying that the rank of the matrix

$$\frac{\partial(\Phi^1, \ldots, \Phi^m)}{\partial(\dot{q}^1, \ldots, \dot{q}^n)}$$

is m for any choice of coordinates $(q^A, \dot{q}^A)$ in TQ. Consequently, by the implicit function theorem, we can locally express the constraints (reordering coordinates if necessary) as

$$\dot{q}^\alpha = \Psi^\alpha(q^A, \dot{q}^a), \tag{2.1}$$

where $1 \leq \alpha \leq m$, $m+1 \leq a \leq n$, and $1 \leq A \leq n$. Then, $(q^A, \dot{q}^a)$ are local coordinates for the submanifold M of TQ.

We denote the set of twice differentiable curves connecting two points $x, y \in Q$ as

$$\mathcal{C}^2(x, y) = \{c : [0, 1] \longrightarrow Q \mid c \text{ is } C^2,\ c(0) = x \text{ and } c(1) = y\}.$$

This set is a differentiable infinite-dimensional manifold [3].

Let c be a curve in $\mathcal{C}^2(x,y)$. A variation of c is a curve c_s in $\mathcal{C}^2(x,y)$, that is a differentiable mapping $c_s : (-\epsilon,\epsilon) \to \mathcal{C}^2(x,y)$, $s \mapsto c_s(t)$, such that $c_0 = c$. An infinitesimal variation of c is the tangent vector of a variation of c, that is,

$$u(t) = \frac{dc_s(t)}{ds}\Big|_{s=0} \in T_{c(t)}Q.$$

The tangent space of $\mathcal{C}^2(x,y)$ at c is then given by

$$T_c\mathcal{C}^2(x,y) = \{u : [0,1] \longrightarrow TQ \mid u \text{ is } C^1,\ u(t) \in T_{c(t)}Q,\ u(0) = 0 \text{ and } u(1) = 0\}.$$

Now, we introduce a special subset $\tilde{\mathcal{C}}^2(x,y)$ of $\mathcal{C}^2(x,y)$ which consists of those curves whose velocities belong to the constraint submanifold M:

$$\tilde{\mathcal{C}}^2(x,y) = \{c \in \mathcal{C}^2(x,y) \mid \dot{c}(t) \in M_{c(t)} = M \cap \tau_Q^{-1}(c(t))\ \forall t \in [0,1]\}.$$

Finally, let us consider the action functional $\mathcal{J}$ defined by

$$\mathcal{J} : \mathcal{C}^2(x,y) \longrightarrow \mathbb{R}, \quad c \mapsto \mathcal{J}(c) = \int_0^1 L(\dot{c}(t))\,dt.$$

DEFINITION 2.1. *The vakonomic problem associated with (Q, L, M, x, y) consists of extremizing the functional $\mathcal{J}$ among the curves satisfying the constraints imposed by M, $c \in \tilde{\mathcal{C}}^2(x,y)$. Hence, a curve $c \in \tilde{\mathcal{C}}^2(x,y)$ will be a solution of the vakonomic problem if c is a critical point of $\mathcal{J}_{|\tilde{\mathcal{C}}^2(x,y)}$.*

Remark 2.2. In this paper, we will assume that the solution curves $c \in \tilde{\mathcal{C}}(x,y)$ admit enough nontrivial variations in $\tilde{\mathcal{C}}(x,y)$. These solutions are called normal in the literature, in contrast to the abnormal ones, which are pathological curves which do not admit sufficient nontrivial variations [1]. Several investigators have shown the existence of C^1, stable under perturbations abnormal $\mathcal{J}$-minimizing solutions [37, 45].

Now, we find a characterization for the solutions of the vakonomic problem.

PROPOSITION 2.3. *A curve $c \in \tilde{\mathcal{C}}^2(x,y)$ is a normal solution of the vakonomic problem if and only if there exists $\mu : [0,1] \to \mathbb{R}^m$ such that*

$$\text{(2.2)} \qquad \begin{cases} \dfrac{d}{dt}\left(\dfrac{\partial \tilde{L}}{\partial \dot{q}^a}\right) - \dfrac{\partial \tilde{L}}{\partial q^a} = \mu_\alpha \left[\dfrac{d}{dt}\left(\dfrac{\partial \Psi^\alpha}{\partial \dot{q}^a}\right) - \dfrac{\partial \Psi^\alpha}{\partial q^a}\right] + \dot{\mu}_\alpha \dfrac{\partial \Psi^\alpha}{\partial \dot{q}^a}, \\ \dot{\mu}_\alpha = \dfrac{\partial \tilde{L}}{\partial q^\alpha} - \mu_\beta \dfrac{\partial \Psi^\beta}{\partial q^\alpha}, \\ \dot{q}^\alpha = \Psi^\alpha(q^A, \dot{q}^a), \end{cases}$$

where $\tilde{L} : M \to \mathbb{R}$ is the restriction of L to M.

Proof. The condition for a curve to be a solution of the vakonomic problem is

$$0 = d\mathcal{J}(c) \cdot u = \frac{d}{ds}\mathcal{J}(c_s)\Big|_{s=0},$$

for any variation c_s in $\tilde{\mathcal{C}}^2(x,y)$ of c, where $u = \frac{dc_s}{ds}|_{s=0}$. Then, we have that

$$0 = \frac{d}{ds}\mathcal{J}(c_s)\Big|_{s=0} = \frac{d}{ds}\left(\int_0^1 L(\dot{c}_s(t))\,dt\right)\Big|_{s=0} = \int_0^1 \frac{d}{ds}L(\dot{c}_s(t))\Big|_{s=0} dt.$$

In local coordinates, we obtain

$$
\begin{aligned}
0 &= \int_0^1 \left(\frac{\partial L}{\partial q^A} u^A + \frac{\partial L}{\partial \dot{q}^a} \dot{u}^a + \frac{\partial L}{\partial \dot{q}^\alpha} \frac{\partial \Psi^\alpha}{\partial q^A} u^A + \frac{\partial L}{\partial \dot{q}^\alpha} \frac{\partial \Psi^\alpha}{\partial \dot{q}^a} \dot{u}^a \right) dt \\
&= \int_0^1 \left(\left[\frac{\partial L}{\partial q^A} + \frac{\partial L}{\partial \dot{q}^\alpha} \frac{\partial \Psi^\alpha}{\partial q^A} \right] u^A + \left[\frac{\partial L}{\partial \dot{q}^a} + \frac{\partial L}{\partial \dot{q}^\alpha} \frac{\partial \Psi^\alpha}{\partial \dot{q}^a} \right] \dot{u}^a \right) dt \\
&= \int_0^1 \left(\frac{\partial \tilde{L}}{\partial q^A} u^A + \frac{\partial \tilde{L}}{\partial \dot{q}^a} \dot{u}^a \right) dt.
\end{aligned} \tag{2.3}
$$

From (2.1) we know that the infinitesimal variations u^A, $1 \le A \le n$, are not arbitrary. Consider the functions μ_α defined as the solutions of the following system of first order differential equations

$$
\dot{\mu}_\alpha = \frac{\partial \tilde{L}}{\partial q^\alpha}\Big|_c - \mu_\beta \frac{\partial \Psi^\beta}{\partial q^\alpha}\Big|_c, \qquad 1 \le \alpha \le m.
$$

Then, using the fact that $\dot{u}^\alpha = \frac{\partial \Psi^\alpha}{\partial q^A} u^A + \frac{\partial \Psi^\alpha}{\partial \dot{q}^a} \dot{u}^a$, we get

$$
\frac{d}{dt}(\mu_\alpha u^\alpha) = \mu_\alpha \dot{u}^\alpha + \left(\frac{\partial \tilde{L}}{\partial q^\alpha} - \mu_\beta \frac{\partial \Psi^\beta}{\partial q^\alpha} \right) u^\alpha = u^\alpha \frac{\partial \tilde{L}}{\partial q^\alpha} + \mu_\alpha \frac{\partial \Psi^\alpha}{\partial q^a} u^a + \mu_\alpha \frac{\partial \Psi^\alpha}{\partial \dot{q}^a} \dot{u}^a,
$$

or, equivalently, $u^\alpha \frac{\partial \tilde{L}}{\partial q^\alpha} = \frac{d}{dt}(\mu_\alpha u^\alpha) - \mu_\alpha \frac{\partial \Psi^\alpha}{\partial q^a} u^a - \mu_\alpha \frac{\partial \Psi^\alpha}{\partial \dot{q}^a} \dot{u}^a$. Substituting the last expression in (2.3) and integrating by parts, we obtain

$$
d\mathcal{J}(c) \cdot u = \int_0^1 \left(\left[\frac{\partial \tilde{L}}{\partial q^a} - \mu_\alpha \frac{\partial \Psi^\alpha}{\partial q^a} \right] u^a + \left[\frac{\partial \tilde{L}}{\partial \dot{q}^a} - \mu_\alpha \frac{\partial \Psi^\alpha}{\partial \dot{q}^a} \right] \dot{u}^a \right) dt.
$$

Now, since

$$
\left[\frac{\partial \tilde{L}}{\partial \dot{q}^a} - \mu_\alpha \frac{\partial \Psi^\alpha}{\partial \dot{q}^a} \right] \dot{u}^a = \frac{d}{dt} \left(\left[\frac{\partial \tilde{L}}{\partial \dot{q}^a} - \mu_\alpha \frac{\partial \Psi^\alpha}{\partial \dot{q}^a} \right] u^a \right) - \frac{d}{dt} \left(\frac{\partial \tilde{L}}{\partial \dot{q}^a} - \mu_\alpha \frac{\partial \Psi^\alpha}{\partial \dot{q}^a} \right) u^a,
$$

using again integration by parts, we can write

$$
0 = \int_0^1 \left[\frac{\partial \tilde{L}}{\partial q^a} - \mu_\alpha \frac{\partial \Psi^\alpha}{\partial q^a} - \frac{d}{dt} \left(\frac{\partial \tilde{L}}{\partial \dot{q}^a} - \mu_\alpha \frac{\partial \Psi^\alpha}{\partial \dot{q}^a} \right) \right] u^a \, dt.
$$

As the infinitesimal variations u^a are arbitrary, the fundamental lemma of the calculus of variations applies and we can assert that $d\mathcal{J}(c) \cdot u = 0$ if and only if c and μ_α satisfy (2.2). □

Remark 2.4. The usual way in which the equations of motion for vakonomic mechanics are presented is the following:

$$
\begin{cases}
\dfrac{d}{dt} \left(\dfrac{\partial L}{\partial \dot{q}^A} \right) - \dfrac{\partial L}{\partial q^A} = \dot{\lambda}_\alpha \dfrac{\partial \Phi^\alpha}{\partial \dot{q}^A} + \lambda_\alpha \left[\dfrac{d}{dt} \left(\dfrac{\partial \Phi^\alpha}{\partial \dot{q}^A} \right) - \dfrac{\partial \Phi^\alpha}{\partial q^A} \right], \\
\Phi^\alpha(q, \dot{q}) = 0, \qquad 1 \le \alpha \le m,
\end{cases} \tag{2.4}
$$

where $\Phi^\alpha = \Psi^\alpha - \dot{q}^\alpha$ and $\lambda_\alpha = \frac{\partial L}{\partial \dot{q}^\alpha} - \mu_\alpha$, $1 \leq \alpha \leq m$. Observe that, in contrast to (2.2), (2.4) are expressed in terms of the ambient Lagrangian $L : TQ \to \mathbb{R}$. Equations (2.2) stress how the information given by L outside M is irrelevant to obtain the vakonomic equations, a fact that it is not promptly deduced from (2.4). This is in contrast with what happens in nonholonomic mechanics (see section 5 below).

Equations (2.4) can be seen as the Euler–Lagrange equations for the extended Lagrangian $\mathcal{L} = L + \lambda_\alpha \Phi^\alpha$. We will not follow this approach here, which has been exploited successfully in [20, 28, 42, 43]. Finally, note that if we consider the extended Lagrangian $\lambda_0 L + \lambda_\alpha \Phi^\alpha$, with $\lambda_0 = 1$ or 0, then we recover all the solutions, both the normal and the abnormal ones [1].

3. Geometric approach to vakonomic mechanics. We will develop a geometric characterization of vakonomic mechanics following an approach similar to the formulation given by Skinner and Rusk [58] for singular Lagrangians (see also [15, 22, 38]). This characterization is specially interesting, for it enables us to study both linear and nonlinear constraints in an intrinsic way. Moreover, as we shall discuss later, this formalism will allow to use ideas from geometric mechanics in the treatment of optimal control problems.

Consider the Whitney sum of T^*Q and TQ, $T^*Q \oplus TQ$, and its canonical projections $pr_1 : T^*Q \oplus TQ \longrightarrow T^*Q$, $pr_2 : T^*Q \oplus TQ \longrightarrow TQ$. Let us take the submanifold $W_0 = pr_2^{-1}(M)$, where M is the constraint submanifold, locally determined by the constraint equations $\Phi^\alpha = 0$, $1 \leq \alpha \leq m$. We will denote $W_0 = T^*Q \times_Q M$ and $\pi_1 = pr_{1|W_0}$, $\pi_2 = pr_{2|W_0}$. Now, define on $T^*Q \times_Q M$ the presymplectic 2-form $\omega = \pi_1^* \omega_Q$, where ω_Q is the canonical symplectic form on T^*Q. Observe that the rank of this presymplectic form is equal to $2n$ everywhere. Define also the function

$$H_{W_0} = \langle \pi_1, \pi_2 \rangle - \pi_2^* \tilde{L},$$

where $\langle \cdot, \cdot \rangle$ denotes the natural pairing between vectors and covectors on Q.

If (q^A) are local coordinates on a neighborhood U of Q, $(q^A, \dot{q}^a)$ coordinates on $TU \cap M$, and (q^A, p_A) the induced coordinates on T^*U, then we have induced coordinates $(q^A, p_A, \dot{q}^a)$ on $T^*U \times_Q (TU \cap M)$. Locally, the Hamiltonian function H_{W_0} reads as

$$H_{W_0}(q^A, p_A, \dot{q}^a) = p_a \dot{q}^a + p_\alpha \Psi^\alpha - \tilde{L}(q^A, \dot{q}^a),$$

and the 2-form ω is $\omega = dq^A \wedge dp_A$.

Now, we will see how the dynamics of the vakonomic system (2.2) is determined by the solutions of the equation

$$i_X \omega = dH_{W_0}. \tag{3.1}$$

This then justifies the use of the following terminology.

DEFINITION 3.1. *The presymplectic Hamiltonian system* $(T^*Q \times_Q M, \omega, H_{W_0})$ *will be called vakonomic Hamiltonian system.*

The system $(T^*Q \times_Q M, \omega, H_{W_0})$ being presymplectic, we may apply Gotay Nester's constraint algorithm [21]. First we consider the set of points W_1 of $T^*Q \times_Q M$ where (3.1) has a solution. This first constraint submanifold is determined by

$$W_1 = \{x \in T^*Q \times_Q M \mid dH_{W_0}(x)(V) = 0 \ \forall V \in \ker \omega(x)\}.$$

Locally, $\ker\omega = \operatorname{span}\langle\partial/\partial\dot{q}^a\rangle$. Therefore, the constraint submanifold W_1 is locally characterized by the vanishing of the constraints

$$\varphi_a = p_a + p_\alpha \frac{\partial \Psi^\alpha}{\partial \dot{q}^a} - \frac{\partial \tilde{L}}{\partial \dot{q}^a} = 0, \qquad m+1 \le a \le n,$$

or, equivalently,

$$p_a = \frac{\partial \tilde{L}}{\partial \dot{q}^a} - p_\alpha \frac{\partial \Psi^\alpha}{\partial \dot{q}^a}, \qquad m+1 \le a \le n. \tag{3.2}$$

Expanding the expressions in (3.1) the equations of motion along W_1 are

$$\dot{q}^A = \frac{\partial H_{W_0}}{\partial p_A}, \qquad \dot{p}_A = -\frac{\partial H_{W_0}}{\partial q^A},$$

which is equivalent to

$$\dot{q}^\alpha = \Psi^\alpha(q^A, \dot{q}^a), \tag{3.3}$$

$$\dot{p}_\alpha = \frac{\partial \tilde{L}}{\partial q^\alpha} - p_\beta \frac{\partial \Psi^\beta}{\partial q^\alpha}, \tag{3.4}$$

$$\frac{d}{dt}\left(\frac{\partial \tilde{L}}{\partial \dot{q}^a} - p_\alpha \frac{\partial \Psi^\alpha}{\partial \dot{q}^a}\right) = \frac{\partial \tilde{L}}{\partial q^a} - p_\beta \frac{\partial \Psi^\beta}{\partial q^a}. \tag{3.5}$$

Observe that these equations are precisely the vakonomic equations of motion (2.2), where now $p_\alpha = \mu_\alpha$, $1 \le \alpha \le m$.

Remark 3.2. The momenta p_α, $1 \le \alpha \le m$, play the role of the Lagrange multipliers, but they do not have any physical meaning (see [61]).

Therefore, a vector field X solution of (3.1) will generally be of the form

$$\begin{aligned} X = {} & \dot{q}^a \left(\frac{\partial}{\partial q^a} + \left(\frac{\partial^2 \tilde{L}}{\partial q^a \partial \dot{q}^b} - p_\gamma \frac{\partial^2 \Psi^\gamma}{\partial q^a \partial \dot{q}^b} \right) \frac{\partial}{\partial p_b} \right) \\ & + \Psi^\alpha \left(\frac{\partial}{\partial q^\alpha} + \left(\frac{\partial^2 \tilde{L}}{\partial q^\alpha \partial \dot{q}^b} - p_\gamma \frac{\partial^2 \Psi^\gamma}{\partial q^\alpha \partial \dot{q}^b} \right) \frac{\partial}{\partial p_b} \right) \\ & + \bar{X}^a \left(\frac{\partial}{\partial \dot{q}^a} + \left(\frac{\partial^2 \tilde{L}}{\partial \dot{q}^a \partial \dot{q}^b} - p_\gamma \frac{\partial^2 \Psi^\gamma}{\partial \dot{q}^a \partial \dot{q}^b} \right) \frac{\partial}{\partial p_b} \right) \\ & + \left(\frac{\partial \tilde{L}}{\partial q^\alpha} - p_\beta \frac{\partial \Psi^\beta}{\partial q^\alpha} \right) \left(\frac{\partial}{\partial p_\alpha} - \frac{\partial \Psi^\alpha}{\partial \dot{q}^b} \frac{\partial}{\partial p_b} \right), \end{aligned}$$

where the coefficients $\bar{X}^a$ are still undetermined. The solution on W_1 may not be tangent to W_1. In such a case, we have to restrict W_1 to the submanifold W_2 where this solution is tangent to W_1. Proceeding further, we obtain a sequence of submanifolds (we are assuming that all the subsets generated by the algorithm are submanifolds)

$$\cdots \hookrightarrow W_k \hookrightarrow \cdots \hookrightarrow W_2 \hookrightarrow W_1 \hookrightarrow W_0 = T^*Q \times_Q M.$$

Algebraically, these constraint submanifolds may be described as

$$W_i = \{ x \in T^*Q \times_Q M \mid dH_{W_0}(x)(v) = 0 \ \forall v \in T_x W_{i-1}^\perp \}, \qquad i \ge 1, \tag{3.6}$$

where $T_xW_{i-1}^{\perp} = \{v \in T_x(T^*Q \times_Q M) \mid \omega(x)(u,v) = 0 \ \forall u \in T_xW_{i-1}\}$. If this constraint algorithm stabilizes, i.e., if there exists a positive integer $k \in \mathbb{N}$ such that $W_{k+1} = W_k \neq W_{k-1}$ and $\dim W_k \neq 0$, then we will have obtained a final constraint submanifold $W_f = W_k$ on which a vector field X exists such that

$$(i_X\omega = dH_{W_0})_{|W_f}.$$

Note that on W_f we will have an explicit solution of the vakonomic dynamics. A very important particular case is when the final constraint submanifold is the first one, i.e., $W_f = W_1$. Observe that the dimension of W_1 is even, $\dim W_1 = 2n$. In what follows, we will investigate when this constraint submanifold is equipped with a symplectic 2-form in order to determine a unique solution X of the vakonomic equations. Obviously, this geometrical study is related to the explicit or implicit character of the second order differential equations obtained in (2.2).

Denote by ω_{W_1} the restriction of the presymplectic 2-form ω to W_1.

PROPOSITION 3.3. *(W_1, ω_{W_1}) is a symplectic manifold if and only if for any point of W_1,*

$$\det\left(\frac{\partial^2\tilde{L}}{\partial\dot{q}^a\partial\dot{q}^b} - p_\alpha\frac{\partial^2\Psi^\alpha}{\partial\dot{q}^a\partial\dot{q}^b}\right) \neq 0. \tag{3.7}$$

Proof. ω_{W_1} is symplectic if and only if $T_xW_1 \cap (T_xW_1)^{\perp} = 0$ for all $x \in W_1$. This condition is satisfied if and only if the matrix $d\varphi_a(\frac{\partial}{\partial\dot{q}^b})$ is regular, that is,

$$\det\left(\frac{\partial^2\tilde{L}}{\partial\dot{q}^a\partial\dot{q}^b} - p_\alpha\frac{\partial^2\Psi^\alpha}{\partial\dot{q}^a\partial\dot{q}^b}\right) \neq 0$$

for all $x \in W_1$. □

In this case, (3.5) can be rewritten in explicit form as

$$\ddot{q}^a = -\bar{\mathcal{C}}^{ab}\left[\dot{q}^A\frac{\partial^2\tilde{L}}{\partial q^A\partial\dot{q}^b} - \dot{q}^A p_\alpha\frac{\partial^2\Psi^\alpha}{\partial q^A\partial\dot{q}^b} - \frac{\partial\tilde{L}}{\partial q^b} + p_\alpha\frac{\partial\Psi^\alpha}{\partial q^b} - \left(\frac{\partial\tilde{L}}{\partial q^\gamma} - p_\beta\frac{\partial\Psi^\beta}{\partial q^\gamma}\right)\frac{\partial\Psi^\gamma}{\partial\dot{q}^b}\right], \tag{3.8}$$

where

$$\bar{\mathcal{C}}_{ab} = \frac{\partial^2\tilde{L}}{\partial\dot{q}^a\partial\dot{q}^b} - p_\alpha\frac{\partial^2\Psi^\alpha}{\partial\dot{q}^a\partial\dot{q}^b}, \tag{3.9}$$

and $(\bar{\mathcal{C}}^{ab})$ denotes the inverse matrix of $(\bar{\mathcal{C}}_{ab})$.

Remark 3.4. The characterization found in Proposition 3.3 for the symplectic nature of the manifold (W_1, ω_{W_1}) implies, by the implicit function theorem, that the constraint equations

$$\varphi_a = p_a + p_\alpha\frac{\partial\Psi^\alpha}{\partial\dot{q}^a} - \frac{\partial\tilde{L}}{\partial\dot{q}^a} = 0, \qquad m+1 \leq a \leq n,$$

locally determine the variables $\dot{q}^a$, $m+1 \leq a \leq n$. That is, we have $\dot{q}^a = \varsigma^a(q^A, p_A)$, $m+1 \leq a \leq n$. Therefore, we can also consider local coordinates (q^A, p_A) on W_1. In such a case, the symplectic form and the restriction of the Hamiltonian H_{W_0} to W_1 have the following local expressions:

$$\omega_{W_1} = dq^A \wedge dp_A, \qquad H_{W_1} = p_a\varsigma^a + p_\alpha\Psi^\alpha - \bar{L}(q^A, p_A),$$

where $\bar{L}(q^A, p_A) = \tilde{L}(q^A, \varsigma^a(q^A, p_A))$. Consequently, (3.3)–(3.5) can be rewritten in Hamiltonian form as

$$\dot{q}^A = \frac{\partial H_{W_1}}{\partial p_A}, \qquad \dot{p}_A = -\frac{\partial H_{W_1}}{\partial q^A}.$$

This choice of coordinates is common in optimal control theory.

Now, observe that, if the constraints are linear in the velocities, we can write $\dot{q}^\alpha = \Psi_a^\alpha(q)\dot{q}^a$. Then, from Proposition 3.3, ω_{W_1} is symplectic if and only if

$$\det\left(\frac{\partial^2 \tilde{L}}{\partial \dot{q}^a \partial \dot{q}^b}\right) \neq 0.$$

PROPOSITION 3.5. *Suppose that the constraints are given by* $\dot{q}^\alpha = \Psi_a^\alpha(q)\dot{q}^a$, $1 \leq \alpha \leq m$, *and the Lagrangian* L *is regular. Denote by* (W^{AB}) *the inverse matrix of the Hessian matrix of* L. *In this case,* ω_{W_1} *is symplectic on* W_1 *if and only if the constraints are compatible, that is, the matrix whose entries are*

$$\mathcal{C}^{\alpha\beta} = W^{ab}\Psi_a^\alpha\Psi_b^\beta - W^{\alpha b}\Psi_b^\beta - W^{a\beta}\Psi_a^\alpha + W^{\alpha\beta}$$

is nonsingular.

Proof. See the geometrical proof of Theorem IV.3 in reference [33]. □

Remark 3.6. The compatibility condition guarantees the existence and uniqueness of the solutions for the nonholonomic problem with Lagrangian L and constraint submanifold M [33, 57].

Before ending this section, we would like to make some remarks concerning this geometric approach to vakonomic dynamics. First of all, we must say that it provides an intrinsic formulation of variational problems subject to both linear and *nonlinear* constraints on *manifolds*. In addition, this formulation belongs to the context of Symplectic Geometry and Geometric Mechanics, following previous work by Bloch and Crouch [4, 8, 9], Jurdjevic [23, 24], and others. There is a whole collection of ideas and methods ensuing from these fields that have been used in the treatment of optimal control problems. Apart from being of use as a tool for an algorithmic study of the existence of optimal solutions and their domains of definition, we think that this formulation has something to contribute in at least three directions: the study of the symmetry properties of constrained problems [8, 18, 24, 43] (infinitesimal, Noether and Cartan symmetries, dynamical symmetries, ...), the study of higher order variational problems [6] (since a generalization of our approach to the higher order case seems to be straightforward), and the development of numerical integrators [19, 55, 65, 66, 67] that take into account the geometry of the problem (2-form, Hamiltonian, momentum) and are competitive with the traditional methods.

An immediate outcome of the formulation on $T^*Q \times_Q M$ is that for the study of problems subject to nonlinear constraints we can use similar techniques to those used for the linear case. Finally, this framework will allow us in section 6 to compare vakonomic dynamics with nonholonomic dynamics within a common setting.

In the following, we aim to illustrate some of the above ideas on two examples.

3.1. Applications in economy. The variational calculus is an indispensable tool in many economic problems [25, 39, 52]. In fact, a typical optimization problem in modern economics deals with extremizing the functional

$$\int_0^T D(t)U[f(t, k, \dot{k})]\, dt$$

subject or not to constraints. Here, $D(t)$ is a discount rate factor, U a utility function, f a consumption function, and k the capital-labor ratio. It is common to find dynamical economic models with nonholonomic constraints.

Example 3.7 (closed von Neumann system [53, 54, 56]). Consider the transformation function F on $\mathbb{R}^{2n}$ which relates n capital goods $K_1, K_2, \dots, K_n$ and the net capital formations $\dot{K}_1, \dot{K}_2, \dots, \dot{K}_n$ as

$$F(K_1, \dots, K_n, \dot{K}_1, \dots, \dot{K}_n) = K_1^{\alpha_1} K_2^{\alpha_2} \cdots K_n^{\alpha_n} - \left[\dot{K}_1^2 + \cdots + \dot{K}_n^2\right]^{1/2},$$

with $\alpha_1 + \alpha_2 + \cdots + \alpha_n = 1$. The von Neumann problem consists of maximizing

$$\int_0^T \dot{K}_n \, dt \quad \text{subject to} \quad F(K_1, \dots, K_n, \dot{K}_1, \dots, \dot{K}_n) = 0,$$

with appropriate initial conditions.

Our formalism makes it possible to write this problem as a presymplectic system on $W_0 = \mathbb{R}^{3n-1}$. The constraint $F = 0$ can be rewritten as

$$\dot{K}_1 = \pm \left(K_1^{2\alpha_1} \cdots K_n^{2\alpha_n} - \sum_{i=2}^{n} \dot{K}_i^2 \right)^{1/2} = \pm \Psi(K_1, \dots, K_n, \dot{K}_2, \dots, \dot{K}_n).$$

Here, we restrict the analysis to the component $\dot{K}_1 = \Psi$. Taking coordinates $(K_1, \dots, K_n, \dot{K}_2, \dots, \dot{K}_n, P^1, \dots, P^n)$ we have that

$$\omega = \sum_{j=1}^{n} dK_j \wedge dP^j, \; H_{W_0} = \sum_{i=2}^{n} P^i \dot{K}_i + P^1 \cdot \left(K_1^{2\alpha_1} K_2^{2\alpha_2} \cdots K_n^{2\alpha_n} - \sum_{i=2}^{n} \dot{K}_i^2 \right)^{1/2} - \dot{K}_n.$$

Applying the Gotay and Nester algorithm, new constraints arise,

$$P^i = P^1 \dot{K}_i \left(K_1^{2\alpha_1} K_2^{2\alpha_2} \cdots K_n^{2\alpha_n} - \sum_{i=2}^{n} \dot{K}_i^2 \right)^{-1/2}, \qquad 2 \le i \le n-1,$$

$$P^n = 1 + P^1 \dot{K}_n \left(K_1^{2\alpha_1} K_2^{2\alpha_2} \cdots K_n^{2\alpha_n} - \sum_{i=2}^{n} \dot{K}_i^2 \right)^{-1/2}.$$

Therefore, from (3.3)–(3.5) the initial system is determined by solving the following n differential equations on the variables $(K_1, \dots, K_n, \dot{K}_2, \dots, \dot{K}_n, P^1)$:

(3.10)

$$\begin{cases} \dot{P}^1 = -P^1 \alpha_1 \left(K_1^{2\alpha_1 - 1} K_2^{2\alpha_2} \cdots K_n^{2\alpha_n} \right) G \\ 0 = \dot{P}^1 \dot{K}_i G \\ \qquad + P^1 \left[\left(\ddot{K}_i + \alpha_i \left(K_1^{2\alpha_1} \cdots K_i^{2\alpha_i - 1} \cdots K_n^{2\alpha_n} \right) \right) G + \dot{K}_i \dfrac{d}{dt}(G) \right], \quad 2 \le i \le n, \end{cases}$$

where $G = 1/\Psi$. The presymplectic context for these optimal equations provides us with some new insights into the problem. On the one hand, the existence of well-defined solutions to (3.10) is not guaranteed in general. It can occur, for instance, that an optimal curve starting from a point in W_1 "escapes" from this phase space

after some time because the dynamical vector field is no longer tangent to W_1. But one can indeed eliminate this possibility. Consider the case $n = 2$ for simplicity. Assume $\Psi \neq 0$. Otherwise the dynamics is fully determined and the optimization problem is trivial (we have abnormal solutions). The determinant (3.7) is equal to

$$\frac{P^1}{\Psi^3}(\Psi^2 + \dot{K}_2^2) = \frac{H_{W_1}}{\Psi^2}. \tag{3.11}$$

Therefore, if the optimal curve starts from any point in $x \in W_1$ such that $H_{W_1}(x) \neq 0$, (3.11) guarantees that the dynamics of the vakonomic problem remains tangent to W_1. On the other hand, the optimal solutions with $H_{W_1} = 0$ are stationary curves, $K_1 = \text{const}$, $K_2 = \text{const}$, and $K_1 K_2 = 0$.

This formulation can also shed light on the aspect of symmetries and conservation laws. It is known [53, 54, 56] that the closed von Neumann system possesses, besides the Hamiltonian H, another conservation law, which is usually found by ad hoc methods. However, it is not difficult to define in our context the notion of *Noether symmetry* and verify that the vector field

$$Y = \sum_{j=1}^{n} K_i \frac{\partial}{\partial K^i} \in \mathfrak{X}(Q)$$

indeed corresponds to such a symmetry. The associated conservation law is precisely given by $\Phi = P^1(K_1\Psi + \sum_{j=2}^{n} K_j \dot{K}_j)/\Psi$. In the same way, one can explore the presence of other types of symmetries, like Cartan symmetries, for example, [32, 34, 48].

Finally, obtaining explicit solutions of (3.10) is, in general, a very difficult task. The use of numerical integrators can help in analyzing the behavior of the system. In the last years there has been an increasing activity in the development of integrators that take into account the geometric structures associated with the problem [19, 55, 65, 66, 67]. The proposed formalism offers the possibility of designing such methods for a variety of optimal control problems.

3.2. LC-circuits. The dynamics of nonlinear LC electric circuits [44] can be given a variational interpretation, as discussed in [46]. Here, we treat this class of systems under our vakonomic formalism and study the well-posedness of the optimal equations.

Consider a circuit consisting of capacitors and inductors, which are charge and current controlled. Let $\mathcal{C}$ be the collection of n-capacitor branches and $\mathcal{L}$ the m-inductor branches. Denote by $q \in Q_{\mathcal{C}}$ the vector of capacitor charges and by $i \in Q_{\mathcal{L}}$ the inductor currents. Kirchhoff's current and voltage laws require that $\dot{q} = A_{\mathcal{C}} u$, $i = A_{\mathcal{L}} u$, where $A_{\mathcal{C}}$ and $A_{\mathcal{L}}$ are appropriate linear maps from a vector space $\mathcal{U}$ to $Q_{\mathcal{C}}$ and $Q_{\mathcal{L}}$ characterizing, respectively, the topology of the network and the chosen current reference directions. The new variables $u \in \mathcal{U}$ are usually thought of as a vector of some independent loop currents. The generality of the interconnection structure of the circuit relies on how general the matrices $A_{\mathcal{C}}$, $A_{\mathcal{L}}$ can be. In the following, we will assume that $A_{\mathcal{L}}$ is nonsingular and then the space $\mathcal{U}$ will be identified with $Q_{\mathcal{L}}$ through $A_{\mathcal{L}}$. Finally, denote by $W_e : Q_{\mathcal{C}} \to \mathbb{R}$ the electric energy and by $W_m^* : Q_{\mathcal{L}} \to \mathbb{R}$ the magnetic coenergy of the circuit.

The dynamics of the circuit is governed by the element equations, the equations arising from Kirchhoff's current law, and those arising from Kirchhoff's voltage law.

After some manipulations, these equations may be reduced to

$$\dot{q} = A_{\mathcal{C}} u, \qquad A^*_{\mathcal{L}} \frac{d}{dt} \left(dW^*_m(A_{\mathcal{L}} u) \right) = -A^*_{\mathcal{C}} dW_e(q), \tag{3.12}$$

where the star superscript denotes the transpose of the corresponding matrix operator. However, the well-posedness of this mathematical model for the electric circuit is not guaranteed in general. It could be, for instance, that some specifications of initial conditions $(q(0), u(0))$ turn out to be incompatible with the algebraic constraints embedded in (3.12).

The theoretical setting described above can bring some new insight into this question. Consider as configuration space the product manifold $Q = Q_{\mathcal{C}} \times Q_{\mathcal{L}}$ with coordinates (q^α, u^a). Let $L : TQ \to \mathbb{R}$, $L = W^*_m(A_{\mathcal{L}} u) - W_e(q)$, be the Lagrangian and define $M \subseteq TQ$ by $\dot{q}^\alpha = (A_{\mathcal{C}})^\alpha_b u^b$ as the submanifold of constraints. Then, the dynamics of the LC-circuit is found to be defined on the tertiary constraint submanifold of the presymplectic Hamiltonian system $(T^*Q \times_Q M, \omega, H)$. This means that all initial conditions in W_3 are compatible in the sense of the previous paragraph.

Let $(q^\alpha, u^a, \xi_\alpha, \zeta_a, \dot{u}^a)$ be the local coordinates in $W_0 = T^*Q \times_Q M$. Then,

$$\omega = dq^\alpha \wedge d\xi_\alpha + du^a \wedge d\zeta_a, \qquad H = \zeta_a \dot{u}^a + \xi_\alpha A^\alpha_{\mathcal{C}b} u^b - W^*_m(A_{\mathcal{L}} u) + W_e(q).$$

The first submanifold of constraints is given by $W_1 = \{ x \in W_0 \,|\, dH_x(\frac{\partial}{\partial \dot{u}^a}) = \zeta_a = 0 \}$. After some computations, we find that

$$TW_1 \cap TW_1^\perp = \operatorname{span} \left\{ \frac{\partial}{\partial u^a}, \frac{\partial}{\partial \dot{u}^a} \right\},$$

and hence we must continue with the constraint algorithm. Following (3.6), we have that W_2 is described by the new constraints

$$\frac{\partial H}{\partial u^a} = \left[A^*_{\mathcal{C}} \xi - A^*_{\mathcal{L}} dW^*_m(A_{\mathcal{L}} u) \right]_a = 0. \tag{3.13}$$

Under the additional assumption of invertibility of dW^*_m, or, equivalently, under the assumption that the LC-circuit is also flux controlled, we can ensure that there exists a magnetic energy W_m such that $dW^*_m(A_{\mathcal{L}} u) = \phi \iff A_{\mathcal{L}} u = dW_m(\phi)$. Then, we can rewrite (3.13) as

$$u = A^{-1}_{\mathcal{L}} dW_m((A^{-1}_{\mathcal{L}})^* A^*_{\mathcal{C}} \xi) \equiv F(\xi)$$

and consider $(q^\alpha, \xi_\alpha, \dot{u}^a)$ as a set of coordinates on W_2. The following step of the algorithm leads us to the constraints

$$\frac{\partial H}{\partial \zeta_a} + \frac{\partial F^a}{\partial \xi_\alpha} \frac{\partial H}{\partial q^\alpha} = \dot{u}^a + \frac{\partial F^a}{\partial \xi_\alpha} dW_e(q) = 0.$$

In this way, we have (q^α, ξ_α) as coordinates on W_3, which turns out to be the final constraint submanifold. The dynamics of the system is described on W_3 by the differential equations

$$\dot{q}^\alpha = A_{\mathcal{C}} F(\xi), \qquad \dot{\xi}_\alpha = -dW_e(q). \tag{3.14}$$

Thus, the application of the algorithm allows us to say that, under the given assumptions, the initial conditions in W_3 provide us with consistent optimal solutions of the dynamics of the LC-circuit.

There are, of course, other optimal control problems that can be interpreted in a vakonomic setting and for which this formulation can be of some help. We mention here the optimal control for nonholonomic systems with symmetry, with interesting applications to the locomotion of kinematic and mixed kinematic and dynamic systems [18, 28, 49] or sub-Riemannian geometry [11].

4. Comparison of the Vershik–Gershkovich and the vakonomic Hamiltonian approaches. In the preceding section we have found an intrinsic geometric approach to vakonomic dynamics. It is possible to give an alternative geometric formulation of the vakonomic equations of motion, related to the one of Vershik and Gershkovich [64]. A key element to obtain this alternative description will be the next fibered morphism

$$\begin{array}{rccc} F: & T^*Q \oplus TQ & \longrightarrow & T^*Q \oplus TQ, \\ & (\alpha, v) & \longmapsto & (\alpha - Leg_L(v), v), \end{array}$$

for any $\alpha \in T_x^*Q$, $v \in T_xQ$, and $x \in Q$. Here, $Leg_L : TQ \to T^*Q$ denotes the Legendre transformation associated with the Lagrangian L, which in local coordinates reads $Leg_L(q^A, \dot{q}^A) = (q^A, \frac{\partial L}{\partial \dot{q}^A})$. It is clear that $F(T^*Q \times_Q M) = T^*Q \times_Q M$. We will see how in the case of linear constraints, we "recover" the Vershik–Gershkovich formulation. As a by-product, we will have obtained a generalization of their formulation to the case of nonlinear constraints.

Consider on $T^*Q \oplus TQ$ the presymplectic 2-form $\Omega = pr_1^*\omega_Q$. Let $\omega_L = -dS^*dL$ be the Poincaré–Cartan 2-form on TQ associated with $L : TQ \to \mathbb{R}$ and E_L its energy function. Take also the presymplectic 2-form $pr_2^*\omega_L$ on $T^*Q \oplus TQ$, and define the functions

$$H = \langle pr_1, pr_2 \rangle - pr_2^*L, \qquad \bar{H} = \langle pr_1, pr_2 \rangle - pr_2^*E_L.$$

LEMMA 4.1. *The morphism $F : T^*Q \oplus TQ \to T^*Q \oplus TQ$ is a presymplectomorphism from $(T^*Q \oplus TQ, \Omega)$ onto $(T^*Q \oplus TQ, \Omega + pr_2^*\omega_L)$, i.e., $F^*(\Omega + pr_2^*\omega_L) = \Omega$. Moreover, it verifies $F^*\bar{H} = H$.*

Proof. F is clearly invertible with inverse

$$\begin{array}{rccc} F^{-1}: & T^*Q \oplus TQ & \longrightarrow & T^*Q \oplus TQ, \\ & (\alpha, v) & \longmapsto & (\alpha + Leg(v), v). \end{array}$$

A direct computation shows that $H \circ F^{-1} = \bar{H}$. Moreover, in local coordinates,

$$(F^{-1})^*(dq^A \wedge dp_A) = dq^A \wedge \left[dp_A + d\left(\frac{\partial L}{\partial \dot{q}^A} \right) \right] = dq^A \wedge dp_A + dq^A \wedge d\left(\frac{\partial L}{\partial \dot{q}^A} \right),$$

which implies $F^*(\Omega + pr_2^*\omega_L) = \Omega$. □

Denote by $j : T^*Q \times_Q M \hookrightarrow T^*Q \oplus TQ$ and $i : M \hookrightarrow TQ$ the respective canonical inclusions. Let us define $\bar{\omega} = j^*(\Omega + pr_2^*\omega_L)$. Since $pr_2 \circ j = i \circ \pi_2$, we have that

$$\bar{\omega} = \omega + (i \circ \pi_2)^*\omega_L.$$

PROPOSITION 4.2. *The solutions of the equations*

$$i_X\omega = dH_{W_0} \tag{4.1}$$

and

$$i_Y\bar{\omega} = d(j^*\bar{H}) \tag{4.2}$$

are $F_{|W_0}$-related; that is, if $x \in T^*Q \times_Q M$ is a point where a solution Y of (4.2) exists, then $TF^{-1}(Y)$ is a solution of (4.1) at $F^{-1}(x)$ and, conversely, if X is a solution of (4.1) at $F^{-1}(x)$, then $TF(X)$ is a solution of (4.2) at x.

Proof. The proof readily follows from Lemma 4.1. □

An immediate consequence is the following corollary.

COROLLARY 4.3. *F preserves the constraint submanifolds provided by the presymplectic systems $(T^*Q \times_Q M, \omega, H_{W_0})$ and $(T^*Q \times_Q M, \bar{\omega}, j^*\bar{H})$. That is, if*

$$\cdots \hookrightarrow W_k \cdots \hookrightarrow W_1 \hookrightarrow W_0 = T^*Q \times_Q M \quad \textit{and}$$

$$\cdots \hookrightarrow P_k \cdots \hookrightarrow P_1 \hookrightarrow P_0 = T^*Q \times_Q M$$

are the sequences of submanifolds generated by Gotay and Nester's algorithm for the first and the second presymplectic Hamiltonian system, respectively, then $F_i = F_{|W_i} : W_i \longrightarrow P_i$, are diffeomorphisms for all i.

In conclusion, Proposition 4.2 and Corollary 4.3 show that solving the vakonomic Hamiltonian equations (4.1) as in section 3 is equivalent to solving (4.2). Locally, if $(q^A(t), p_A(t), \dot{q}^a(t))$ is an integral curve of X, then

$$\left(q^A(t), p_A - i^* \frac{\partial L}{\partial \dot{q}^A}(q^B(t), \dot{q}^b(t)), \dot{q}^a(t) \right)$$

is an integral curve of Y.

4.1. Vershik–Gershkovich approach. In [64], Vershik and Gershkovich gave a formulation for the "nonholonomic variational problem," i.e., the vakonomic problem, within the framework of the so-called mixed bundle picture, which we briefly review in the following (see also [7]).

If $\mathcal{D} : Q \longrightarrow TQ$ is a differentiable distribution along Q, then the mixed bundle over Q associated with $\mathcal{D}$ is given by $\mathcal{D} \oplus \mathcal{D}^o$, where $\mathcal{D}^o$ is the codistribution annihilating $\mathcal{D}$; the fibers of $\mathcal{D} \oplus \mathcal{D}^o \longrightarrow Q$ are $\mathcal{D}_q \oplus \mathcal{D}_q^o$.

Let $\{ \Phi^\alpha(q^A, \dot{q}^A) = \Psi_a^\alpha(q)\dot{q}^a - \dot{q}^\alpha, 1 \leq \alpha \leq m \}$ be a set of independent functions whose annihilation defines the distribution $\mathcal{D}$, and let $\{ \eta^\alpha = \Psi_a^\alpha dq^a - dq^\alpha, 1 \leq \alpha \leq m \}$ be the corresponding basis of $\mathcal{D}^o$. Regarding $\mathcal{D} \subset TQ$ as the set of admissible velocities, Vershik and Gershkovich write the equations of motion (2.4) for the vakonomic problem $(L, \mathcal{D})$ as follows:

$$\text{(4.3)} \qquad \begin{cases} \left(\dfrac{d}{dt}\left(\dfrac{\partial L}{\partial \dot{q}^A} \right) - \dfrac{\partial L}{\partial q^A} \right) dq^A = \dot{\lambda}_\alpha \eta^\alpha + \lambda_\alpha (i_{\dot{q}} d\eta^\alpha), \\ \\ \langle \dot{q}, \eta^\alpha \rangle = 0, \qquad 1 \leq \alpha \leq m. \end{cases}$$

In this particular case, we obtain that P_1, the first constraint submanifold for the presymplectic Hamiltonian system $(T^*Q \times_Q M, \bar{\omega}, j^*\bar{H})$, is just $\mathcal{D}^o \oplus \mathcal{D}$, since we get $\lambda_a + \lambda_\alpha \Psi_a^\alpha = 0,\ 1 \leq \alpha \leq m$.

If $(P_1 = \mathcal{D}^o \oplus \mathcal{D}, \omega_{P_1})$ is a symplectic manifold (see Proposition 3.5), then the equations of motion (4.3) determine a unique vector field on $\mathcal{D}^o \oplus \mathcal{D}$ and the Lagrange multipliers λ_α are coordinates in $\mathcal{D}^o$ with respect to the basis η^α.

Consequently, the geometrical picture we have developed in section 3 is equivalent to the Vershik–Gershkovich approach. As said above, we have obtained a generalization of the Vershik–Gershkovich formulation to the case of nonlinear constraints, just "translating" things from our approach by the diffeomorphism F.

In the nonlinear case, under the admissibility condition, one can verify that the first constraint submanifold $P_1 = F(W_1)$ can be identified with the manifold $S^*(TM^o) \times_Q M$. In fact, we have that $S^*(TM^o)$ is generated by the 1-forms

$$S^* d\Phi^\alpha = dq^\alpha - \frac{\partial \Psi^\alpha}{\partial \dot{q}^a} dq^a, \qquad 1 \leq \alpha \leq m.$$

If $(q^A, \lambda_A, \dot{q}^a) \in P_1$, then the 1-form $\lambda_A dq^A$ is a linear combination of the 1-forms $S^* d\Phi^\alpha$ in the following manner: $\lambda_A dq^A = \lambda_\alpha S^* d\Phi^\alpha$.

5. Geometric approach to nonholonomic mechanics. A nonholonomic Lagrangian system consists of a Lagrangian $L : TQ \to \mathbb{R}$ subject to nonholonomic constraints defined by m local functions $\Phi^\alpha(q^A, \dot{q}^A)$, $1 \leq \alpha \leq m$. The equations of motion for nonholonomic mechanics are derived assuming that the constraints satisfy d'Alembert's principle, in the linear or affine case. In the nonlinear case, there does not seem to exist a general consensus concerning the correct principle to adopt [41, 51]. The most widely used model is based on Chetaev's principle, which will also be adopted in the present paper. The equations of motion are then given by

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}^A}\right) - \frac{\partial L}{\partial q^A} = \lambda_\alpha \frac{\partial \Phi^\alpha}{\partial \dot{q}^A}, \tag{5.1}$$

together with the algebraic equations $\Phi^\alpha(q^A, \dot{q}^A) = 0$. The functions λ_α, $1 \leq \alpha \leq m$, are some Lagrange multipliers to be determined. As in the vakonomic case, we assume the admissibility condition, so it is possible to write the constraints as $\dot{q}^\alpha = \Psi^\alpha(q^A, \dot{q}^a)$, where $1 \leq \alpha \leq m$, $m+1 \leq a \leq n$, and $1 \leq A \leq n$.

The study of nonholonomic systems in the realm of geometric mechanics started with the work by Vershik and Faddeev [62, 63] and has been an active area of research since then, with many contributions from different authors (see [16] for a recent survey). In particular, the role of symmetry has been treated extensively in the literature, starting with the work by Koiller [27] and going through the use of the Hamiltonian formalism [2], Lagrangian reduction [10], the geometry of the tangent bundle [12, 13, 17, 33], or Poisson methods [40], among others.

Nonholonomic mechanics also admits a nice geometrical description on the space $T^*Q \oplus TQ$ inspired on the Skinner and Rusk formalism [58]. In addition, this novel description will be appropriate to compare the solutions of the dynamics between the vakonomic and nonholonomic mechanics. In the following, we will prove that (5.1) can be intrinsically written as

$$\begin{cases} (i_X \Omega - dH)_{|T^*Q \times_Q M} \in F^o, \\ X_{|T^*Q \times_Q M} \in T(T^*Q \times_Q M), \end{cases} \tag{5.2}$$

where Ω is the presymplectic 2-form $\Omega = pr_1^* \omega_Q$ on $T^*Q \oplus TQ$, H the Hamiltonian function $H = \langle pr_1, pr_2 \rangle - pr_2^* L$, and F^o the subbundle of $T^*(T^*Q \oplus TQ)$ along $T^*Q \times_Q M$ defined by $F^o = pr_2^*(S^*(TM^o))$, representing the constraint forces.

Indeed, we have in local coordinates

$$\Omega = dq^A \wedge dp_A, \qquad dH = \dot{q}^A dp_A + p_A d\dot{q}^A - \frac{\partial L}{\partial q^A} dq^A - \frac{\partial L}{\partial \dot{q}^A} d\dot{q}^A,$$

and F^o is generated by the 1-forms

$$\frac{\partial \Phi^\alpha}{\partial \dot{q}^A} dq^A = \frac{\partial \Psi^\alpha}{\partial \dot{q}^a} dq^a - dq^\alpha, \qquad 1 \leq \alpha \leq m.$$

If $X = X^A \frac{\partial}{\partial q^A} + Y^A \frac{\partial}{\partial \dot{q}^A} + Z_A \frac{\partial}{\partial p_A}$ was a solution of (5.2), then

$$X^A = \dot{q}^A, \qquad Z_A = \frac{\partial L}{\partial q^A} + \lambda_\alpha \frac{\partial \Phi^\alpha}{\partial \dot{q}^A}, \tag{5.3}$$

along with the constraints

$$p_A - \frac{\partial L}{\partial \dot{q}^A} = 0, \qquad \Phi^\alpha(q^A, \dot{q}^A) = 0. \tag{5.4}$$

Observe that these constraints determine a submanifold $\tilde{M}$ of $T^*Q \times_Q M$. The submanifold $\tilde{M}$ is diffeomorphic to M since

$$\begin{array}{ccc} M & \longrightarrow & \tilde{M}, \\ m & \longmapsto & (Leg_L(m), m) \end{array}$$

is a diffeomorphism. $\tilde{M}$ is the first constraint submanifold provided by the constraint algorithm applied to (5.2). This algorithm will lead to a final constraint submanifold on which there exists a well-defined dynamics. Obviously, (5.3) and (5.4) are equivalent to the nonholonomic equations of motion (5.1).

In terms of the Ψ^α's the above equations can be written as

$$X^A = \dot{q}^A, \quad Z_a = \frac{\partial L}{\partial q^a} + \lambda_\alpha \frac{\partial \Psi^\alpha}{\partial \dot{q}^a}, \quad Z_\beta = \frac{\partial L}{\partial q^\beta} - \lambda_\beta,$$

together with the constraints

$$p_A - \frac{\partial L}{\partial \dot{q}^A} = 0, \qquad \dot{q}^\alpha - \Psi^\alpha(q^A, \dot{q}^a) = 0. \tag{5.5}$$

Therefore, a solution X of (5.2) is of the form

$$\begin{aligned} X = {} & \dot{q}^a \left(\frac{\partial}{\partial q^a} + \frac{\partial \Psi^\alpha}{\partial q^a} \frac{\partial}{\partial \dot{q}^\alpha} + \left(\frac{\partial^2 L}{\partial \dot{q}^A \partial q^a} + \frac{\partial \Psi^\alpha}{\partial q^a} \frac{\partial^2 L}{\partial \dot{q}^A \partial \dot{q}^\alpha} \right) \frac{\partial}{\partial p_A} \right) \\ & + \Psi^\gamma \left(\frac{\partial}{\partial q^\gamma} + \frac{\partial \Psi^\alpha}{\partial q^\gamma} \frac{\partial}{\partial \dot{q}^\alpha} + \left(\frac{\partial^2 L}{\partial \dot{q}^A \partial q^\gamma} + \frac{\partial \Psi^\alpha}{\partial q^\gamma} \frac{\partial^2 L}{\partial \dot{q}^A \partial \dot{q}^\alpha} \right) \frac{\partial}{\partial p_A} \right) \\ & + Y^a \left(\frac{\partial}{\partial \dot{q}^a} + \frac{\partial \Psi^\alpha}{\partial \dot{q}^a} \frac{\partial}{\partial \dot{q}^\alpha} + \left(\frac{\partial^2 L}{\partial \dot{q}^A \partial \dot{q}^a} + \frac{\partial \Psi^\alpha}{\partial \dot{q}^a} \frac{\partial^2 L}{\partial \dot{q}^A \partial \dot{q}^\alpha} \right) \frac{\partial}{\partial p_A} \right). \end{aligned}$$

Under the regularity assumption, which here means that the matrix

$$\tilde{\mathcal{C}}_{ab} = \frac{\partial^2 \tilde{L}}{\partial \dot{q}^a \partial \dot{q}^b} - i^* \left(\frac{\partial L}{\partial \dot{q}^\alpha} \right) \frac{\partial^2 \Psi^\alpha}{\partial \dot{q}^a \partial \dot{q}^b} \tag{5.6}$$

is invertible (see [57]), there is a unique solution of the dynamics on $\tilde{M}$. In particular, after some computations we obtain

$$Y^a = -\tilde{\mathcal{C}}^{ab} \left[\dot{q}^A \frac{\partial^2 \tilde{L}}{\partial q^A \partial \dot{q}^b} - \dot{q}^A i^* \left(\frac{\partial L}{\partial \dot{q}^\alpha} \right) \frac{\partial^2 \Psi^\alpha}{\partial q^A \partial \dot{q}^b} - \frac{\partial \tilde{L}}{\partial q^b} + i^* \left(\frac{\partial L}{\partial q^\alpha} \right) \left(\frac{\partial \Psi^\alpha}{\partial q^b} - \frac{\partial \Psi^\alpha}{\partial \dot{q}^b} \right) \right],$$

where $i : M \to TQ$ is the canonical inclusion and $(\tilde{\mathcal{C}}^{ab})$ the inverse matrix of $(\tilde{\mathcal{C}}_{ab})$.

Taking coordinates $(q^A, \dot{q}^a)$ on $\tilde{M}$, the equations of motion for a nonholonomic system will be

(5.7)
$$\begin{cases} \dot{q}^\alpha = \Psi^\alpha(q^A, \dot{q}^a), \\ \ddot{q}^a = -\mathcal{C}^{ab}\left[\dot{q}^A \dfrac{\partial^2 \tilde{L}}{\partial q^A \partial \dot{q}^b} - \dot{q}^A i^*\left(\dfrac{\partial L}{\partial \dot{q}^\alpha}\right) \dfrac{\partial^2 \Psi^\alpha}{\partial q^A \partial \dot{q}^b} - \dfrac{\partial \tilde{L}}{\partial q^b} + i^*\left(\dfrac{\partial L}{\partial q^\alpha}\right)\left(\dfrac{\partial \Psi^\alpha}{\partial q^b} - \dfrac{\partial \Psi^\alpha}{\partial \dot{q}^b}\right)\right]. \end{cases}$$

6. Vakonomic and nonholonomic mechanics: Equivalence of dynamics. In this section, we shall investigate the relation between vakonomic and nonholonomic dynamics. Consider a physical system with Lagrangian $L : TQ \to \mathbb{R}$ and constraint submanifold $M \subset TQ$. Let us assume that the vakonomic problem lives on the first constraint submanifold, W_1, and that the nonholonomic one lives on $\tilde{M}$ (this will be the case if the constraints are linear and the admissibility and compatibility conditions are satisfied). As a consequence, we have well-defined vector fields X_{vk} on W_1 and X_{nh} on $\tilde{M}$. It is clear that the mapping $(\pi_2)_{|W_1} : W_1 \to M$ is a surjective submersion and that we can define the mapping $\Upsilon : W_1 \to \tilde{M}$ as

$$\begin{array}{rccc} \Upsilon : & W_1 & \longrightarrow & \tilde{M}, \\ & (\alpha, v) & \longmapsto & (Leg_L(v), v). \end{array}$$

In coordinates, Υ reads as $\Upsilon(q^A, \dot{q}^a, p_\alpha) = (q^A, \dot{q}^a)$.

Our aim is to know whether, given a solution of the nonholonomic problem, we can find initial conditions in the vakonomic Lagrange multipliers, p_α, so that the curve can also be seen as a solution of the vakonomic problem. In order to capture the common solutions to both systems, we have developed the following algorithm. It is inspired on the idea of the Υ-relation of X_{vk} and X_{nh} and the constraint algorithm developed by Krupková [30]. If both fields were Υ-related, then the projection to $\tilde{M}$ of all the vakonomic solutions would be nonholonomic. So, selecting those points where both vector fields are related, we are picking up all the possible good candidates. We write $W_1 = S_0$ and define

$$S_1 = \{w \in S_0 \mid T_w\Upsilon(X_{vk}(w)) = X_{nh}(\Upsilon(w))\}.$$

In general, S_1 is not a submanifold. If $S_1 = \emptyset$, there is no relation between the vakonomic and nonholonomic dynamics. If $S_1 \neq \emptyset$, we apply the following algorithm:

- **Step 1**: For any $w \in S_1$, consider $C_{(w)} = \cup_i C_{(w)i}$, the union of all connected submanifolds $C_{(w)i}$ of maximal dimension lying in S_1, contained in a neighborhood U of w, and passing through w. (Maximal dimension means that if N is a connected submanifold lying in $S_1 \cap U$ passing through w and $C_{(w)i} \subseteq N$, then $C_{(w)i} = N$.)
  Suppose that $C_{(w)} \neq \{w\}$. For each i we consider the subset of $C_{(w)i}$,

  $$\tilde{C}_{(w)i} = \{v \in C_{(w)i} \mid X_{vk}(v) \in T_v C_{(w)i}\}.$$

  If $\tilde{C}_{(w)i} = C_{(w)i}$ then we call the submanifold $C_{(w)i}$ a *final constraint submanifold at* w. If $\tilde{C}_{(w)i} = \emptyset$, we exclude $C_{(w)i}$ from the collection $C_{(w)}$. If $\emptyset \subsetneq \tilde{C}_{(w)i} \subsetneq C_{(w)i}$, then we proceed to the next step.
- **Step 2**: Repeat the Step 1 with $\tilde{C}_{(w)i}$ instead of S_1.

After a sufficient number of steps in this algorithm either we obtain a collection of final constraint submanifolds at w or we find that there is no final constraint submanifold passing through w. Collecting all the points where there exist such final constraint submanifolds, we obtain the subset of W_1 where there is equivalence between vakonomic and nonholonomic dynamics.

Suppose that the constraints Φ^α, $1 \leq \alpha \leq m$, are linear in the velocities so we can write them as $\dot{q}^\alpha = \Psi_a^\alpha(q)\dot{q}^a$. In such a case, the matrices $\mathcal{C}$ and $\tilde{\mathcal{C}}$ defined in (3.9) and (5.6), respectively, are the same (even for constraints affine in the velocities).

PROPOSITION 6.1. *S_1 is locally characterized by the vanishing of the $n-m$ constraints functions on W_1:*

(6.1)
$$g_b = \dot{q}^a \left(p_\alpha - i^* \frac{\partial L}{\partial \dot{q}^\alpha} \right) \left[\frac{\partial \Psi_b^\alpha}{\partial q^a} - \frac{\partial \Psi_a^\alpha}{\partial q^b} + \Psi_a^\beta \frac{\partial \Psi_b^\alpha}{\partial q^\beta} - \Psi_b^\beta \frac{\partial \Psi_a^\alpha}{\partial q^\beta} \right], \qquad m+1 \leq b \leq n.$$

Proof. The comparison between the vector fields X_{vk} and X_{nh} consists of taking the difference between $\ddot{q}^a$'s in the expressions (3.8) and (5.7) and equating the result to zero. □

Consider the local projection $\rho(q^a, q^\alpha) = (q^\alpha)$ and the connection Γ on ρ such that the horizontal distribution $\mathcal{H}$ is given by prescribing its annihilator to be $\mathcal{H}^o = \langle dq^\alpha - \Psi_a^\alpha dq^a, 1 \leq \alpha \leq m \rangle$. Then the curvature R of this connection (see [34]) is given by $R(\frac{\partial}{\partial q^a}, \frac{\partial}{\partial q^b}) = R_{ab}^\alpha \frac{\partial}{\partial q^\alpha}$, where

$$R_{ab}^\alpha = \frac{\partial \Psi_b^\alpha}{\partial q^a} - \frac{\partial \Psi_a^\alpha}{\partial q^b} + \Psi_a^\beta \frac{\partial \Psi_b^\alpha}{\partial q^\beta} - \Psi_b^\beta \frac{\partial \Psi_a^\alpha}{\partial q^\beta}.$$

We say that Γ is flat if the curvature R vanishes identically. The tensor R measures the lack of integrability of the horizontal distribution $\mathcal{H}$, which in our case is the constraint manifold. Then, we can write the constraints determining S_1 as

$$g_b = \dot{q}^a \left(p_\alpha - i^* \frac{\partial L}{\partial \dot{q}^\alpha} \right) R_{ab}^\alpha, \qquad m+1 \leq b \leq n.$$

From this expression we deduce that if the constraints are holonomic, then $R = 0$ and the final constraint submanifold is equal to $S_0 = W_1$. Therefore, every solution of the nonholonomic problem is also a vakonomic solution. Indeed, (3.3)–(3.5) will read as

(6.2)
$$\begin{cases} \dot{q}^\alpha = \Psi_a^\alpha \dot{q}^a, \\ \dot{p}_\alpha = \dfrac{\partial \tilde{L}}{\partial q^\alpha} - p_\beta \dfrac{\partial \Psi_a^\beta}{\partial q^\alpha} \dot{q}^a, \\ \dfrac{d}{dt}\left(\dfrac{\partial \tilde{L}}{\partial \dot{q}^a} \right) - \dfrac{\partial \tilde{L}}{\partial q^a} = \Psi_a^\alpha \dfrac{\partial \tilde{L}}{\partial q^a}. \end{cases}$$

The first and the third set of equations determine the trajectory in M. The Lagrange multipliers p_α are determined by the second set of equations once we know the solution in M. This is the typical behavior of the holonomic case [1, 36]. But, in general, for linear constraints, the first constraint subset in the algorithm is determined by

$$S_1 = \{ g_b = 0, \ m+1 \leq b \leq n \},$$

where $g_b(q^A, \dot{q}^a, p_\alpha) = \dot{q}^a R_{ab}^\alpha(q)(p_\alpha - \frac{\partial L}{\partial \dot{q}^\alpha})$. Note that S_1 will not be a submanifold, because 0 is not a regular value of the functions g_b, $b = m+1, \ldots, n$. Anyway, the

geometric context we have developed can be very useful to tackle the problem of the comparison of the two methods.

PROPOSITION 6.2. *If $c(t) = (q^A(t))$ is a solution of the unconstrained problem which, in addition, verifies all the constraints, i.e.,*

$$\dot{q}^\alpha(t) = \Psi_a^\alpha(q(t))\dot{q}^a(t), \qquad 1 \leq \alpha \leq m,$$

then $c(t)$ is a solution of the nonholonomic and vakonomic problems simultaneously.

Proof. Let us consider the submanifold $S := \{p_\alpha = i^*(\frac{\partial L}{\partial \dot{q}^\alpha})\}$, which is contained in S_1. A natural question is whether the vakonomic vector field will be tangent to S, that is, $X_{vk} \in TS$. From (3.3)–(3.5), we have along any integral curve of the vakonomic vector field

$$X_{vk} \in TS \Longleftrightarrow \frac{d}{dt}\left(p_\alpha - i^*\frac{\partial L}{\partial \dot{q}^\alpha}\right) = 0 \Longleftrightarrow \dot{p}_\alpha = \dot{q}^A \frac{\partial^2 L}{\partial q^A \partial \dot{q}^\alpha} + \ddot{q}^a \frac{\partial^2 L}{\partial \dot{q}^a \partial \dot{q}^\alpha}.$$

On S, we have that

$$\dot{p}_\alpha = \frac{\partial \tilde{L}}{\partial q^\alpha} - p_\beta \frac{\partial \Psi^\beta}{\partial q^\alpha} = \frac{\partial \tilde{L}}{\partial q^\alpha} - \frac{\partial L}{\partial \dot{q}^\beta}\frac{\partial \Psi^\beta}{\partial q^\alpha} = \frac{\partial L}{\partial q^\alpha}.$$

Then the above condition can be rewritten as

$$\frac{\partial L}{\partial q^\alpha} = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}^\alpha}\right),$$

that with (3.5) are precisely the Euler–Lagrange equations. Then, we have proved that a solution $c(t)$ of the unconstrained problem satisfies the constraints if and only if

$$\left(q^A(t), i^*\left(\frac{\partial L}{\partial \dot{q}^\alpha}\right), \dot{q}^a(t)\right)$$

is a solution of the vakonomic equations (3.3). Since the constraints $g_b = 0$ are automatically satisfied for all the points in S we deduce that $c(t)$ is also a solution of the nonholonomic problem. □

Remark 6.3. As a consequence of Proposition 6.2 we obtain that if g is a Riemannian metric on Q, with kinetic energy $L = \frac{1}{2}g$, and if we assume that we are given a distribution $\mathcal{D}$ on Q which is geodesically invariant with respect to the Levi–Civita connection ∇^g, then all the nonholonomic solutions can be seen as vakonomic ones. In fact, they all are solutions of the free problem. This last result was first stated in [20, Theorem 3.2] with additional hypothesis on the nature of the metric g and the integrability of $\mathcal{D}^{\perp_g}$, which are not essential, as we have seen.

Remark 6.4. Let $\Theta : G \times Q \longrightarrow Q$ be a free and proper action on Q. Then $\pi : Q \longrightarrow Q/G$ is a principal G-bundle. Assume that the Lagrangian $L : TQ \longrightarrow \mathbb{R}$ is G-invariant and is subject to equivariant affine constraints, M, such that its linear part $\mathcal{D}$ is the horizontal distribution of a principal connection γ on $\pi : Q \longrightarrow Q/G$. Then, we have the following result, which is an adaptation of Theorem 3.1 in [20] to our geometric description of vakonomic and nonholonomic mechanics.

PROPOSITION 6.5. *Assume that the admissibility and compatibility conditions hold. Then, the following are equivalent:*

1. *The solution of the nonholonomic problem $(q^A(t), \dot{q}^a(t)) \in \tilde{M}$ verifies the condition $g_b(q^A(t), \dot{q}^a(t), p_0) = 0$ for some p_0, $m+1 \leq b \leq n$.*

2. *The curve* $(q^A(t), \dot{q}^a(t), p_0) \in W_1$ *is a vakonomic solution.*

Example 6.6 (rolling penny [4]). Consider a vertical penny constrained to roll without slipping on a horizontal plane. Let (x, y) denote the position of contact of the disk in the plane, θ the orientation of a chosen material point P with respect to the vertical, and ϕ the heading angle of the penny. The configuration space is then $Q = \mathbb{R}^2 \times \mathbb{S}^1 \times \mathbb{S}^1$. The Lagrangian may be written as $L = (\dot{x}^2 + \dot{y}^2 + \dot{\theta}^2 + \dot{\phi}^2)/2$ and the constraints are given by $\dot{x} = \dot{\theta}\cos\phi$, $\dot{y} = \dot{\theta}\sin\phi$. For simplicity, we assume that the mass m, the moments of inertia I, J, and the radius of the penny R are 1.

Applying the algorithm, we obtain the final constraint submanifolds

$$C_{f_1} = \{w \in W_1 \mid \dot{\phi} = 0\}, \qquad C_{f_2} = \{w \in W_1 \mid 2\dot{\theta} = p_x \cos\phi + p_y \sin\phi\},$$
$$C_{f_3} = \{w \in W_1 \mid \dot{\theta} = 0,\ \dot{\phi} = 0\}.$$

The nonholonomic solutions living on C_{f_1} are motions of the penny along a straight line in the horizontal plane. The nonholonomic solutions in C_{f_3} are stationary positions. However, any nonholonomic solution can be seen as a vakonomic one contained in C_{f_2}, with Lagrange multipliers $p_x = 2\dot{\theta}\cos\phi$ and $p_y = 2\dot{\theta}\sin\phi$. In terms of the extended Lagrangian formalism mentioned in Remark 2.4, we have the following Lagrange multipliers:

$$\lambda_x = \frac{\partial L}{\partial x} - p_x = \dot{x} - p_x = -\dot{\theta}\cos\phi, \qquad \lambda_y = \frac{\partial L}{\partial y} - p_y = \dot{y} - p_y = -\dot{\theta}\sin\phi,$$

which is just the result of Bloch and Crouch [4].

Example 6.7 (planar mobile robot). Consider the motion of a two-wheeled planar mobile robot which is able to move in the direction in which it points and, in addition, can spin about a vertical axis [26, 29, 33]. Let P be the intersection point of the horizontal symmetry axis of the robot and the horizontal line connecting the centers of the two wheels. The position and orientation of the robot is determined by $(x, y, \theta) \in SE(2)$, where $\theta \in \mathbb{S}^1$ is the heading angle and the coordinates $(x, y) \in \mathbb{R}^2$ locate the point P. Let ψ_1, $\psi_2 \in \mathbb{S}^1$ denote the rotation angles of the wheels which are assumed to be controlled independently and roll without slipping on the floor. The configuration space of this system is $Q = \mathbb{S}^1 \times \mathbb{S}^1 \times SE(2)$.

The Lagrangian function is the kinetic energy of the system

$$L = \frac{1}{2}m(\dot{x}^2 + \dot{y}^2) + m_0 l\dot{\theta}(\cos\theta\dot{y} - \sin\theta\dot{x}) + \frac{1}{2}J\dot{\theta}^2 + \frac{1}{2}J_2(\dot{\psi}_1^2 + \dot{\psi}_2^2),$$

where $m = m_0 + 2m_1$, m_0 is the mass of the robot without the wheels, J its moment of inertia with respect to the vertical axis, m_1 the mass of each wheel, J_2 the axial moments of inertia of the wheels, and l the distance between the center of mass C of the robot and the point P. The constraints are induced by the conditions that there is no lateral sliding of the robot and that the motion of the wheels also consists of a rolling without sliding,

$$\dot{x} = -R\cos\theta(\dot{\psi}_1 + \dot{\psi}_2)/2, \quad \dot{y} = -R\sin\theta(\dot{\psi}_1 + \dot{\psi}_2)/2, \quad \dot{\theta} = R(\dot{\psi}_2 - \dot{\psi}_1)/(2c),$$

where R is the radius of the wheels and $2c$ the lateral length of the robot.

This example is very interesting because its qualitative behavior changes completely depending on the parameters. If $l = 0$ (namely, the point P is the center of mass of the robot), application of the algorithm yields the following constraint

submanifolds:

$$C_{f_1} = \{w \in W_1 \mid p_x \sin\theta - p_y \cos\theta = 0,\ \dot{\psi}_1 = \dot{\psi}_2\},$$
$$C_{f_2} = \{w \in W_1 \mid p_x = 0,\ p_y = 0\}, \qquad C_{f_3} = \{w \in W_1 \mid \dot{\psi}_1 = 0,\ \dot{\psi}_2 = 0\}.$$

If $l \neq 0$ and $K_1 \neq K_2^2$, with $K_1 = J_2(J_2 + mR^2/2 + R^2J/2c^2) + mR^3J/4c^2$, $K_2 = m_0lR^2/2c$, we find that

$$C_{f_1} = \{w \in W_1 \mid p_x \sin\theta - p_y \cos\theta = 0,\ \dot{\psi}_1 = \dot{\psi}_2\},$$
$$C_{f_2} = \{w \in W_1 \mid \dot{\psi}_1 = 0,\ \dot{\psi}_2 = 0\},$$

whereas if $K_1 = K_2^2$, we obtain an additional final constraint submanifold

$$C_{f_3} = \{w \in W_1 \mid p_x = K_2(\dot{\psi}_1 - \dot{\psi}_2)\sin\theta/R - 2K_2^2(\dot{\psi}_1 + \dot{\psi}_2)\cos\theta/R(2K_1/J_2 - mR^2),$$
$$p_y = -2K_2^2(\dot{\psi}_1 + \dot{\psi}_2)\sin\theta/R(2K_1/J_2 - mR^2) - K_2(\dot{\psi}_1 - \dot{\psi}_2)\cos\theta/R\}.$$

Therefore, in the cases $l = 0$ and $l \neq 0$, $K_1 = K_2^2$, every nonholonomic solution can be seen as a vakonomic one. This has the following interesting physical interpretation: under an appropriate design of the robot (i.e., choice of the parameters), the trajectories that it describes between two points are optimal, in the sense that they minimize the energy cost among all the other possible trajectories satisfying the constraints and connecting the given points.

Example 6.8 (ball on a rotating table [36]). Applying the algorithm to this example, one can recover the result found in [36]. The configuration space is $Q = \mathbb{R}^2 \times SO(3)$ with coordinates (x, y, R). We denote the spatial angular velocity by $\xi \in \mathbb{R}^3$, where $\hat{\xi} = \dot{R}R^T$. The Lagrangian is $L = I((\xi^1)^2 + (\xi^2)^2 + (\xi^3)^2)/2 + m(\dot{x}^2 + \dot{y}^2)/2$, where I and m are the moment of inertia and mass of the ball, respectively. The constraints are $\dot{x} = r\xi^2 - \Omega y$, $\dot{y} = -r\xi^1 + \Omega x$, where r is the radius of the ball and Ω is the angular velocity of the table.

Applying the algorithm, one finds the following final constraint submanifolds

$$C_{f_1} = \{w \in W_1 \mid \dot{x} = \dot{y} = p_x = p_y = 0\}, \qquad C_{f_2} = \{w \in W_1 \mid \xi^3 = \Omega\}.$$

There are nonholonomic solutions that can not be seen as vakonomic ones (see [36]).

Acknowledgments. J. Cortés, S. Martínez, and D. Martín de Diego would like to thank the Department of Mathematical Physics and Astronomy of the University of Ghent for its kind hospitality. We would like to thank F. Cantrijn and E. Martínez for several helpful conversations and the referees for their careful reading of the manuscript and their useful remarks that helped us improve the quality of the paper.

REFERENCES

[1] V.I. Arnold, *Dynamical Systems*, Vol. III, Springer–Verlag, New York, Heidelberg, Berlin, 1988.

[2] L. Bates and J. Śniatycki, *Nonholonomic reduction*, Rep. Math. Phys., 32 (1992), pp. 99–115.

[3] E. Binz, J. Śniatycki, and H. Fisher, *The Geometry of Classical Fields*, North–Holland Math. Stud. 154, North–Holland, Amsterdam, 1988.

[4] A.M. Bloch and P.E. Crouch, *Nonholonomic and vakonomic control systems on Riemannian manifolds*, in Dynamics and Control of Mechanical Systems, Michael J. Enos, ed., Fields Inst. Commun. 1, AMS, Providence, RI, 1993, pp. 25–52.

[5] A.M. Bloch and P.E. Crouch, *Nonholonomic control systems on Riemannian manifolds*, SIAM J. Control Optim., 33 (1995), pp. 126–148.

[6] A.M. Bloch and P.E. Crouch, *On the equivalence of higher order variational problems and optimal control problems*, in Proceedings of the IEEE International Conference on Decision and Control, Kobe, Japan, 1996, pp. 1648–1653.

[7] A.M. Bloch and P.E. Crouch, *Newton's law and integrability of nonholonomic systems*, SIAM J. Control Optim., 36 (1998), pp. 2020–2039.

[8] A.M. Bloch and P.E. Crouch, *Optimal Control, Optimization, and Analytical Mechanics*, in Mathematical Control Theory, J. Baillieul and J.C. Willems, eds., Springer–Verlag, New York, 1998, pp. 268–321.

[9] A.M. Bloch and P.E. Crouch, *Constrained variational principles on manifolds*, in Proceedings of the IEEE International Conference on Decision and Control, Phoenix, AZ, 1999, pp. 1–6.

[10] A.M. Bloch, P.S. Krishnaprasad, J.E. Marsden, and R.M. Murray, *Nonholonomic mechanical systems with symmetry*, Arch. Ration. Mech. Anal., 136 (1996), pp. 21–99.

[11] R.W. Brockett, *Control Theory and Singular Riemannian Geometry*, in New Directions in Applied Mathematics, P.J. Hilton and G.S. Young, eds., Springer–Verlag, New York, 1982, pp. 11–27.

[12] F. Cantrijn, J. Cortés, M. de León, and D. Martín de Diego, *On the geometry of generalized Chaplygin systems*, Math. Proc. Cambridge Philos. Soc., 132 (2002), pp. 323–351.

[13] F. Cantrijn, M. de León, J.C. Marrero, and D. Martín de Diego, *Reduction of nonholonomic mechanical systems with symmetries*, Rep. Math. Phys., 42 (1998), pp. 25–45.

[14] F. Cardin and M. Favretti, *On nonholonomic and vakonomic dynamics of mechanical systems with nonintegrable constraints*, J. Geom. Phys., 18 (1996), pp. 295–325.

[15] J.F. Cariñena, C. López, and M.F. Rañada, *Geometric Lagrangian approach to first-order systems and applications*, J. Math. Phys., 29 (1988), pp. 1134–1142.

[16] H. Cendra, J.E. Marsden, and T.S. Ratiu, *Geometric mechanics, Lagrangian reduction and nonholonomic systems*, in Mathematics Unlimited-2001 and Beyond, B. Enguist and W. Schmid, eds., Springer–Verlag, New York, 2001, pp. 221–273.

[17] J. Cortés and M. de León, *Reduction and reconstruction of the dynamics of nonholonomic systems*, J. Phys. A, 32 (1999), pp. 8615–8645.

[18] J. Cortés and S. Martínez, *Optimal control for nonholonomic systems with symmetry*, in Proceedings of the IEEE International Conference on Decision and Control, Sydney, Australia, 2000, pp. 5216–5218.

[19] J. Cortés and S. Martínez, *Nonholonomic integrators*, Nonlinearity, 14 (2001), pp. 1365–1392.

[20] M. Favretti, *Equivalence of dynamics for nonholonomic systems with transverse constraints*, J. Dynam. Differential Equations, 10 (1998), pp. 511–536.

[21] M. Gotay and J. Nester, *Presymplectic Lagrangian systems* I: *The constraint algorithm and the equivalence theorem*, Ann. Inst. H. Poincaré Sect. A (N.S.), 30 (1978), pp. 129–142.

[22] A. Ibort and J. Marín-Solano, *A Geometric Approach to Optimal Control Theory and the Inverse Problem of the Calculus of Variations*, preprint, 1998; available upon request from jmarin@eco.ub.es, Departament de Matemàtica Econòmica Financera i Actuarial, Universitat de Barcelona, Barcelona.

[23] V. Jurdjevic, *Geometric Control Theory*, Cambridge Stud. Adv. Math. 52, Cambridge University Press, Cambridge, UK, 1997.

[24] V. Jurdjevic, *Optimal control, geometry and mechanics*, in Mathematical Control Theory, J. Baillieul and J.C. Willems, eds., Springer–Verlag, New York, 1998, pp. 227–267.

[25] H. Kataoka and H. Hashimoto, *New conservation laws in a neoclassical von Neumann model*, J. Math. Econom., 36 (1995), pp. 271–280.

[26] S.D. Kelly and R.M. Murray, *Geometric phases and robotic locomotion*, J. Robotic Systems, 12 (1995), pp. 417–431.

[27] J. Koiller, *Reduction of some classical non-holonomic systems with symmetry*, Arch. Ration. Mech. Anal., 118 (1992), pp. 113–148.

[28] W.-S. Koon and J.E. Marsden, *Optimal control for holonomic and nonholonomic mechanical systems with symmetry and Lagrangian reduction*, SIAM J. Control Optim., 35 (1997), pp. 901–929.

[29] V.V. Kozlov, *Realization of nonintegrable constraints in classical mechanics*, Dokl. Akad. Nauk SSSR, 272 (1983), pp. 550–554 (in Russian); Sov. Phys. Dokl., 28 (1983), pp. 735–737 (in English).

[30] O. Krupková, *The Geometry of Ordinary Variational Equations*, Lectures Notes in Math. 1678, Springer–Verlag, New York, Heidelberg, Berlin, 1997.

[31] M. de León, J.C. Marrero, and D. Martín de Diego, *Vakonomic mechanics versus nonholonomic mechanics: A unified geometrical approach*, J. Geom. Phys., 35 (2000), pp. 126–144.
[32] M. de León and D. Martín de Diego, *Symmetries and constants of the motion for singular Lagrangian systems*, Internat. J. Theoret. Phys., 35 (1996), pp. 975–1011.
[33] M. de León and D. Martín de Diego, *On the geometry of non-holonomic Lagrangian systems*, J. Math. Phys., 37 (1996), pp. 3389–3414.
[34] M. de León and P.R. Rodrigues, *Methods of Differential Geometry in Analytical Mechanics*, North–Holland Math. Stud. 152, North–Holland, Amsterdam, 1989.
[35] A.D. Lewis, *Simple mechanical control systems with constraints*, IEEE Trans. Automat. Control, 45 (2000), pp. 1420–1436.
[36] A.D. Lewis and R.M. Murray, *Variational principles for constrained systems: Theory and experiments*, Internat. J. Non-linear Mech., 30 (1995), pp. 793–815.
[37] W. Liu and H.J. Sussmann, *Abnormal sub-Riemannian minimizers*, in Differential Equations, Dynamical Systems, and Control Science, Lecture Notes in Pure and Appl. Math. 152, Marcel Dekker, New York, 1994, pp. 705–716.
[38] C. López and E. Martínez, *Sub-Finslerian metric associated to an optimal control system*, SIAM J. Control Optim., 39 (2000), pp. 798–811.
[39] M.J.P. Magill, *On a General Economic Theory of Motion*, Springer–Verlag, Berlin, 1970.
[40] C.-M. Marle, *Reduction of constrained mechanical systems and stability of relative equilibria*, Comm. Math. Phys., 174 (1995), pp. 295–318.
[41] C.-M. Marle, *Various approaches to conservative and nonconservative nonholonomic systems*, Rep. Math. Phys., 42 (1998), pp. 211–229.
[42] S. Martínez, J. Cortés, and M. de León, *The geometrical theory of constraints applied to the dynamics of vakonomic mechanical systems*: *The vakonomic bracket*, J. Math. Phys, 41 (2000), pp. 2090–2120.
[43] S. Martínez, J. Cortés, and M. de León, *Symmetries in vakonomic dynamics*: *Applications to optimal control*, J. Geom. Phys., 38 (2001), pp. 343–365.
[44] B.M. Maschke, A.J. van der Schaft, and P.C. Breedveld, *An intrinsic Hamiltonian formulation of the dynamics of LC-circuits*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 42 (1995), pp. 73–82.
[45] R. Montgomery, *Abnormal minimizers*, SIAM J. Control Optim., 32 (1994), pp. 1605–1620.
[46] L. Moreau and D. Aeyels, *A variational principle for nonlinear LC-circuits with arbitrary interconnection structure*, in Proceedings of the IFAC Symposium on Nonlinear Control Systems, 2001, pp. 54–59.
[47] R.M. Murray and S.S. Sastry, *Nonholonomic motion planning: Steering using sinusoids*, IEEE Trans. Automat. Control, 38 (1993), pp. 700–716.
[48] P.J. Olver, *Applications of Lie groups to Differential Equations*, Grad. Texts in Math. 107, Springer–Verlag, New York, 1986.
[49] J.P. Ostrowski, *Optimal controls for kinematic systems on Lie groups*, IFAC World Congress, Beijing, China, 1999.
[50] J.P. Ostrowski and J.W. Burdick, *The geometric mechanics of undulatory robotic locomotion*, Internat. J. Robotics Res., 17 (1998), pp. 683–702.
[51] Y. Pironneau, *Sur les liaisons non holonomes non linéaires, déplacements virtuels à travail nul, conditions de Chetaev*, Atti Accad. Sci. Torino Cl. Sci. Fis. Mat. Natur., 117 (1983), pp. 671–686.
[52] F.P. Ramsey *A mathematical theory of saving*, Economic J., 38 (1928), pp. 543–559.
[53] P.A. Samuelson, *Law of conservation of the capital-output ratio*, Proc. Natl. Acad. Sci. USA, 67 (1970), pp. 1477–1479.
[54] P.A. Samuelson, *Law of conservation of the capital-output ratio in closed von Neumann systems*, in Conservation Laws and Symmetry, Kluwer Academic Publishers, Boston, 1990, pp. 53–56.
[55] J.M. Sanz-Serna and M. Calvo, *Numerical Hamiltonian Problems*, Chapman and Hall, London, 1994.
[56] R. Sato and R.V. Ramachandran, eds., *Conservation Laws and Symmetry: Applications to Economics and Finance*, Kluwer Academic Publishers, Boston, 1990.
[57] D.J. Saunders, F. Cantrijn, and W. Sarlet, *Regularity aspects and Hamiltonisation of non-holonomic systems*, J. Phys. A, 32 (1999), pp. 6869–6890.
[58] R. Skinner and R. Rusk, *Generalized Hamiltonian dynamics* I. *Formulation on* $T^*Q \oplus TQ$, J. Math. Phys., 24 (1983), pp. 2589–2594.
[59] H.J. Sussmann, ed., *Nonlinear Controllability and Optimal Control*, Marcel Dekker, New York, 1990.
[60] H.J. Sussmann, *Geometry and optimal control*, in Mathematical Control Theory, J. Baillieul

and J.C. Willems, eds., Springer–Verlag, New York, 1998, pp. 140–198.

[61] I.A. Taimanov, *Integrable geodesics flows of non-holonomic metrics*, J. Dynam. Control Systems, 3 (1997), pp. 129–147.

[62] A.M. Vershik and L.D. Faddeev, *Differential geometry and Lagrangian mechanics with constraints*, Sov. Phys. Dokl., 17 (1972), pp. 34–36.

[63] A.M. Vershik and L.D. Faddeev, *Lagrangian mechanics in invariant form*, Sel. Math. Sov., 1 (1981), pp. 339–350.

[64] A.M. Vershik and V.Ya. Gershkovich, I. *Nonholonomic Dynamical Systems, Geometry of Distributions and Variational Problems*, Dynamical Systems VII, Springer–Verlag, New York, Heidelberg, Berlin, 1994.

[65] A.P. Veselov, *Integrable discrete-time systems and difference operators*, Funktsional Anal. i Prilozhen., 22 (1988), pp. 1–13.

[66] A.P. Veselov, *Integrable Lagrangian correspondences and the factorization of matrix polynomials*, Funktsional Anal. i Prilozhen., 25 (1991), pp. 38–49.

[67] J.M. Wendlandt and J.E. Marsden, *Mechanical integrators derived from a discrete variational principle*, Phys. D, 106 (1997), pp. 223–246.

[68] G. Zampieri, *Nonholonomic versus vakonomic dynamics*, J. Differential Equations, 163 (2000), pp. 335–347.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 5, pp. 1413–1439

ALGEBRAIC AND SPECTRAL PROPERTIES OF GENERAL TOEPLITZ MATRICES*

A. BULTHEEL† AND P. CARRETTE‡

Abstract. We consider problems related to the generalization of the classical Fourier basis to a basis of rational functions with prescribed poles outside the unit disk. We give some generalizations about the convergence and estimation of the Fourier coefficients with respect to this generalized basis. We also consider a rational generalization of the classical Toeplitz matrices and consider the asymptotic distribution of their spectrum. These bases and general Toeplitz matrices were considered by Ninness et al. in the context of least-squares system estimation where the prescribed poles allow to incorporate a priori knowledge into the system dynamics of the model.

Key words. general Toeplitz matrix, orthogonal rational functions, system identification

AMS subject classifications. 42C05, 42C15, 47B35, 93E12

PII. S0363012900377183

1. Introduction.

It is well known that $\{e^{j n\omega}\}_{n\in\mathbb{Z}}$ forms an orthonormal basis in the Hilbert space $\mathcal{L}_2(\mathbb{T})$ of square integrable functions on the unit circle. This basis is also complete in the space $\mathcal{C}_{2\pi}$ of 2π-periodic functions. Moreover, if such a function $f \in \mathcal{C}_{2\pi}$ has absolutely integrable nth derivative, then its Fourier expansion $f(\omega) = \sum_{r\in\mathbb{Z}} c_r e^{jr\omega}$ with $c_r = \langle f, e^{jr\omega}\rangle$ converges uniformly and $|c_{\pm r}| \le \|f^{(n)}\|_1/r^n$ for $r > 0$.

For $f \in \mathcal{C}_{2\pi}$, one can define the Toeplitz matrices associated with the symbol f as

$$M_p(f) = \frac{1}{2\pi}\int_0^{2\pi} \Gamma_p(\omega) f(\omega) \Gamma_p^*(\omega) d\omega, \qquad p \ge 1,$$

where $\Gamma_p(\omega) = [1,\ e^{-j\omega}, \dots, e^{-j(p-1)\omega}]^*$. (The superscript * denotes complex conjugate transpose.) This integral can be approximated by a quadrature formula whose nodes are, for example, the pth roots of unity $\omega_0, \dots, \omega_{p-1}$. This leads to an approximation of the form $\Upsilon_p F_p \Upsilon_p^*$, where Υ_p is a FFT matrix with entries $(\Upsilon_p)_{k,l} = e^{jk\omega_l}/\sqrt{p}$, for $k, l = 0, \dots, p-1$, and $F_p = \mathrm{diag}(f(\omega_0), \dots, f(\omega_{p-1}))$. A bound for the approximation error in this matrix problem can be used in the study of the following asymptotic problems for the matrices.

For general functions $f \in \mathcal{C}_{2\pi}$, the convergence of the Fourier series may not be uniform, and then other summation techniques like Cesàro means are used. It takes only a little thought to see that such a sum can be written as (see, for example, [14])

$$\frac{1}{p}\Gamma_p^*(\omega) M_p(f) \Gamma_p(\omega) = \sum_{k=-(p-1)}^{p-1} \left(1 - \frac{|k|}{p}\right) c_k e^{j\omega k},$$

*Received by the editors August 29, 2000; accepted for publication (in revised form) June 27, 2002; published electronically January 3, 2003. This work is partially supported by the Belgian Program on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture. The work was also partially supported by Vetenskapsrådet (Swedish Research Council) when P. Carrette was making a post-doctoral visit to the Automatic Control Department of Linköping University (Sweden). The scientific responsibility rests with the authors.

http://www.siam.org/journals/sicon/41-5/37718.html

†Department of Computer Science, K. U. Leuven, Celestijnenlaan 200 A, B-3001 Leuven, Belgium (adhemar.bultheel@cs.kuleuven.ac.be).

‡Shell Oil Company, 3333 Highway 6 South, Houston, TX 77082 (pcarrette@EquilonTech.com).

where the c_k are the Fourier coefficients of f. We know that the limit as $p \to \infty$ in the right-hand side (RHS) gives $f(\omega)$, so that we know the limit in the left-hand side (LHS). This leads to the study of asymptotics for bilinear forms

$$\lim_{p\to\infty} \left(\frac{1}{p}\right) \Gamma_p^*(\sigma) M_p(f) \Gamma_p(\mu),$$

or even more general forms where $M_p(f)$ is replaced by $T(M_p(f))$ with T being an analytic function. Even more general forms are needed. For example, when $M_p(f)$ is replaced by $M_p(f)M_p(g)$, then the limit gives $f(\omega)g(\omega)$ for $\sigma = \mu = \omega$ under appropriate conditions for f and g. So we should be interested in replacing $M_p(f)$ by some analytic *multivariate* function $T(M_p(f^{[1]}), \ldots, M_p(f^{[p]}))$, involving several, possibly different, Toeplitz matrices.

In many applications (see, e.g., [14] and the references mentioned there) it is important to know the asymptotic distribution of the spectrum of these Toeplitz matrices. It is well known that under appropriate conditions on f the eigenvalues of $M_p(f)$ are in the range of f for all p. In this paper we shall generalize results such as the fact that the average of the spectrum of $M_p(f)$ converges to the average value of f, a result due to Grenander and Szegő [11, p. 65].

In this paper we discuss technical results concerning general Toeplitz matrices related to the unifying construction of orthonormal bases presented in [13, 14]. The generalization consists of using rational basis functions where it is allowed to have poles somewhere in the complex plane outside the unit circle. The orthonormal basis functions $\mathrm{e}^{\mathrm{j}n\omega}$ of classical Fourier analysis shall be replaced by orthonormal functions of the form (1.2) below, where the ξ_i are assumed to be arbitrary points in the open unit disk. The motivation for this generalized basis and for the interest in generalizing the above-mentioned properties can be found in papers by Ninness and coworkers [12, 13, 14, 15]; see also [5].

The properties in their generalized form will be summarized in the next section and will be analyzed and proved in the subsequent ones. But let us first recall some of the notation used in [14].

We consider the complex unit circle $\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$ and the Hilbert space $\mathcal{L}_2(\mathbb{T})$ of square integrable complex functions defined on $\mathbb{T}$ with inner product

$$\langle F, G\rangle = \frac{1}{2\pi}\int_0^{2\pi} F(\mathrm{e}^{\mathrm{j}\omega})\overline{G(\mathrm{e}^{\mathrm{j}\omega})}\mathrm{d}\omega = \langle f, g\rangle = \frac{1}{2\pi}\int_0^{2\pi} f(\omega)\overline{g(\omega)}\mathrm{d}\omega, \tag{1.1}$$

where the overbar stands for complex conjugation and $f(\omega) = F(\mathrm{e}^{\mathrm{j}\omega})$ and $g(\omega) = G(\mathrm{e}^{\mathrm{j}\omega})$. For simplicity reasons, we shall identify the function $F(z)$ for $z = \mathrm{e}^{\mathrm{j}\omega} \in \mathbb{T}$ with the 2π-periodic function $f(\omega)$, i.e., $f(\omega) = F(\mathrm{e}^{\mathrm{j}\omega})$. We shall, for example, abuse the notation and write $f \in \mathcal{L}_2(\mathbb{T})$.

Now, with $\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\}$, let $\mathcal{H}_2(\mathbb{D}) \subset \mathcal{L}_2(\mathbb{T})$ be the Hardy space of complex functions analytic in $\mathbb{D}$, i.e., for which all the negative Fourier coefficients vanish: $\mathcal{H}_2(\mathbb{D}) = \{f \in \mathcal{L}_2(\mathbb{T}) : \langle f, \mathrm{e}^{-\mathrm{j}k\omega}\rangle = 0,\ k = 1, 2, \ldots\}$. Obviously, $\|f\|_2 < \infty$ for all $f \in \mathcal{H}_2(\mathbb{D})$ where $\|\cdot\|_2$ is the norm induced by the scalar product in (1.1). Also here we identified the 2π-periodic boundary function $f(\omega) = F(\mathrm{e}^{\mathrm{j}\omega})$ and the function F, depending on a complex variable (which is perfectly justified if F is continuous in $\mathbb{D} \cup \mathbb{T}$).

The basis function $\mathcal{B}_n(z)$ with integer $n \geq 0$ is defined as follows:

$$\mathcal{B}_n(z) = \frac{\sqrt{1-|\xi_n|^2}}{1-\overline{\xi}_n z}\prod_{k=0}^{n-1}\frac{z-\xi_k}{1-\overline{\xi}_k z}, \qquad z \in \mathbb{C}, \tag{1.2}$$

where it is assumed that $|\xi_n| \leq c < 1$ for all $n \geq 0$. This ξ_n is called the nth zero. Let us assume that $\xi_0 = 0$, so that $\mathcal{B}_0$ is just the constant 1. Hereby we guarantee that $\{\mathcal{B}_n\}_{n\geq 0}$ can be a basis for $\mathcal{H}_2(\mathbb{D})$ given some technical condition (see below). It is called the Malmquist basis [16, p. 227]. Note that if all the ξ_i are equal to zero, then $\mathcal{B}_n(z) = z^n$ and we are back in the classical Fourier case. It is easily seen that these general basis functions satisfy $\langle \mathcal{B}_n, \mathcal{B}_m \rangle = \delta_{n-m}$ with δ_i, the Kronecker symbol. In other words, they form an orthonormal set.

An important property of the basis functions $\mathcal{B}_n$ is that the set $\{\mathcal{B}_n(\mathrm{e}^{\mathrm{j}\omega}) : n \geq 0\}$ constitutes an orthonormal basis for $\mathcal{H}_2(\mathbb{D})$ if and only if $\sum_{n\geq 0}(1 - |\xi_n|) = \infty$. This means that, under this condition, the following expansion converges on $\mathbb{T}$:

$$f(\omega) = \sum_{n=0}^{\infty} \langle f, \mathcal{B}_n \rangle \, \mathcal{B}_n(\mathrm{e}^{\mathrm{j}\omega}),$$

where f is any function in $\mathcal{H}_2(\mathbb{D})$. This means that the orthonormal basis $\{\mathcal{B}_n(z), n \geq 0\}$ is complete in $\mathcal{H}_2(\mathbb{D})$. For a proof see [1, p. 244], [5, Chap. 7], or [13, App. C].

Moreover, from these references, it also follows that, for any function $f \in \mathcal{L}_2(\mathbb{T})$, the following expansion holds:

$$f(\omega) = \sum_{n\in\mathbb{Z}} \langle f, \mathcal{B}_n \rangle \, \mathcal{B}_n(\mathrm{e}^{\mathrm{j}\omega}), \quad \text{where} \quad \mathcal{B}_{-n}(\mathrm{e}^{\mathrm{j}\omega}) = \overline{\mathcal{B}_n(\mathrm{e}^{\mathrm{j}\omega})}. \tag{1.3}$$

This means that the infinite set $\{\mathcal{B}_k : k \in \mathbb{Z}\}$ constitutes an orthonormal basis for $\mathcal{L}_2(\mathbb{T})$ under the same condition on the zeros ξ_n; see also [3]. One of our concerns will be to prove convergence and order of convergence properties of the above series given some smoothness conditions of the function f, generalizing the classical Fourier results that we mentioned in the beginning of this section.

However, our main concern will be the spectral properties of general Toeplitz matrices. First we introduce the vector

$$\Gamma_p(\omega) = \left[\overline{\mathcal{B}_0(\mathrm{e}^{\mathrm{j}\omega})}, \ldots, \overline{\mathcal{B}_{p-1}(\mathrm{e}^{\mathrm{j}\omega})}\right]^*, \tag{1.4}$$

where the superscript * stands for the conjugate transpose. Then a general Toeplitz matrix is defined as

$$M_p(f) = \frac{1}{2\pi} \int_0^{2\pi} \Gamma_p(\omega) f(\omega) \Gamma_p^*(\omega) \mathrm{d}\omega, \tag{1.5}$$

where $f(\omega)$ is a 2π-periodic function. It is easily seen that, if $f(\omega)$ is real-valued, then the matrix $M_p(f)$ is Hermitian, i.e., $M_p(f) = M_p^*(f)$.

Let $\gamma_p(\omega) = \Gamma_p^*(\omega)\Gamma_p(\omega)$ be the squared 2-norm of the vector Γ_p. It will be shown that the set of associated normalized vectors

$$\tilde{\Gamma}_p(\omega) = \frac{\Gamma_p(\omega)}{\gamma_p^{1/2}(\omega)}, \tag{1.6}$$

evaluated at frequencies ω_i (with $0 \leq i < p$) for which the phase of the Blaschke product

$$\varphi_p(z) = \prod_{k=0}^{p-1} \frac{z - \xi_k}{1 - \bar{\xi}_k z} \tag{1.7}$$

is constant (modulo 2π), induces an orthonormal basis in $\mathbb{C}^p$. The matrix containing these normalized vectors as columns will be denoted Υ_p, i.e., the ith column is $(\Upsilon_p)_i = \tilde{\Gamma}_p(\omega_i)$ with $\omega_i \in [0, 2\pi)$. This notion generalizes the classical situation where the ω_i are the roots of unity (or a constant rotation thereof) and the matrices Υ_p are the (unitary) FFT matrices.

Here, we intend to study the approximation of this general Toeplitz matrix $M_p(f)$ by the matrix $\Upsilon_p F_p \Upsilon_p^*$ with $F_p = \operatorname{diag}(f_p(\omega_0), \ldots, f_p(\omega_{p-1}))$, where $f_p(\omega)$ is the truncated expansion

$$f_p(\omega) = \sum_{|n|<p} \langle f, \mathcal{B}_n \rangle \, \mathcal{B}_n(\mathrm{e}^{\mathrm{j}\omega}). \tag{1.8}$$

The tool for this will be quadrature formulas on the unit circle. Indeed, we can consider

$$\Upsilon_p F_p \Upsilon_p^* = \sum_{i=0}^{p-1} \tilde{\Gamma}_p(\omega_i) f_p(\omega_i) \tilde{\Gamma}_p^*(\omega_i) \tag{1.9}$$

as a quadrature formula with nodes ω_i for the integral (1.5). It is a special case of a rational Szegő formula as studied in [3, 4, 5, 6].

These results will be used in analyzing the asymptotics of bilinear forms whose matrix is a generalized Toeplitz matrix. For reasons that we have given above in the classical Toeplitz case, it is useful to give more general versions of this problem. In summary, it will be shown that

$$\lim_{p\to\infty} \tilde{\Gamma}_p^*(\sigma) T(M_p(f^{[1]}), \ldots, M_p(f^{[n]})) \tilde{\Gamma}_p(\mu) = \begin{cases} \left[T(f^{[1]}, \ldots, f^{[n]})(\mu)\right], & \text{if } \sigma = \mu, \\ 0, & \text{otherwise,} \end{cases} \tag{1.10}$$

where $T(\cdot)$ is an n-variable analytic function (i.e., having convergent Taylor expansion) on the range of the $f^{[k]}$'s and these $f^{[k]}$'s are 2π-periodic functions satisfying appropriate smoothness assumptions that, at least, make the associated error functions $e_{p,k}(\omega) = f^{[k]}(\omega) - f_p^{[k]}(\omega)$ converge to zero uniformly.

Concerning the asymptotics of the average of the spectrum, it will be furthermore shown that if $f(\omega)$ is also positive (real-valued), then it turns out that the eigenvalue distribution of $M_p(f)$ is given by the distribution of the underlying function, i.e.,

$$\lim_{p\to\infty} \frac{1}{p} \sum_{i=0}^{p-1} T\left(\lambda_i\left(M_p(f)\right)\right) = \frac{1}{2\pi} \int_0^{2\pi} T\left(f\left(\tilde{\chi}^{-1}(\omega)\right)\right) \mathrm{d}\omega, \tag{1.11}$$

where now $T(\cdot)$ stands for a continuous function on the range of $f(\omega)$ and $\tilde{\chi}(\omega)$ denotes the asymptotic normalized phase of the Blaschke product evaluated over the (fundamental) unit circle, i.e., $\tilde{\chi}(\omega) = \lim_{p\to\infty} \chi_p(\omega)/p$ with $\chi_p(\omega)$, the phase of $\varphi_p(\mathrm{e}^{\mathrm{j}\omega})$ for $\omega \in [0, 2\pi)$ where $\varphi_p(z)$ is the Blaschke product (1.7). In the classical Fourier case where all $\xi_i = 0$, $\tilde{\chi}(\omega)$ is just ω because $\varphi_p(z) = z^p$, and hence $\chi_p(\omega) = p\omega$. Thus if T is the identity, this property says that the average of the eigenvalues of $M_p(f)$ converges to the average of $f \in \mathcal{C}_{2\pi}$.

Obviously all these results are of crucial importance when studying these specific rational generalized Fourier series. Thus they are important for rational approximation in general, but our main interest is motivated by the paper [14] and the arguments given there, particularly in the context of system identification and models derived

from orthonormal rational bases as they were introduced by several authors with different degrees of generality (see, e.g., [14] for some history). The equation (1.10) is a far reaching generalization of the simple Cesàro sum that we started out with in the classical Fourier case. But as formulated in section 2, we also get bounds for the generalized Fourier coefficients and therefrom obtain the convergence of the generalized Fourier series, including an estimate for the speed of convergence. The spectral properties of the matrix $M_p(f)$ are important from a numerical point of view because the condition number of this matrix will dictate the numerical robustness of the least-squares estimation problem in the underlying model identification.

The structure of the paper is as follows. In section 2, we give a detailed formulation of the results we have obtained. Their proofs are presented in the subsequent sections. In section 3, we consider the convergence of the approximation of 2π-periodic functions by their expansion over the rational basis functions $\mathcal{B}_n(z)$. In section 4, we extract from reproducing kernel properties orthonormal bases of $\mathbb{C}^p$ made of sets of the normalized vector $\tilde{\Gamma}_p(\omega)$ for appropriate ω. This is performed by analyzing the phase of the Blaschke product $\varphi_p(z)$ evaluated over the unit circle. In section 5, we introduce the concept of quadrature formulas on the unit circle and apply it to approximate particular integrals of the basis functions $\mathcal{B}_n(z)$ by use of the derived orthonormal basis of $\mathbb{C}^p$. In section 6, we apply these tailored quadrature formulas to approximate general Toeplitz matrices. Several upper bounds for the norm of the approximation error are also given. By using these bounds, we derive the results about bilinear forms based on functionals of general Toeplitz matrices in section 7. Finally, in section 8, we consider the eigenvalue distribution of general Toeplitz matrices.

2. Main results. In this section, we summarize the results we have achieved concerning the algebraic and spectral properties of general Toeplitz matrices as introduced in (1.5). We give more detailed formulations here. The derivation of these results will be obtained in the subsequent sections.

Before proceeding, let us state a fundamental assumption concerning the poles of the rational basis functions.

Assumption 1. The points ξ_i defining the basis functions $\mathcal{B}_p(z)$ lie in a closed disk of radius $c < 1$, i.e., $|\xi_i| \leq c$ for $i = 0, 1, \ldots$. Further, $\xi_0 = 0$. □

In fact, this is a key assumption for deriving all the results presented in this paper.

2.1. Rational function approximation. Here, we aim at characterizing the functional approximation properties of the rational basis functions. Therefore, we consider the orthonormal set

$$\mathcal{S}_p = \{\mathcal{B}_k : |k| < p\} \quad \text{with} \quad \mathcal{B}_0 \equiv 1 \quad \text{and} \quad \mathcal{B}_{-n}(z) = \overline{\mathcal{B}_n(1/\overline{z})}.$$

Let $a_n = \langle f, \mathcal{B}_n \rangle$ stand for the generalized Fourier coefficient of the function with respect to the rational basis function $\mathcal{B}_n(\mathrm{e}^{\mathrm{j}\omega})$. The convergence analysis of the expansion coefficients $a_{\pm n}$ has led us to the following theorem.

THEOREM 1. *Let Assumption* 1 *be satisfied. Let $f(\omega)$ be a 2π-periodic function having a continuous qth derivative with $q > 2$. Then the generalized Fourier coefficients satisfy*

$$|a_0| \leq \|f\|_1 \quad \textit{and} \quad |a_{\pm n}| \leq \left(K_1 \frac{1}{\ln \tilde{\rho}} + K_2(c) \frac{\epsilon\, n}{q-1} \right) \frac{\|f^{(q)}\|_1}{(\epsilon\, n)^q}, \qquad n \geq 1,$$

where $K_2(0^+) = 0$. *Furthermore,* $\epsilon \triangleq \epsilon(\rho, c) = \ln((\rho + c)/(1 + \rho\, c))/\ln \rho$ *for*

$$\rho \triangleq \rho(c, q, n) = \min\left(\left[\left(\frac{q}{n-1}\,\frac{-4}{\ln c}\right)^{q/4} + \left(\frac{n}{2}\right)^{q/(n-2)}\right] + \left[\frac{1+c}{\sqrt{cq}} + \frac{1}{c(q-1)}\right], e^{q/2}\right)$$

and $\tilde{\rho} \triangleq \tilde{\rho}(c, q) = \lim_{n\to\infty} \rho(c, q, n)$.

This result can be seen as the generalization of the Fourier case for which c is identically zero. It will not be a surprise that the proof is based on the Fourier expansion of the function $f(\omega)$. The proof is given in section 3.

Now, let us denote by $f_p(\omega)$ the partial sum in the expansion of a 2π-periodic function $f \in \mathcal{L}_2(\mathbb{T})$ (see expression (1.8)). Then, as a consequence of Theorem 1, the approximation error

$$e_p(\omega) = f(\omega) - f_p(\omega)$$

of the truncated expansion of the function $f(\omega)$ over the orthonormal set $\mathcal{S}_p$ has the following convergence property.

COROLLARY 2. *With the same assumptions as in Theorem* 1, *the approximation error* $e_p(\omega)$ *converges to zero uniformly in* ω *with increasing values of* p, *i.e.,*

$$\lim_{p\to\infty} |e_p(\omega)| = 0 \qquad \textit{for all } \omega$$

with a convergence rate at least as fast as $1/p^{q-2}$.

Similar to the Fourier case, this result is based on the fact the expansion (1.3) is absolutely convergent and that the limiting set $\mathcal{S} = \bigcup_{p=1}^{\infty} \mathcal{S}_p$ constitutes an orthonormal basis for $\mathcal{L}_2(\mathbb{T})$ (see [3] where no result about the convergence of $e_p(\omega)$ to zero is found).

2.2. General Toeplitz matrix properties. Now, let us concentrate on our contributions to the analysis of the general Toeplitz matrix properties. Let $\tilde{\Gamma}_p(\omega) \in \mathbb{C}^p$ be the normalized vector corresponding to $\Gamma_p(\omega)$ as defined in (1.4) and (1.6). Note that $\gamma_p(\omega) = K_p(\omega, \omega)$, where

$$K_p(\mu, \sigma) = \sum_{n=0}^{p-1} \mathcal{B}_n(\mathrm{e}^{\mathrm{j}\mu})\overline{\mathcal{B}_n(\mathrm{e}^{\mathrm{j}\sigma})}$$

is the *reproducing kernel* [2] of the p-dimensional subspace spanned by the basis functions in $\Gamma_p(\omega)$. By using the Blaschke product $\varphi_p(e^{j\omega})$ as defined in (1.7), it can be brought into a Christoffel–Darboux form (see, e.g., [14, 5]):

$$K_p(\mu, \sigma) = \frac{1 - \varphi_p(\mathrm{e}^{\mathrm{j}\mu})\overline{\varphi_p(\mathrm{e}^{\mathrm{j}\sigma})}}{\mathrm{e}^{\mathrm{j}(\sigma-\mu)} - 1}. \tag{2.1}$$

In section 4, we shall analyze some properties of $K_p(\mu, \sigma)$. The main result is as follows.

LEMMA 3. *Let* $\chi_p(\omega)$ *be the phase of the Blaschke product* $\varphi_p(e^{j\omega})$ *with its zeros satisfying Assumption* 1. *Define for arbitrary real* θ

$$\Omega_p(\theta) = \{\omega \in [0, 2\pi) \,:\, \chi_p(\omega) = \theta \bmod 2\pi\}.$$

Then, the set $\{\tilde{\Gamma}_p(\omega_i) : \omega_i \in \Omega_p(\theta)\}$ *constitutes an orthonormal basis for* $\mathbb{C}^p$. ∎

Note that there is an infinite number of such orthonormal bases because the value of the reference phase $\theta \in [0, 2\pi)$ has been left free.

Before formulating our main results on general Toeplitz matrices, let us state the following assumption on the regularity of the function $f(\omega)$.

Assumption 2. Let $f(\omega)$ be a 2π-periodic function having a continuous qth derivative with $q > 4$. □

THEOREM 4. *Let Assumption* 1 *be satisfied. Let the function* $f^{[k]}(\omega)$ *(with* $k = 1, \ldots, n$*) satisfy Assumption* 2. *Let* $T(\cdot)$ *be an* n*-variable analytic function (i.e., having convergent Taylor expansion) on the range of the* $f^{[k]}(\omega)$*'s. Then,*

$$\lim_{p\to\infty} \tilde{\Gamma}_p^*(\sigma) T\Big(M_p(f^{[1]}), \ldots, M_p(f^{[n]})\Big) \tilde{\Gamma}_p(\mu) = \begin{cases} \left[T(f^{[1]}, \ldots, f^{[n]})(\mu)\right], & \text{if } \sigma = \mu, \\ 0, & \text{otherwise.} \end{cases}$$

The convergence rate to the limiting value is at least as fast as $1/p$ *if* $\sigma = \mu$ *and as* $\ln p/p$ *otherwise.*

Note that the order in which the matrices appear in the serial expansion does not matter for the asymptotic expression since $M_p(f^{[i]})M_p(f^{[j]})$ converges to $M_p(f^{[i]}f^{[j]})$ in the Hilbert Schmidt norm as $p \to \infty$.

Let $\lambda_i(X)$ denote the ith (in any order) eigenvalue of the Hermitian matrix X.

THEOREM 5. *Let Assumption* 1 *be satisfied. Let the real function* $f(\omega)$ *satisfy Assumption* 2. *Let* $T(\cdot)$ *stand for a continuous function on the range of* $f(\omega)$. *Furthermore, let* $\chi_p(\omega)$ *be the phase of* $\varphi_p(\mathrm{e}^{\mathrm{j}\omega})$ *for* $\omega \in [0, 2\pi)$ *where* $\varphi_p(z)$ *is the Blaschke product* (1.7). *Then,* $\tilde{\chi}(\omega) = \lim_{p\to\infty} \chi_p(\omega)/p$ *exists and*

$$\lim_{p\to\infty} \frac{1}{p} \sum_{i=0}^{p-1} T\Big(\lambda_i\big(M_p(f)\big)\Big) = \frac{1}{2\pi} \int_0^{2\pi} T\Big(f\big(\tilde{\chi}^{-1}(\omega)\big)\Big) \mathrm{d}\omega.$$

The convergence rate to the limiting value is at least as fast as $1/p$.

These results are seen as straightforward generalizations of the properties of the Toeplitz matrices based on the Fourier basis functions.

3. Expansion with respect to rational bases. In this section, we analyze the functional approximation properties of the orthonormal basis set $\mathcal{S}_p = \{\mathcal{B}_k : |k| < p\}$. Note from the introduction that, under Assumption 1, the set

$$\mathcal{S} = \{\mathcal{B}_k : k \in \mathbb{Z}\} = \{\ldots, \mathcal{B}_{-p}, \ldots, \mathcal{B}_{-1}, \mathcal{B}_0, \mathcal{B}_1, \ldots, \mathcal{B}_p, \ldots\}, \tag{3.1}$$

where $\mathcal{B}_0 = 1$ and $\mathcal{B}_{-n}(z) = \overline{\mathcal{B}_n(1/\overline{z})}$, is an orthonormal basis for $\mathcal{L}_2(\mathbb{T})$.

More precisely, we show that, under suitable assumptions on the function $f(\omega)$, the approximation error $e_p(\omega) = f(\omega) - f_p(\omega)$ with $f_p(\omega)$ as in (1.8) converges to zero uniformly as p becomes unbounded.

First, we focus our attention on the expansion coefficients

$$a_n = \langle f, \mathcal{B}_n \rangle, \qquad n \in \mathbb{Z},$$

and the asymptotic behavior of these coefficients as presented in Theorem 1. Then, we consider the convergence of the approximation error $e_p(\omega)$ to zero for increasing values of p as in Corollary 2.

Therefore, we should recall some results about Fourier series of 2π-periodic functions (see Edwards [8, 9]). Namely (from Assertions 2.3.2, 2.3.5, and 2.4.3 in

[8, Chap. 2]), for any function $f(\omega) \in \mathcal{C}^q$, i.e., with continuous qth derivative (with $q \geq 2$), we have that

$$f(\omega) = \sum_{k\in\mathbb{Z}} w_k \mathrm{e}^{\mathrm{j}k\omega} \qquad \text{with} \quad w_k = \frac{1}{2\pi}\int_0^{2\pi} f(\omega)\mathrm{e}^{-\mathrm{j}k\omega}\mathrm{d}\omega.$$

The convergence of the Fourier series is uniform in ω and the Fourier coefficients are bounded as

$$|w_0| \leq \|f\|_1 \quad \text{and} \quad |w_{\pm k}| \leq \|f^{(q)}\|_1/k^q, \qquad k > 0, \tag{3.2}$$

where $f^{(q)}(\omega)$ stands for the qth derivative of the function $f(\omega)$.

Prior to prove Theorem 1, let us state the following working result.

LEMMA 6. *Let $x_2 > x_1 > 0$ and $q \geq 2$; then*

$$q\int_{x_1}^{x_2} e^{q(x-\ln x)}\,dx \leq I_{x_1<1}\left[\alpha_1^{-1}\left(1-(x_1/\tilde{x})^{(q-1)\alpha_1}\right)\right]\frac{q}{q-1}\,e^{qx_1-(q-1)\ln x_1} + I_{x_2>1}\left[\alpha_2^{-1}\left(1-e^{-q\alpha_2(x_2-\tilde{x})}\right)\right]\,e^{q(x_2-\ln x_2)},$$

where $\tilde{x} = \max(x_1, \min(1, x_2))$ while

$$\alpha_1 = 1 - \frac{q}{q-1}\tilde{x}\,\frac{1-x_1/\tilde{x}}{\ln(\tilde{x}/x_1)} \quad \textit{and} \quad \alpha_2 = 1 - x_2^{-1}\,\frac{\ln(x_2/\tilde{x})}{1-\tilde{x}/x_2}.$$

The indicator function $I_{x<y}$ is one if $x < y$ and zero otherwise.

Proof. Let us separate the integration interval in two parts, i.e.,

$$q\int_{x_1}^{x_2} e^{q(x-\ln x)}\,dx = q\int_{x_1}^{\tilde{x}} e^{q(x-\ln x)}\,dx + q\int_{\tilde{x}}^{x_2} e^{q(x-\ln x)}\,dx$$

with $\tilde{x}$ as defined above. Note that the cases where $\tilde{x} = x_2$ (resp., x_1) correspond to the situations where the second (resp., first) term in the above RHS is trivially zero because the associated integration interval reduces to one point only.

More generally, the first term is upper bounded as

$$\begin{aligned} q\int_{x_1}^{\tilde{x}} e^{q(x-\ln x)}\,dx &= q\,e^{qx_1-(q-1)\ln x_1}\int_{x_1/\tilde{x}}^{1} e^{(q-2)\ln t - qx_1(1-1/t)}dt \\ &\leq q\,e^{qx_1-(q-1)\ln x_1}\int_{x_1/\tilde{x}}^{1} e^{((q-1)\alpha_1-1)\ln t}dt \\ &= \left[\alpha_1^{-1}\left(1-(x_1/\tilde{x})^{(q-1)\alpha_1}\right)\right]\frac{q}{q-1}\,e^{qx_1-(q-1)\ln x_1} \end{aligned}$$

with $t = x_1/x$. Similarly, an upper bound for the second term is found as follows:

$$\begin{aligned} q\int_{\tilde{x}}^{x_2} e^{q(x-\ln x)}\,dx &\leq q\,e^{q(x_2-\ln x_2)}\int_{\tilde{x}}^{x_2} e^{q\alpha_2(x-x_2)}dx \\ &= \left[\alpha_2^{-1}\left(1-e^{-q\alpha_2(x_2-\tilde{x})}\right)\right]\,e^{q(x_2-\ln x_2)}. \end{aligned}$$

The proof is complete after summing up the above two upper bounds. ∎

Proof of Theorem 1. The statement for a_0 is trivial, so we only consider a_n with $|n| > 0$. Without loss of generality, let us focus on the expansion coefficient a_n with $n > 0$. Indeed, a_{-n} is simply the conjugate of the expansion coefficient of $\overline{f(\omega)}$ with respect to $\mathcal{B}_n(\mathrm{e}^{\mathrm{j}\omega})$. Hence, finding a upper bound for $|a_{-n}|$ is similar to finding an upper bound for $|a_n|$.

First, we consider three parts in the Fourier series of $f(\omega)$, i.e.,

$$f(\omega) = \sum_{k\le k_1} w_k \mathrm{e}^{\mathrm{j}k\omega} + \sum_{k=k_1+1}^{k_3-1} w_k \mathrm{e}^{\mathrm{j}k\omega} + \sum_{k\ge k_3} w_k \mathrm{e}^{\mathrm{j}k\omega}$$

for integers $k_3 > k_1 > 0$ to be specified below. So, using $\mathcal{B}_{-n}(\mathrm{e}^{\mathrm{j}\omega}) = \overline{\mathcal{B}_n(\mathrm{e}^{\mathrm{j}\omega})}$ we successively obtain

$$\begin{aligned} a_n &= \frac{1}{2\pi}\int_0^{2\pi} f(\omega)\overline{\mathcal{B}_n(\mathrm{e}^{\mathrm{j}\omega})}\mathrm{d}\omega \\ &= \frac{1}{2\pi\mathrm{j}}\oint_{|z|=1}\left[\sum_{k\le k_1} w_k z^k + \sum_{k=k_1+1}^{k_3-1} w_k z^k + \sum_{k\ge k_3} w_k z^k\right]\mathcal{B}_{-n}(z)\frac{\mathrm{d}z}{z} \\ &= \frac{1}{2\pi\mathrm{j}}\left[\sum_{k=2}^{k_1} w_k \oint_{|z|=\rho_1} z^k\mathcal{B}_{-n}(z)\frac{\mathrm{d}z}{z} + \sum_{k=k_1+1}^{k_3-1} w_k \oint_{|z|=1} z^k\mathcal{B}_{-n}(z)\frac{\mathrm{d}z}{z}\right. \\ &\qquad\left. + \sum_{k\ge k_3} w_k \oint_{|z|=\rho_3} z^k\mathcal{B}_{-n}(z)\frac{\mathrm{d}z}{z}\right] \end{aligned}$$

with appropriate $\rho_1 \ge 1$ and $c < \rho_3 \le 1$. Note that the summation terms for $k < 2$ have been cancelled out by the residue theorem using the fact that $\xi_0 = 0$.

Then, we upper bound each sum in the brackets separately:
(*First sum*) The modulus of the first sum can thus be bounded as follows:

$$\begin{aligned} \left|\frac{1}{2\pi\mathrm{j}}\sum_{k=2}^{k_1} w_k \oint_{|z|=\rho_1} z^k\mathcal{B}_{-n}(z)\frac{\mathrm{d}z}{z}\right| &\le \sum_{k=2}^{k_1} |w_k|\rho_1^k \left(\frac{1+\rho_1 c}{\rho_1 + c}\right)^{n-1} \frac{1}{2\pi}\int_0^{2\pi} \frac{\beta_n}{|\rho_1 \mathrm{e}^{\mathrm{j}\omega} - \xi_n|}\mathrm{d}\omega \\ &\le \|f^{(q)}\|_1 \left(\frac{1+\rho_1 c}{\rho_1 + c}\right)^{n-1} \rho_1^{-1} \sum_{k=2}^{k_1} \frac{\rho_1^k}{k^q}, \end{aligned} \tag{3.3}$$

where we made use of the fact that if $z = \rho\,\mathrm{e}^{\mathrm{j}\omega}$ with $|\xi| < c < 1 \le \rho$, then

$$\left|\frac{1-\overline{\xi}z}{z-\xi}\right| \le \frac{1+\rho|\xi|}{\rho+|\xi|} \le \frac{1+\rho c}{\rho + c} \le 1,$$

and that, by the Cauchy–Schwarz inequality,

$$\begin{aligned} \left[\frac{1}{2\pi}\int_0^{2\pi} \frac{\beta_n}{|\rho\mathrm{e}^{\mathrm{j}\omega} - \xi_n|}\mathrm{d}\omega\right]^2 &\le \frac{1}{2\pi}\int_0^{2\pi} \frac{\beta_n^2}{|\rho\mathrm{e}^{\mathrm{j}\omega} - \xi_n|^2}\mathrm{d}\omega \\ &= \frac{\beta_n^2}{\rho^2 - |\xi_n|^2} \le \rho^{-2} \quad \text{for } \rho \ge 1 \end{aligned}$$

with $\beta_n = \sqrt{1-|\xi_n|^2}$. We have also used the upper bound for the Fourier coefficients, i.e., $|w_k| \le \|f^{(q)}\|_1/k^q$ for $k > 0$.

The summation factor in expression (3.3) is first simplified into

$$\sum_{k=2}^{k_1} \frac{\rho_1^k}{k^q} \approx q\,(\ln \rho_1)^{q-1}\, e^{-q \ln q} \int_{2 \ln \rho_1/q}^{k_1 \ln \rho_1/q} e^{q(x-\ln x)}\, dx$$

with $x = k \ln \rho_1/q$. By use of Lemma 6 (see notation therein), we then obtain

$$\begin{aligned} & q\,(\ln \rho_1)^{q-1}\, e^{-q \ln q} \int_{x_1}^{x_2} e^{q(x-\ln x)}\, dx \\ & \qquad \le I_{x_1<1} \left[\alpha_1^{-1} \left(1 - \left(\frac{x_1}{\tilde{x}}\right)^{(q-1)\alpha_1}\right)\right] \frac{2}{q-1}\, \frac{\rho_1^2}{2^q} \\ & \qquad\qquad + I_{x_2>1} \left[\alpha_2^{-1} \left(1 - e^{-q\alpha_2(x_2-\tilde{x})}\right)\right] \frac{1}{\ln \rho_1}\, \frac{\rho_1^{k_1}}{{k_1}^q}, \end{aligned} \tag{3.4}$$

where $x_1 = 2 \ln \rho_1/q$ and $x_2 = k_1 \ln \rho_1/q$ while $\tilde{x} = \max(x_1, \min(1, x_2))$. Note that increasing values of ρ_1 make x_1 (resp., x_2) closer to (resp., farther away from) one from below (resp., above) so that the second term in the RHS of expression (3.4) becomes dominant.

The last step of the derivation is to choose the values of ρ_1 and k_1. The choice of k_1 is such that it makes the quantity

$$\rho_1^{k_1-1} \left(\frac{1+\rho_1 c}{\rho_1 + c}\right)^{n-1}$$

less than one, independently of n. This is done by imposing $k_1 = \lfloor \epsilon_1 (n-1) + 1 \rfloor$ with

$$0 < \epsilon_1 = [\ln(\rho_1 + c) - \ln(1 + \rho_1 c)]/\ln \rho_1 \le 1 \quad \text{(because } c < 1 \le \rho_1\text{)}$$

and $\lfloor x \rfloor$ denoting the largest integer not greater than x. Note that if $\rho_1 \gg 1/c$, then ϵ_1 becomes close to zero.

The choice for ρ_1 is

$$\rho_1 = \min\left([(\rho_{11} - 1) + \rho_{12}] + (\rho_{13} - 1), e^{q/2}\right), \tag{3.5}$$

where

$$\rho_{11} = 1 + \left(\frac{q}{n-1}\, \frac{-4}{\ln c}\right)^{q/4}, \quad \rho_{12} = \left(\frac{n}{2}\right)^{q/(n-2)}, \quad \text{and} \quad \rho_{13} = 1 + \frac{1+c}{\sqrt{cq}} + \frac{1}{c(q-1)}.$$

The reason for this choice is the following. For $n \gg (-4/\ln c)\, q$, the upper bound (3.3) receives its main contribution from the second term in the RHS expression (3.4) for which a minimum is obtained for $\rho_1 = \rho_{13}$ (given $k_1(\epsilon_1, n)$ as above). In the other cases, i.e., $n \not\gg (-4/\ln c)\, q$, a small value of the upper bound (3.3) is obtained by compromising the contributions of the two terms in the RHS expression (3.4) while adjusting ρ_1. Depending on values of c, q, and n, the resulting ρ_1 may become very large or stay close to one: namely, ρ_{11} (resp., ρ_{12}) describes the behavior of ρ_1 when it is far from (resp., close to) one. It is further seen that $\rho_{11} \gg \rho_{12} \gg \rho_{13}$ for $n \ll (-4/\ln c)\, q$. The minimum operator in expression (3.5) limits the value of ρ_1 to a maximum of $e^{q/2}$ for which $x_1 = 1$ in expression (3.4) so that only its second term remains. Finally, the expression for ρ_1 sums up the different $\rho_{1\cdot}$'s whose contributions apply in particular n intervals.

Provided these choices for k_1 and ρ_1, the expression (3.3) leads to

$$\left|\frac{1}{2\pi\mathrm{j}}\sum_{k=2}^{k_1} w_k \oint_{|z|=\rho_1} z^k \mathcal{B}_{-n}(z)\frac{\mathrm{d}z}{z}\right| \leq \left[\frac{C_1}{\ln\tilde{\rho}_1}\right] \|f^{(q)}\|_1/(\epsilon_1 n)^q, \tag{3.6}$$

where $\tilde{\rho}_1(c,q) = \lim_{n\to\infty}\rho_1(c,q,n)$. Note that ϵ_1 explicitly appears in the upper bounding expression (3.6). The reason for that is that it is smaller than one and depends on the triplet (c,q,n) via ρ_1.

(*Third sum*) Similarly, the modulus of the third sum in the expression of a_n can be upper bounded as follows:

$$\begin{aligned}\left|\frac{1}{2\pi\mathrm{j}}\sum_{k\geq k_3} w_k \oint_{|z|=\rho_3} z^k \mathcal{B}_{-n}(z)\frac{\mathrm{d}z}{z}\right| &\leq \sum_{k\geq k_3}|w_k|\rho_3^k\left(\frac{1-\rho_3 c}{\rho_3-c}\right)^n \frac{1}{2\pi}\int_0^{2\pi}\frac{\beta_n}{|\rho_3\mathrm{e}^{\mathrm{j}\omega}-\xi_n|}\mathrm{d}\omega \\ &\leq \|f^{(q)}\|_1\left(\frac{1-\rho_3 c}{\rho_3-c}\right)^{n+1/2}\sum_{k\geq k_3}\frac{\rho_3^{k-1/2}}{k^q} \\ &\leq \|f^{(q)}\|_1\left(\frac{1-\rho_3 c}{\rho_3-c}\right)^{n+1/2}\frac{1}{1-\rho_3}\frac{\rho_3^{k_3-1/2}}{k_3^q}, \qquad (3.7)\end{aligned}$$

where we have used the fact that for $z=\rho\mathrm{e}^{\mathrm{j}\omega}$ and $|\xi|<c\leq\rho<1$ we have

$$\left|\frac{1-\bar{\xi}z}{z-\xi}\right| \leq \frac{1-\rho|\xi|}{\rho-|\xi|} \leq \frac{1-\rho c}{\rho-c}$$

and also (by the Cauchy–Schwarz inequality as above)

$$\left[\frac{1}{2\pi}\int_0^{2\pi}\frac{\beta_n}{|\rho\mathrm{e}^{\mathrm{j}\omega}-\xi_n|}\mathrm{d}\omega\right]^2 \leq \frac{1-c^2}{\rho^2-c^2} \leq \rho^{-1}\left(\frac{1-\rho c}{\rho-c}\right) \quad \text{for } c<\rho\leq 1.$$

The choice for k_3 is such that it makes

$$\rho_3^{k_3-1/2}\left(\frac{1-\rho_3 c}{\rho_3-c}\right)^{n+1/2} \leq 1$$

independently of n. This is done by imposing $k_3 = \lceil \epsilon_3(n+1/2)+1/2\rceil$ with

$$\epsilon_3 = [\ln(\rho_3-c)-\ln(1-\rho_3 c)]/\ln\rho_3 \geq 1 \quad (\text{because } c<\rho_3\leq 1)$$

and $\lceil x\rceil$ denoting the smallest integer not smaller than x.

Then, by choosing $\rho_3 = 1-(1-c)^q$ (as it has been left free so far) with $q>2$, the expression (3.7) becomes

$$\begin{aligned}\left|\frac{1}{2\pi\mathrm{j}}\sum_{k\geq k_3} w_k \oint_{|z|=\rho_3} z^k \mathcal{B}_{-n}(z)\frac{\mathrm{d}z}{z}\right| &\leq \|f^{(q)}\|_1\frac{(1-(1-c)^q)^{\tau_3}}{(1-c)^q}\left[\epsilon_3\left(n+\frac{1}{2}\right)+\frac{1}{2}\right]^{-q} \\ &\leq \left[\frac{2}{1+\epsilon_3}\frac{1}{1-c}\right]^q \frac{\|f^{(q)}\|_1}{n^q} \leq \frac{\|f^{(q)}\|_1}{n^q}, \qquad (3.8)\end{aligned}$$

where $0\leq\tau_3 = \lceil\epsilon_3(n+1/2)+1/2\rceil - [\epsilon_3(n+1/2)+1/2] < 1$. The last expression comes from the fact that $\epsilon_3 \geq (1+c)/(1-c)$ for the chosen ρ_3 value. It further tends

to this lower limit for increasing q. Note that, as before, the rate of convergence to zero with n is $1/n^q$.

(*Second sum*) The second sum in expression of a_n is treated as follows:

$$\left|\frac{1}{2\pi \mathrm{j}}\sum_{k=k_1+1}^{k_3-1} w_k \oint_{|z|=1} z^k \mathcal{B}_{-n}(z)\frac{\mathrm{d}z}{z}\right| \le \|f^{(q)}\|_1 \sum_{k=k_1+1}^{k_3-1}\frac{1}{k^q}$$

$$\le \|f^{(q)}\|_1 \frac{C_2}{q-1}\left[\frac{1-\epsilon_1^{q-1}}{\epsilon_1^{q-1}} + \frac{\epsilon_3^{q-1}-1}{\epsilon_3^{q-1}}\right]\frac{1}{n^{q-1}}, \tag{3.9}$$

where we divided the summation interval $[k_1+1, k_3-1]$ into two subintervals, i.e., $[k_1+1, n]$ and $[n+1, k_3-1]$.

(*Overall*) When putting the expression (3.6), (3.8), and (3.9) together, we finally obtain the following upper bound:

$$|a_n| \le \left[\left(\frac{C_1}{\ln\tilde{\rho}_1}+\epsilon_1^q\right) + C_2\,\frac{\epsilon_1 n}{q-1}\left((1-\epsilon_1^{q-1}) + \left(\frac{\epsilon_1}{\epsilon_3}\right)^{q-1}(\epsilon_3^{q-1}-1)\right)\right]\frac{\|f^{(q)}\|_1}{(\epsilon_1 n)^q}$$

$$\le \left[K_1\,\frac{1}{\ln\tilde{\rho}_1} + K_2(c)\,\frac{\epsilon_1 n}{q-1}\right]\frac{\|f^{(q)}\|_1}{(\epsilon_1 n)^q}, \tag{3.10}$$

where K_1 and $K_2(c)$ are defined so that $K_2(0^+) = 0$.

The proof is completed when identifying ϵ, ρ, and $\tilde{\rho}$ to ϵ_1, ρ_1, and $\tilde{\rho}_1$, respectively. □

Proof of Corollary 2. As the function $f(\omega)$ is continuous, the proof requires only the uniform convergence of $\lim_{p\to\infty} f_p(\omega)$. This immediately implies that the limiting function is f, and thus that the error e_p converges uniformly to zero. Indeed, by completeness of the basis (3.1) on which this expansion is constructed, we can derive that if $f(\omega)$ is continuous and all the expansion coefficients are zero, then the function must be identically zero (similar to Assertion 2.4.1 in [8]). Then, in case the expansion $f_p(\omega)$ converges uniformly in ω, this limiting function is continuous and has expansion coefficients identical to those of the expansion itself. Thus, this limiting function and the original function $f(\omega)$ coincide identically (see [8, Assertion 2.4.3]).

Hence, it remains to show that the convergence $\lim_{p\to\infty} f_p(\omega) = f(\omega)$ is uniform in ω. For p large enough, we have that

$$|e_p(\omega)| = \left|\lim_{r\to\infty} f_r(\omega) - f_p(\omega)\right| = \left|\sum_{n\ge p}\left(a_n\mathcal{B}_n(\mathrm{e}^{\mathrm{j}\omega}) + a_{-n}\overline{\mathcal{B}_n(\mathrm{e}^{\mathrm{j}\omega})}\right)\right|$$

$$\le 2\sum_{n\ge p}\max(|a_n|,|a_{-n}|)|\mathcal{B}_n(\mathrm{e}^{\mathrm{j}\omega})|$$

$$\le 2\|f^{(q)}\|_1 \sum_{n\ge p}\left(K_1\,\frac{1}{\ln\tilde{\rho}} + K_2\,\frac{\epsilon\, n}{q-1}\right)/(\epsilon\, n)^q,$$

where K_1, K_2, and $\tilde{\rho}$ do not depend on n or ω (see the notation in Theorem 1). As $q > 2$, the sum in the RHS converges to zero for $p \to \infty$. Furthermore, the convergence rate is at least as fast as $1/p^{q-2}$. This completes the proof. □

4. Reproducing kernel and orthonormal bases of $\mathbb{C}^p$. In section 2.2, we have introduced vectors $\tilde{\Gamma}_p(\omega) \in \mathbb{C}^p$. Their unit norm easily follows from the definition

of $\gamma_p(\omega)$, i.e.,

$$\tilde{\Gamma}_p^*(\omega)\tilde{\Gamma}_p(\omega) = \frac{1}{\gamma_p(\omega)} \sum_{n=0}^{p-1} \left|\mathcal{B}_n(\mathrm{e}^{\mathrm{j}\omega})\right|^2 = 1.$$

Now, in order to prove Lemma 3, let us show how to construct orthonormal bases of $\mathbb{C}^p$ consisting of p such vectors. Therefore we remember that $K_p(\omega,\sigma) = \Gamma_p^*(\sigma)\Gamma_p(\omega)$ so that by (2.1)

$$\Gamma_p^*(\sigma)\Gamma_p(\omega) = 0 \qquad \Longleftrightarrow \qquad \overline{\varphi_p(\mathrm{e}^{\mathrm{j}\sigma})}\varphi_p(\mathrm{e}^{\mathrm{j}\omega}) = 1$$

for appropriate $\sigma \neq \omega$. By the Blaschke product expression, this implies the following condition:

$$\Phi\left(\varphi_p(\mathrm{e}^{\mathrm{j}\sigma})\right) = \Phi\left(\varphi_p(\mathrm{e}^{\mathrm{j}\omega})\right),$$

where $\Phi(\xi)$ denotes the phase of $\xi \in \mathbb{C}$. The phase of $\varphi_p(\mathrm{e}^{\mathrm{j}\omega})$, that we denote from now on by $\chi_p(\omega)$, is written as

$$\chi_p(\omega) = p\omega - 2\sum_{n=0}^{p-1} \Phi\left(1 - \overline{\xi}_n \mathrm{e}^{\mathrm{j}\omega}\right).$$

It satisfies $\chi_p(\omega + 2\pi) = p2\pi + \chi_p(\omega)$. Furthermore, the following result holds for its derivative.

LEMMA 7. *Let the zeros of the Blaschke product $\varphi_p(\mathrm{e}^{\mathrm{j}\omega})$ lie inside the unit disk, i.e., $|\xi_n| \leq c < 1$ with $n = 0, \ldots, p-1$. Then,*

$$0 < \frac{p}{K_c} \leq \frac{\mathrm{d}\chi_p(\omega)}{\mathrm{d}\omega} \leq pK_c \tag{4.1}$$

with $1 \leq K_c := (1+c)/(1-c) < \infty$.

Proof. From the expression of $\chi_p(\omega)$, we successively have

$$\begin{aligned}\frac{\mathrm{d}\chi_p(\omega)}{\mathrm{d}\omega} &= p - 2\sum_{n=0}^{p-1} \frac{\mathrm{d}\Phi\left(1 - \overline{\xi}_n \mathrm{e}^{\mathrm{j}\omega}\right)}{\mathrm{d}\omega} \\ &= p - 2\sum_{n=0}^{p-1} |\xi_n| \frac{|\xi_n| - \cos(\omega - \Phi(\xi_n))}{1 + |\xi_n|^2 - 2|\xi_n|\cos(\omega - \Phi(\xi_n))},\end{aligned}$$

where $\Phi(\xi_n)$ stands for the phase of ξ_n. Hence

$$p - 2\sum_{n=0}^{p-1} \frac{|\xi_n|}{1+|\xi_n|} \leq \frac{\mathrm{d}\chi_p(\omega)}{\mathrm{d}\omega} \leq p + 2\sum_{n=0}^{p-1} \frac{|\xi_n|}{1-|\xi_n|}.$$

Then, the proof is completed by noting that the worst lower and upper bounding case occurs when $|\xi_n| = c$ for all n. ∎

The consequence of this is twofold. First, for a given $\theta \in [0, 2\pi)$, there are exactly p distinct frequencies $\omega_0, \ldots, \omega_{p-1}$ in $[0, 2\pi)$ for which the phase $\chi_p(\omega)$ of the Blaschke product takes this value θ modulo 2π, i.e.,

$$\Omega_p(\theta) = \{\omega \in [0, 2\pi) : \chi_p(\omega) = \theta \bmod 2\pi\} = \{\omega_0, \ldots, \omega_{p-1}\}. \tag{4.2}$$

Second, the distance Δ between two successive frequencies in $\Omega_p(\theta)$ is lower bounded by $2\pi/(pK_c)$. Indeed,

$$2\pi = \chi_p(\mu+\Delta) - \chi_p(\mu) = \int_{\mu}^{\mu+\Delta} \frac{\mathrm{d}\chi_p(\omega)}{\mathrm{d}\omega}\mathrm{d}\omega \leq K_c p\Delta \tag{4.3}$$

for any $\mu, \mu+\Delta \in \Omega_p(\theta)$.

As p increases, this distance decreases because the phase of $\varphi_p(\mathrm{e}^{\mathrm{j}\omega})$ is turning (under the modulo operator) more rapidly in the interval $[0, 2\pi)$. In the limit as p tends to infinity, there is a (countably) infinite number of frequencies in $\Omega_p(\theta)$ (included within $[0, 2\pi)$) and they become infinitely close to each other without inducing any limiting point.

Hence, we have proven the result in Lemma 3.

From now on, we denote by Υ_p the matrix representing such an orthonormal basis. Its ith column is given by $(\Upsilon_p)_i = \tilde{\Gamma}_p(\omega_i)$. This is a unitary matrix so that, by the orthonormality of its rows, we derive that

$$\sum_{i=0}^{p-1} \frac{1}{\gamma_p(\omega_i)} \mathcal{B}_n(\mathrm{e}^{\mathrm{j}\omega_i})\overline{\mathcal{B}_m(\mathrm{e}^{\mathrm{j}\omega_i})} = \delta_{n-m}. \tag{4.4}$$

Finally, as the function $\chi_p(\omega)$ becomes unbounded for increasing p, it is worth defining the following normalized limiting function as mentioned in Theorem 5:

$$\tilde{\chi}(\omega) = \lim_{p\to\infty} \chi_p(\omega)/p.$$

This function $\tilde{\chi}(\omega)$ has values in the interval

$$[\tilde{\chi}(0), \tilde{\chi}(0)+2\pi) \quad \text{with} \quad \tilde{\chi}(0) = -2\lim_{p\to\infty} \sum_{n=0}^{p-1} \Phi(1-\overline{\xi}_n)/p$$

when ω lies in $[0, 2\pi)$. Note that if each complex-valued ξ_n has a conjugate counterpart, then this initial value of $\tilde{\chi}(\omega)$ vanishes.

Before ending this section, let us evaluate the decomposition of any normalized vector $\tilde{\Gamma}_p(\omega)$ in the orthonormal basis $\{\tilde{\Gamma}_p(\omega_i) : \omega_i \in \Omega_p(\theta)\}$.

LEMMA 8. *Let $\alpha_{p,i}(\omega)$ denote the ith decomposition coefficient of $\tilde{\Gamma}_p(\omega)$ in the orthonormal set $\{\tilde{\Gamma}_p(\omega_i) : \omega_i \in \Omega_p(\theta)\}$, i.e., $\alpha_{p,i}(\omega) = \tilde{\Gamma}_p^*(\omega_i)\tilde{\Gamma}_p(\omega)$. Then, $\alpha_{p,i}(\omega_k) = \delta_{k-i}$ and*

$$|\alpha_{p,i}(\omega)| \leq \min\left(1, \frac{K_c}{p}\left|\sin\frac{\omega_i-\omega}{2}\right|^{-1}\right) \qquad \textit{for } \omega \notin \Omega_p(\theta)$$

with $K_c = (1+c)/(1-c)$.

Proof. The case where $\omega = \omega_k$ belongs to $\Omega_p(\theta)$ is trivial. When $\omega \notin \Omega_p(\theta)$, we have by 2-norm properties that $|\alpha_{p,i}(\omega)| \leq \|\tilde{\Gamma}_p^*(\omega_i)\|_2 \|\tilde{\Gamma}_p(\omega)\|_2 \leq 1$. But, by definition of $\tilde{\Gamma}_p(\omega)$, we also have that

$$|\alpha_{p,i}(\omega)| = \frac{|K_p(\omega,\omega_i)|}{\gamma_p^{1/2}(\omega)\gamma_p^{1/2}(\omega_i)} \leq \left|\sin\frac{\omega_i-\omega}{2}\right|^{-1} \max_{\omega} \gamma_p^{-1}(\omega) \tag{4.5}$$

with

$$\gamma_p(\omega) = \sum_{i=0}^{p-1} \frac{1-|\xi_i|^2}{|1-\overline{\xi_i}\mathrm{e}^{\mathrm{j}\omega}|^2} \geq \frac{p}{K_c}$$

so that the proof is completed. □

Furthermore, it is worth evaluating the 1-norm of the sequence of the decomposition coefficients. We have the following result.

LEMMA 9. *With a notation similar to Lemma* 8 *and* $\alpha_p(\omega) = \{\alpha_{p,i}(\omega) : 0 \leq i < p\}$, *we have that*

$$\|\alpha_p(\omega)\|_1 \leq 2 - 2(K_c^2/\pi)\ln[\tan(\pi/2pK_c)] \qquad \textit{for} \;\; p \geq 4,$$

where $\|\alpha_p\|_1$ *stands for the* 1*-norm* $\sum_{k=0}^{p-1}|\alpha_{p,i}|$.

Proof. First, assume that the two frequency points surrounding ω are ω_k and ω_{k+1}. Without loss of generality, we may assume they are somewhere in the middle of the set $\Omega_p(\theta)$ for $p \geq 4$. So, we can successively write

$$\begin{aligned}
\sum_{i=0}^{p-1}|\alpha_{p,i}| &= \sum_{i=k-1}^{k+2}|\alpha_{p,i}| + \left[\sum_{i=0}^{k-2} + \sum_{i=k+3}^{p-1}\right]|\alpha_{p,i}| \\
&\leq 2 + \frac{K_c}{p}\left[\sum_{i'=-k}^{-2} + \sum_{i'=2}^{p-k-2}\right]\left|\sin\frac{i'\pi}{pK_c}\right|^{-1} \\
&\leq 2 + 2\frac{K_c}{p}\sum_{i'=2}^{p/2}\frac{1}{\sin(i'\pi/pK_c)} \leq 2 + 2\frac{K_c^2}{\pi}\int_{\pi/pK_c}^{\pi/2}\frac{\mathrm{d}x}{\sin x} \\
&\leq 2 - 2\left(\frac{K_c^2}{\pi}\right)\ln\left[\tan\left(\frac{\pi}{2pK_c}\right)\right],
\end{aligned}$$

where we have used the fact that $\sum_i |\alpha_{p,i}|^2 = 1$ in upper bounding the first sum in the first expression. We have also taken advantage of

$$\frac{p}{K_c}|\alpha_{p,i}| \leq \left|\sin\left(\frac{\omega_i-\omega}{2}\right)\right|^{-1} \leq \left|\sin\left(\frac{(k-i)\pi}{pK_c}\right)\right|^{-1} \quad \text{or} \quad \left|\sin\left(\frac{(i-(k+1))\pi}{pK_c}\right)\right|^{-1}$$

as the difference between two successive frequency points ω_i is larger than $2\pi/pK_c$. □

Hence, for large p, the 1-norm of the sequence $\alpha_p(\omega)$ diverges like $2(K_c^2/\pi)\ln p$.

5. Quadrature formulas. In the present section, we intend to find approximate quadrature formulas to evaluate integrals of the form[1]

$$\frac{1}{2\pi}\int_0^{2\pi} \mathcal{B}_k(\mathrm{e}^{\mathrm{j}\omega})\overline{\mathcal{B}_l(\mathrm{e}^{\mathrm{j}\omega})}\mathcal{B}_r(\mathrm{e}^{\mathrm{j}\omega})\mathrm{d}\omega, \tag{5.1}$$

where $0 \leq k, l, r < p$.

[1]This generic integral originates from the definition (1.5) of general Toeplitz matrices with the expansion (1.3) for the function $f(\omega)$.

To recall, quadrature formulas [7] are used to approximate the integral of functions as the sum of values of these functions at appropriate points of the integration interval, i.e., generically

$$\int_a^b f(x)\mathrm{d}x \approx \sum_n A_n f(x_n),$$

where x_n are the sampling points inside $[a,b]$ and A_n are appropriate constants related to the functional class to which $f(x)$ belongs. In the case where the integration corresponds to an integration over the complex unit circle, we refer to papers by Bultheel et al. [3, 4, 6].

In the preceding section, we have already derived the quadrature formula corresponding to the integral (5.1) in the particular case where r is zero. Indeed, from the orthonormality of the functions $\mathcal{B}_n(z)$ and of the rows of the unitary matrix Υ_p in (4.4), we obtain that

$$\frac{1}{2\pi}\int_0^{2\pi} \mathcal{B}_n(\mathrm{e}^{\mathrm{j}\omega})\overline{\mathcal{B}_m(\mathrm{e}^{\mathrm{j}\omega})}\mathrm{d}\omega = \delta_{n-m} = \sum_{i=0}^{p-1} \frac{1}{\gamma_p(\omega_i)}\mathcal{B}_n(\mathrm{e}^{\mathrm{j}\omega_i})\overline{\mathcal{B}_m(\mathrm{e}^{\mathrm{j}\omega_i})} \tag{5.2}$$

with ω_i in $\Omega_p(\theta)$ for some θ. The rightmost sum obviously forms a quadrature formula for the leftmost integral.

Now, for $z \in \mathbb{T}$, let us write the basis function $\mathcal{B}_n(z)$ as

$$\mathcal{B}_n(z) = \beta_n \frac{1}{\mu_p(z)}\frac{\mu_p(z)}{\mu_{n+1}(z)}\nu_n(z),$$

where $\beta_n^2 = 1 - |\xi_n|^2$ while

$$\nu_n(z) = \prod_{i=0}^{n-1}(z-\xi_i) \quad \text{and} \quad \mu_n(z) = \prod_{i=0}^{n-1}(1-\overline{\xi}_i z).$$

Similarly, we have

$$\overline{\mathcal{B}_m(z)} = \beta_m \frac{z}{\nu_p(z)}\mu_m(z)\frac{\nu_p(z)}{\nu_{m+1}(z)}.$$

Hence, we can write

$$\mathcal{B}_n(z)\overline{\mathcal{B}_m(z)} = \frac{z}{\mu_p(z)\nu_p(z)}\left[(\beta_n\beta_m)\frac{\mu_p(z)\mu_m(z)}{\mu_{n+1}(z)}\frac{\nu_n(z)\nu_p(z)}{\nu_{m+1}(z)}\right] \tag{5.3}$$

while

$$\mathcal{B}_k(z)\overline{\mathcal{B}_l(z)}\mathcal{B}_r(z) = \frac{z}{\mu_p(z)\nu_p(z)}\left[(\beta_k\beta_l\beta_r)\frac{\mu_p(z)\mu_l(z)}{\mu_{k+1}(z)\mu_{r+1}(z)}\frac{\nu_k(z)\nu_r(z)\nu_p(z)}{\nu_{l+1}(z)}\right].$$

In order to have this last expression written as a sum of terms as in (5.3), i.e.,

$$\mathcal{B}_k(z)\overline{\mathcal{B}_l(z)}\mathcal{B}_r(z) = \sum_{n,m=0}^{p-1}\kappa_{n,m}\mathcal{B}_n(z)\overline{\mathcal{B}_m(z)},$$

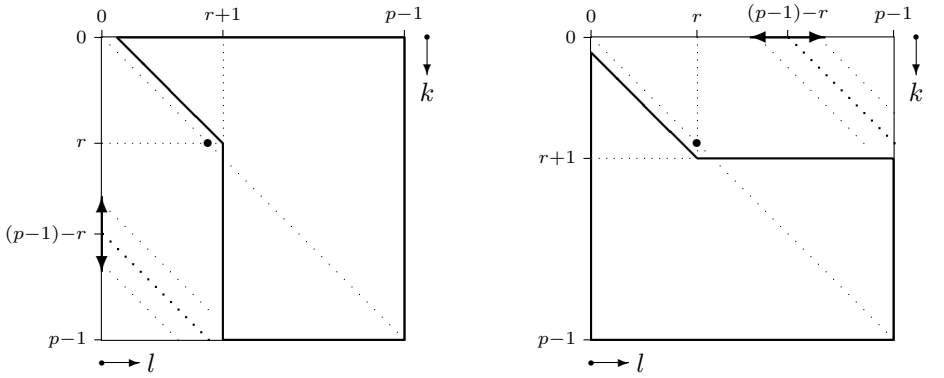


FIG. 5.1. *Admissible (k,l) areas for the quadrature formulas of $\mathcal{B}_k\overline{\mathcal{B}_l}\mathcal{B}_r$ (left) and $\mathcal{B}_k\overline{\mathcal{B}_l}\mathcal{B}_{-r}$ (right) to hold, as functions of r.*

we need that $l > \min(k,r)$ so that the expression between the brackets of its RHS reduces to a polynomial in z only. The coefficients $\kappa_{n,m}$ of this expansion are found by imposing that the zeros of the LHS (being either ξ_i or $1/\overline{\xi}_i$) are reproduced in the RHS.

Hence, the quadrature formula in (5.2) can be applied as follows:

$$\begin{aligned}\frac{1}{2\pi}\int_0^{2\pi}\mathcal{B}_k(\mathrm{e}^{\mathrm{j}\omega})\overline{\mathcal{B}_l(\mathrm{e}^{\mathrm{j}\omega})}\mathcal{B}_r(\mathrm{e}^{\mathrm{j}\omega})\mathrm{d}\omega &= \sum_{n,m=0}^{p-1}\kappa_{n,m}\frac{1}{2\pi}\int_0^{2\pi}\mathcal{B}_n(\mathrm{e}^{\mathrm{j}\omega})\overline{\mathcal{B}_m(\mathrm{e}^{\mathrm{j}\omega})}\mathrm{d}\omega\\ &= \sum_{n,m=0}^{p-1}\kappa_{n,m}\sum_{i=0}^{p-1}\frac{1}{\gamma_p(\omega_i)}\mathcal{B}_n(\mathrm{e}^{\mathrm{j}\omega_i})\overline{\mathcal{B}_m(\mathrm{e}^{\mathrm{j}\omega_i})}\\ &= \sum_{i=0}^{p-1}\frac{1}{\gamma_p(\omega_i)}\mathcal{B}_k(\mathrm{e}^{\mathrm{j}\omega_i})\overline{\mathcal{B}_l(\mathrm{e}^{\mathrm{j}\omega_i})}\mathcal{B}_r(\mathrm{e}^{\mathrm{j}\omega_i}).\end{aligned}\tag{5.4}$$

To summarize, we state the following result.

LEMMA 10. *For $0 \le k,r,l < p$, we have that*

$$\langle\mathcal{B}_k\mathcal{B}_r,\mathcal{B}_l\rangle = \sum_{i=0}^{p-1}\frac{1}{\gamma_p(\omega_i)}\mathcal{B}_k(\mathrm{e}^{\mathrm{j}\omega_i})\overline{\mathcal{B}_l(\mathrm{e}^{\mathrm{j}\omega_i})}\mathcal{B}_r(\mathrm{e}^{\mathrm{j}\omega_i})$$

provided that $l > \min(k,r)$. □

With the help of $\mathcal{B}_k(z)\overline{\mathcal{B}_l(z)}\mathcal{B}_{-r}(z) = \overline{\mathcal{B}_l(z)\overline{\mathcal{B}_k(z)}\mathcal{B}_r(z)}$ (for $z \in \mathbb{T}$) and provided that $k > \min(l,r)$, we can similarly write the following quadrature formula:

$$\langle\mathcal{B}_k\mathcal{B}_{-r},\mathcal{B}_l\rangle = \sum_{i=0}^{p-1}\frac{1}{\gamma_p(\omega_i)}\mathcal{B}_k(\mathrm{e}^{\mathrm{j}\omega_i})\,\overline{\mathcal{B}_l(\mathrm{e}^{\mathrm{j}\omega_i})}\,\overline{\mathcal{B}_r(\mathrm{e}^{\mathrm{j}\omega_i})}.\tag{5.5}$$

Before considering the situation where the index l (resp., k) does not satisfy the previously mentioned condition for $\langle\mathcal{B}_k\mathcal{B}_r,\mathcal{B}_l\rangle$ (resp., $\langle\mathcal{B}_k\mathcal{B}_{-r},\mathcal{B}_l\rangle$), let us draw a picture representing the zones in the (k,l) plane where it indeed satisfies it. It appears in Figure 5.1 that the smaller the value of r, the larger the admitted area in the plane.

Now, let us focus on the other situations, e.g., $l \leq \min(k,r)$ when evaluating the quantity in (5.1). Therefore, we first express the product $\mathcal{B}_k(z)\mathcal{B}_r(z)$ lying in $\mathcal{H}_2(\mathbb{D})$ as an expansion over the basis function $\mathcal{B}_n(z)$ with integer $n \geq 0$, i.e.,

$$\mathcal{B}_k(z)\mathcal{B}_r(z) = \sum_{n=0}^{p-1} \langle \mathcal{B}_k\mathcal{B}_r, \mathcal{B}_n\rangle\, \mathcal{B}_n(z) + \sum_{n\geq p} \langle \mathcal{B}_k\mathcal{B}_r, \mathcal{B}_n\rangle\, \mathcal{B}_n(z).$$

Then, we evaluate this expression at $z = \mathrm{e}^{\mathrm{j}\omega_i}$, we multiply both sides by $[\overline{\mathcal{B}_l(\mathrm{e}^{\mathrm{j}\omega_i})}/\gamma_p(\omega_i)]$ and we sum over $i = 0, \ldots, p-1$. Hence, by taking advantage of the quadrature formula in (5.2), we obtain that

$$\begin{aligned}\frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{B}_k(\mathrm{e}^{\mathrm{j}\omega})\overline{\mathcal{B}_l(\mathrm{e}^{\mathrm{j}\omega})}\mathcal{B}_r(\mathrm{e}^{\mathrm{j}\omega})\mathrm{d}\omega &= \sum_{i=0}^{p-1} \frac{1}{\gamma_p(\omega_i)}\mathcal{B}_k(\mathrm{e}^{\mathrm{j}\omega_i})\overline{\mathcal{B}_l(\mathrm{e}^{\mathrm{j}\omega_i})}\mathcal{B}_r(\mathrm{e}^{\mathrm{j}\omega_i}) \\ &\quad - \sum_{n\geq p} \langle \mathcal{B}_k\mathcal{B}_r, \mathcal{B}_n\rangle \sum_{i=0}^{p-1} \frac{1}{\gamma_p(\omega_i)}\mathcal{B}_n(\mathrm{e}^{\mathrm{j}\omega_i})\overline{\mathcal{B}_l(\mathrm{e}^{\mathrm{j}\omega_i})}. \qquad (5.6)\end{aligned}$$

In order to analyze the last summation term in the RHS of this expression, it is helpful to make the following (temporary) assumption.

Assumption 3. The zeros in $\mathcal{B}_n(z)$ for $n \geq p$ are obtained by cyclically repeating the p zeros of $\varphi_p(z)$ in their original order. □

Under this assumption, the basis functions $\mathcal{B}_n(z)$ for $n \geq p$ can easily be written as $\mathcal{B}_n(z) = \mathcal{B}_s(z)\varphi_p^{n'}(z)$ for $n = s + n'p$ with $0 \leq s < p$. Furthermore, we have the following result.

THEOREM 11. *For $0 \leq k, r, l < p$, Assumption* 3 *leads to*

$$\langle \mathcal{B}_k\mathcal{B}_r, \mathcal{B}_l\rangle = \sum_{i=0}^{p-1} \frac{1}{\gamma_p(\omega_i)}\mathcal{B}_k(\mathrm{e}^{\mathrm{j}\omega_i})\overline{\mathcal{B}_l(\mathrm{e}^{\mathrm{j}\omega_i})}\mathcal{B}_r(\mathrm{e}^{\mathrm{j}\omega_i}) - \mathrm{e}^{\mathrm{j}\theta} D_{(k,r),l}, \qquad (5.7)$$

where $D_{(k,r),l} = \langle \mathcal{B}_k\mathcal{B}_r, \mathcal{B}_{l+p}\rangle$ for $l \leq \min(k,r)$ and is identical to zero otherwise.

Proof. First, we have that

$$\sum_{i=0}^{p-1} \frac{1}{\gamma_p(\omega_i)}\mathcal{B}_n(\mathrm{e}^{\mathrm{j}\omega_i})\overline{\mathcal{B}_l(\mathrm{e}^{\mathrm{j}\omega_i})} = \mathrm{e}^{\mathrm{j}n'\theta}\sum_{i=0}^{p-1} \frac{1}{\gamma_p(\omega_i)}\mathcal{B}_s(\mathrm{e}^{\mathrm{j}\omega_i})\overline{\mathcal{B}_l(\mathrm{e}^{\mathrm{j}\omega_i})} = \mathrm{e}^{\mathrm{j}n'\theta}\delta_{l-s},$$

where θ is the reference phase of the Blaschke product $\varphi_p(z)$ while constructing the orthonormal bases of $\mathbb{C}^p$ (see section 4). This implies that n is restricted to $n = l + n'p$ in expression (5.6). Second, it is not difficult to show that

$$\left\langle \mathcal{B}_k\mathcal{B}_r, \mathcal{B}_{\min(k,r)+p+k'}\right\rangle = 0, \qquad \text{for } \; k' > 0,$$

because the integrand corresponding to this scalar product has no poles outside the unit circle so that the residue formula makes the associated integral vanish.

By putting these two results together, we have that n is restricted to $n = l + p$. Then, the proof is completed by making use of Lemma 10. □

As Assumption 3 has no consequences on the first p points ξ_i, the quantity $D_{(k,r),l}$ actually stands for a closed form expression of the last term in the RHS of the expression (5.6).

By complex conjugation, we similarly derive that

$$\langle \mathcal{B}_k \mathcal{B}_{-r}, \mathcal{B}_l \rangle = \sum_{i=0}^{p-1} \frac{1}{\gamma_p(\omega_i)} \mathcal{B}_k(\mathrm{e}^{\mathrm{j}\omega_i}) \overline{\mathcal{B}_l(\mathrm{e}^{\mathrm{j}\omega_i}) \mathcal{B}_r(\mathrm{e}^{\mathrm{j}\omega_i})} - \mathrm{e}^{-\mathrm{j}\theta} \overline{D}_{(l,r),k},$$

for $k \leq \min(l, r)$.

Finally, it is of interest to derive an upper bound on the modulus of $D_{(k,r),l}$ (identical to $D_{(r,k),l}$ by product commutativity) in the case that it is nonzero. This is the purpose of the two following lemmas.

LEMMA 12. *For $0 \leq l \leq k < p$, we have*

$$|D_{(k,l),l}| \leq c\, \beta_c^{-1} \eta^{(p-1)-k},$$

where $\beta_c^2 = 1 - c^2$ and $\eta = 2c/(1 + c^2) < 1$ as $0 \leq c < 1$.

Proof. This comes from the residue theorem. Without loss of generality, we shall again make use of Assumption 3. In detail, we successively have

$$\begin{aligned}
\overline{D}_{(k,l),l} &= \frac{1}{2\pi} \int_0^{2\pi} \overline{\mathcal{B}_k(\mathrm{e}^{\mathrm{j}\omega}) \mathcal{B}_l(\mathrm{e}^{\mathrm{j}\omega})} \mathcal{B}_{p+l}(\mathrm{e}^{\mathrm{j}\omega}) \mathrm{d}\omega \\
&= \frac{1}{2\pi \mathrm{j}} \oint_{|z|=1} \left[\frac{\beta_k}{1 - \overline{\xi}_k z} \prod_{i=k+1}^{p-1} \frac{z - \xi_i}{1 - \overline{\xi}_i z} \right] \frac{\beta_l^2 z}{(z - \xi_l)(1 - \overline{\xi}_l z)} \mathrm{d}z \\
&= \frac{\beta_k}{1 - \overline{\xi}_k \xi_l} \prod_{i=k+1}^{p-1} \frac{\xi_l - \xi_i}{1 - \overline{\xi}_i \xi_l} \xi_l
\end{aligned}$$

because of the contribution of the pole at $z = \xi_l$ only. Then, we easily find that

$$|D_{(k,l),l}| = \frac{\beta_k}{|1 - \overline{\xi}_k \xi_l|} \prod_{i=k+1}^{p-1} \left| \frac{\xi_l - \xi_i}{1 - \overline{\xi}_i \xi_l} \right| |\xi_l| \leq c\, \beta_c^{-1} \eta^{(p-1)-k}.$$

This completes the proof. ∎

Before going into the second lemma, let us introduce the following notation. While denoting by $\tilde{c}_i$ the ith largest value within the set of the zero moduli, i.e., $\{|\xi_k|, 0 \leq k < p\}$, we define

$$\zeta_m^-(\rho) = \prod_{i=1}^{m} \frac{\rho + \tilde{c}_i}{1 + \rho \tilde{c}_i} \quad \text{and} \quad \zeta_n^+(\rho) = \prod_{i=1}^{n} \frac{1 - \rho \tilde{c}_i}{\rho - \tilde{c}_i}. \tag{5.8}$$

Then, we have the following result.

LEMMA 13. *For $0 \leq l < r \leq k < p$, we have*

$$|D_{(k,r),l}| \leq K_c \min_{c < \rho \leq 1} \left[\zeta_m^-(\rho) \zeta_n^+(\rho) I(\rho, c) \right],$$

where $K_c = (1 + c)/(1 - c)$ while

$$I(\rho, c) = \frac{(1 - c)^2 \sqrt{1 - c^2}}{(1 - c\rho)^2 \sqrt{1 - c^2/\rho^2}}.$$

Further, the pair (m, n) (with $m \geq n$) is as follows:

$$(m, n) = \begin{cases} (p - k, r - l), & \text{if } p - k > r - l, \\ (r - l - 1, p - k - 1), & \text{otherwise.} \end{cases}$$

Proof. The derivation of the proof originates from the upper bounding of the value of the unimodular functions (appearing in the definition of the basis functions) outside the unit circle, i.e., $z = \rho\, \mathrm{e}^{\mathrm{j}\omega}$ with $\rho \neq 1$.

First, we can write

$$\overline{\mathcal{B}_k(z)\mathcal{B}_r(z)}\mathcal{B}_{p+l}(z) = \left[\frac{\beta_k z}{1-\overline{\xi}_k z}\prod_{i=k+1}^{p-1}\frac{z-\xi_i}{1-\overline{\xi}_i z}\right]\left[\frac{\beta_r z}{z-\xi_r}\frac{\beta_l}{1-\overline{\xi}_l z}\prod_{i=l}^{r-1}\frac{1-\overline{\xi}_i z}{z-\xi_i}\right].$$

Now, let us consider the case where $p-k > r-l$ for which we choose $c < \rho \leq 1$ in the definition of z. It is easily seen that

$$\frac{\rho-|\xi|}{1-\rho|\xi|} \leq \left|\frac{z-\xi}{1-\overline{\xi}z}\right| \leq \frac{\rho+|\xi|}{1+\rho|\xi|} \leq 1$$

for any $|\xi| \leq c$. The LHS (resp., RHS) inequality of this expression decreases (resp., increases) with $|\xi|$ going from zero to c. It yields

$$\begin{aligned}
|D_{(k,r),l}| &\leq \frac{1}{2\pi\mathrm{j}}\oint_{|z|=\rho}\left|\overline{\mathcal{B}_k(z)\mathcal{B}_r(z)}\mathcal{B}_{p+l}(z)\right|\frac{\mathrm{d}z}{z}\\
&\leq \zeta^-_{(p-1)-k}(\rho)\zeta^+_{r-l}(\rho)\frac{\rho^2}{2\pi\mathrm{j}}\oint_{|z|=\rho}\frac{\beta_k}{|1-\overline{\xi}_k z|}\frac{\beta_r}{|z-\xi_r|}\frac{\beta_l}{|1-\overline{\xi}_l z|}\frac{\mathrm{d}z}{z}\\
&\leq \zeta^-_{(p-1)-k}(\rho)\zeta^+_{r-l}(\rho)\frac{\beta_c^2\rho^2}{(1-c\rho)^2}\left[\frac{1}{2\pi}\int_0^{2\pi}\frac{\beta_r^2}{|\rho\mathrm{e}^{\mathrm{j}\omega}-\xi_r|^2}\mathrm{d}\omega\right]^{1/2}
\end{aligned}$$

by use of $\beta_i/|1-\overline{\xi}_i z| \leq \beta_c/(1-c\rho)$ for $|z| = \rho \in (c,1]$ and the Cauchy–Schwarz inequality for the last expression. With the help of the residue theorem

$$\frac{1}{2\pi}\int_0^{2\pi}\frac{\beta_i^2}{|\rho\mathrm{e}^{\mathrm{j}\omega}-\xi_i|^2}\mathrm{d}\omega = \frac{\beta_i^2}{\rho^2-|\xi_i|^2} \leq \frac{\beta_c^2}{\rho^2-c^2}, \tag{5.9}$$

we obtain

$$\begin{aligned}
|D_{(k,r),l}| &\leq \zeta^-_{p-k}(\rho)\zeta^+_{r-l}(\rho)\frac{\beta_c^3\rho}{(1-c\rho)^2\sqrt{\rho^2-c^2}}\left[\rho\,\frac{1+c\rho}{\rho+c}\right]\\
&\leq \zeta^-_{p-k}(\rho)\zeta^+_{r-l}(\rho)\frac{\beta_c^3}{(1-c\rho)^2\sqrt{1-c^2/\rho^2}} \qquad (5.10)
\end{aligned}$$

as the expression in brackets is less than or identical to one for all $\rho \in (c,1]$. This last expression is straightforwardly put into the form stated in the lemma.

In the case $p-k \leq r-l$, we choose $1 \leq \rho' < 1/c$ for defining $z = \rho'\mathrm{e}^{\mathrm{j}\omega}$. Then, by using the fact that

$$1 \leq \frac{\rho'+|\xi|}{1+\rho'|\xi|} \leq \left|\frac{z-\xi}{1-\overline{\xi}z}\right| \leq \frac{\rho'-|\xi|}{1-\rho'|\xi|}$$

for which the LHS (resp., RHS) inequality of this expression decreases (resp., increases) with $|\xi|$ going from zero to c, we derive that

$$\begin{aligned}
|D_{(k,r),l}| &\leq \zeta^+_{(p-1)-k}\left(\frac{1}{\rho'}\right)\zeta^-_{(r-l)-1}\left(\frac{1}{\rho'}\right)\frac{(\rho')^2}{2\pi\mathrm{j}}\oint_{|z|=\rho'}\frac{\beta_k}{|1-\overline{\xi}_k z|}\frac{\beta_r}{|z-\xi_r|}\frac{\beta_l}{|z-\xi_l|}\frac{\mathrm{d}z}{z}\\
&\leq \zeta^+_{(p-1)-k}\left(\frac{1}{\rho'}\right)\zeta^-_{(r-l)-1}\left(\frac{1}{\rho'}\right)\frac{\beta_c^2(\rho')^2}{(\rho'-c)^2}\left[\frac{1}{2\pi}\int_0^{2\pi}\frac{\beta_k^2}{|1-\overline{\xi}_k\rho'\mathrm{e}^{\mathrm{j}\omega}|^2}\mathrm{d}\omega\right]^{1/2}.
\end{aligned}$$

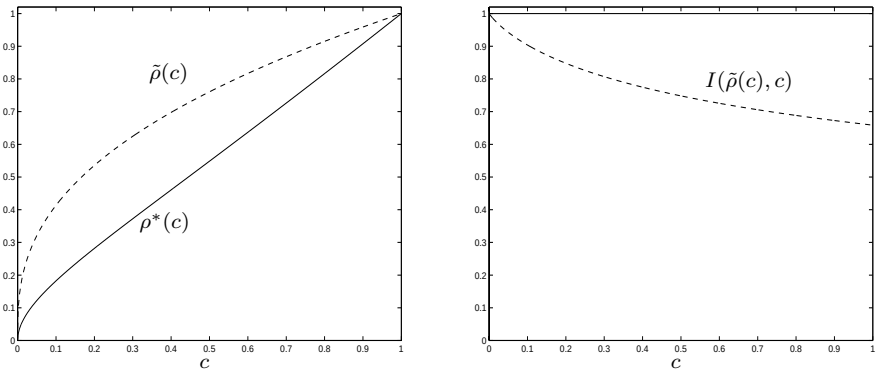


FIG. 5.2. *Characteristics of the convex function $I(\rho, c)$ as functions of c . Left: $\rho^*(c)$ such that $I(\rho^*, c) = 1$ (—) and $\tilde{\rho}(c)$ such that $I(\tilde{\rho}, c)$ is minimal (−−). Right: The value of $I(\tilde{\rho}(c), c)$ (−−).*

From this with $\rho = 1/\rho'$, we obtain that

$$|D_{(k,r),l}| \leq \zeta^+_{(p-1)-k}(\rho)\zeta^-_{(r-l)-1}(\rho)\frac{\beta_c^3}{(1-c\rho)^2\sqrt{1-c^2/\rho^2}}. \tag{5.11}$$

Finally, as ρ is chosen arbitrarily in the interval $(c, 1]$, we can ask for the minimum of the RHS expression (5.10) and (5.11) with respect to such ρ. This completes the proof. □

A characterization of the function $I(\rho, c)$ is given in the following proposition.

PROPOSITION 14. *The function $I(\rho, c)$ is convex for $c < \rho \leq 1$. Furthermore, $I(\rho, c) = 1$ at $\rho^* \in (c, 1)$ for appropriate values of $\rho^* = \rho^*(c)$. Thus, $I(\rho, c) < 1$ for $\rho \in (\rho^*, 1)$.* □

In the left part of Figure 5.2, we have drawn the function $\rho^*(c)$ as well as the value of $\rho \in [\rho^*(c), 1]$, denoted $\tilde{\rho}(c)$, at which $I(\rho, c)$ is minimal. In its right part, we have displayed the value of this minimum, i.e., $I(\tilde{\rho}(c), c)$.

From this proposition, we can state the following lemma.

LEMMA 15. *Let us take the same notation as in Lemma* 13 *and in Proposition* 14. *Then,*

$$|D_{(k,r),l}| \leq K_c\ \eta_0^{m-\min(m,n\epsilon_0)},$$

where $\rho_0 \in [\rho^*, 1)$, $\eta_0 = (\rho_0 + c)/(1 + \rho_0 c) < 1$ and $\epsilon_0 = \epsilon(\rho_0, c)$ is defined as

$$\epsilon(\rho_0, c) = \left\lceil \frac{\ln(1-\rho_0 c) - \ln(\rho_0 - c)}{\ln(1+\rho_0 c) - \ln(\rho_0 + c)} \right\rceil$$

with $\lceil x \rceil$ denoting the smallest integer not less than x.

Proof. First, from the discussion about the function $I(\rho, c)$, we have that $I(\rho_0, c) \leq 1$. Thus, Lemma 13 yields

$$|D_{(k,r),l}| \leq K_c\ \min(1, \zeta^-_m(\rho_0)\zeta^+_n(\rho_0)).$$

Then, by the definition of $\zeta^-_m(\rho)$ and $\zeta^+_n(\rho)$ (see expression (5.8)), the second argument in the RHS minimization is upper bounded as

$$\zeta^-_m(\rho_0)\zeta^+_n(\rho_0) \leq \left(\frac{\rho_0 + c}{1 + \rho_0 c}\right)^m \left(\frac{1-\rho_0 c}{\rho_0 - c}\right)^n.$$

It is easily seen that the RHS expression is less (or identical) to one for $m = n\epsilon_0$ as defined above. Hence, for $m \geq n\epsilon_0$, we successively have

$$\begin{aligned}\zeta_m^-(\rho_0)\zeta_n^+(\rho_0) &= \left[\zeta_m^-(\rho_0)/\zeta_{n\epsilon_0}^-(\rho_0)\right]\left[\zeta_{n\epsilon_0}^-(\rho_0)\zeta_n^+(\rho_0)\right] \\ &\leq \zeta_m^-(\rho_0)/\zeta_{n\epsilon_0}^-(\rho_0) \leq \eta_0^{m-n\epsilon_0}.\end{aligned}$$

Together with the above RHS minimization, this completes the proof. ∎

In summary, in this section we have given exact quadrature formulas (5.7) for the integral of $\mathcal{B}_k\mathcal{B}_r\overline{\mathcal{B}}_l$ if $l > \min(k, r)$. In the case where $l \leq \min(k, r)$, we assume a cyclic repetition of the ξ_k (Assumption 3), and then some error term $D_{(k,r),l}$ appears, which has been bounded in subsequent lemmas. This bound depends on how well the ξ_k are bounded away from the unit circle, which is measured by the paramater c from Assumption 1.

Similar results were obtained for integrals of the form $\mathcal{B}_k\mathcal{B}_{-r}\overline{\mathcal{B}}_l$, in which case a distinction has to be made between $k > \min(l, r)$ and $k \leq \min(l, r)$.

6. General Toeplitz matrix approximation. In this section, we consider general Toeplitz matrices constructed by use of a 2π-periodic function $f(\omega)$. Such a matrix was introduced in section 1 as

$$M_p(f) = \frac{1}{2\pi}\int_0^{2\pi} \Gamma_p(\omega) f(\omega) \Gamma_p^*(\omega) \mathrm{d}\omega.$$

With the help of the quadrature formulas derived in the preceding section, we shall show that any general Toeplitz matrix can be approximated by the matrix $\Upsilon_p F_p \Upsilon_p^*$ of (1.9). Note that the eigenpairs of this matrix are precisely the orthonormal vectors $\tilde{\Gamma}_p(\omega_i)$ with corresponding eigenvalue $f_p(\omega_i)$ (see definition (1.8)) for those particular ω values $\omega_i \in \Omega_p(\theta)$ (defined in (4.2)) that were generalizations of the pth roots of unity. The eigenvalue decomposition is in fact the quadrature formula (1.9), namely,

$$\Upsilon_p F_p \Upsilon_p^* = \sum_{i=0}^{p-1} \tilde{\Gamma}_p(\omega_i) f_p(\omega_i) \tilde{\Gamma}_p^*(\omega_i). \tag{6.1}$$

The approximation error is written as

$$\Delta_p(f) = M_p(f) - \Upsilon_p F_p \Upsilon_p^*.$$

Its (k, l)-element is written as

$$\begin{aligned}[\Delta_p(f)]_{k,l} &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{B}_k(\mathrm{e}^{\mathrm{j}\omega}) f(\omega) \overline{\mathcal{B}_l(\mathrm{e}^{\mathrm{j}\omega})} \mathrm{d}\omega - \sum_{i=0}^{p-1} \frac{1}{\gamma_p(\omega_i)} \mathcal{B}_k(\mathrm{e}^{\mathrm{j}\omega_i}) f_p(\omega_i) \overline{\mathcal{B}_l(\mathrm{e}^{\mathrm{j}\omega_i})} \\ &= -\sum_{r=0}^{p-1} \left(a_r \mathrm{e}^{\mathrm{j}\theta} D_{(k,r),l} + a_{-r} \mathrm{e}^{-\mathrm{j}\theta} \overline{D}_{(l,r),k}\right), \end{aligned} \tag{6.2}$$

where we have used the fact that $M_p(e_p) = 0$ by the residue theorem (recall $e_p = f - f_p$). So, it is seen that the closer the position of the $\Delta_p(f)$-elements to the top-right and/or bottom-left corners of the matrix, the larger their modulus because of the contribution of larger $|D_{(k,r),l}|$ terms.

Finally, for future use, let us upper bound the scalar quantities derived from the error matrix $\Delta_p(f)$. The following results hold.

LEMMA 16. *Given expression* (6.2) *together with Lemma* 15*, we have that*

$$\left|\tilde{\Gamma}_p^*(\sigma)\Delta_p(f)\tilde{\Gamma}_p(\mu)\right| \leq K_c^2 K_{f,p}/p \qquad \textit{for any} \quad \sigma,\mu \in [0,2\pi) \tag{6.3}$$

as well as

$$\|\Delta_p(f)\|_F \leq K_{f,p} \quad \textit{and} \quad \|\Delta_p(f)\tilde{\Gamma}_p(\omega)\|_2 \leq K_c K_{f,p}/p^{1/2} \qquad \textit{for} \quad \omega \in [0,2\pi),$$

where $\|.\|_F$ *denotes the Frobenius norm operator while*

$$K_{f,p} = 2K_c \sum_{r=0}^{p-1} K_0(r) \max(|a_r|,|a_{-r}|)$$

with $K_0(r) = [r^2(\epsilon_0 - 1) + (2r+1)/(1-\eta_0)]$.

Proof. First, we write

$$\begin{aligned}\left|\tilde{\Gamma}_p^*(\sigma)\Delta_p(f)\tilde{\Gamma}_p(\mu)\right| &\leq \sum_{k,l} \left|\left[\tilde{\Gamma}_p^*(\sigma)\right]_k [\Delta_p(f)]_{k,l} \left[\tilde{\Gamma}_p(\mu)\right]_l\right| \\ &\leq \max_{\omega,l} \frac{\left|\mathcal{B}_l(\mathrm{e}^{\mathrm{j}\omega})\right|^2}{\gamma_p(\omega)} \sum_{k,l} \left|[\Delta_p(f)]_{k,l}\right| \leq \frac{K_c^2 K_{f,p}}{p},\end{aligned}$$

where $K_{f,p}$ is evaluated as follows:

$$\begin{aligned}\sum_{k,l} \left|[\Delta_p(f)]_{k,l}\right| &\leq \sum_{r=0}^{p-1} \tilde{a}_r \sum_{k,l} \left(|D_{(k,r),l}| + |D_{(l,r),k}|\right) \leq 2\sum_{r=0}^{p-1} \tilde{a}_r \sum_{k,l} |D_{(k,r),l}| \\ &\leq 2K_c \sum_{r=0}^{p-1} K_0(r)\tilde{a}_r =: K_{f,p},\end{aligned}$$

where $\tilde{a}_r = \max(|a_r|,|a_{-r}|)$. This last expression has been obtained by taking advantage of Figure 5.1 in order to evaluate the contribution of the nonzero elements $D_{(k,r),l}$. It is found that

$$\begin{aligned}\sum_{k,l} |D_{(k,r),l}| &= \sum_{l=0}^{r-1} \sum_{k=l+1}^{p-1} |D_{(k,r),l}| + \sum_{k=r}^{p-1} |D_{(k,r),r}| + \sum_{l=0}^{r-1} |D_{(l,r),l}| \\ &\leq \sum_{l=0}^{r-1} \left[\sum_{k=l+1}^{(p-1)-(r-l)} |D_{(k,r),l}| + \sum_{k=(p-1)-(r-l)+1}^{p-1} |D_{(k,r),l}| \right] \\ &\qquad + c\,\beta_c^{-1} \left[\frac{1}{1-\eta} + r\eta^{(p-1)-r}\right] \\ &\leq K_c r \left[r(\epsilon_0 - 1) + \frac{1+\eta_0}{1-\eta_0}\right] + c\,\beta_c^{-1} \left[\frac{1}{1-\eta} + r\eta^{(p-1)-r}\right] \\ &\leq K_c \left[r^2(\epsilon_0 - 1) + \frac{2r+1}{1-\eta_0}\right] \leq K_c K_0(r),\end{aligned}$$

where we have used Lemma 12 and the results in Lemma 15 as follows:

$$\sum_{l=0}^{r-1} \sum_{k=l+1}^{(p-1)-(r-l)} |D_{(k,r),l}| \leq K_c \sum_{n=1}^{r} \sum_{m=n+1}^{(p-1)-(r-n)} \eta_0^{m-\min(m,n\epsilon_0)}$$

$$\leq K_c \sum_{n=1}^{r} \left[n(\epsilon_0 - 1) + \sum_{m=n\epsilon_0+1}^{(p-1)-(r-n)} \eta_0^{m-n\epsilon_0} \right]$$
$$\leq K_c\, r \left[\frac{r+1}{2}(\epsilon_0 - 1) + \frac{\eta_0}{1-\eta_0} \right]$$

as well as

$$\sum_{l=0}^{r-1} \sum_{k=(p-1)-(r-l)+1}^{p-1} |D_{(k,r),l}| \leq K_c \sum_{n=0}^{r-1} \sum_{m=n}^{r-1} \eta_0^{m-\min(m,n\epsilon_0)}$$
$$\leq K_c \sum_{n=0}^{r-1} \left[n(\epsilon_0 - 1) + \sum_{m=n\epsilon_0}^{r-1} \eta_0^{m-n\epsilon_0} \right]$$
$$\leq K_c r \left[\frac{r-1}{2}(\epsilon_0 - 1) + \frac{1}{1-\eta_0} \right].$$

Similarly, it is easy to show that the Frobenius norm of $\Delta_p(f)$ is readily upper bounded by $K_{f,p}$ as well as

$$\left\| \Delta_p(f) \tilde{\Gamma}_p(\omega) \right\|_2 \leq K_c K_{f,p} / p^{1/2}$$

so that the proof is completed. □

Note that this lemma puts into light the dependence between the asymptotic (with p) behavior of the upper bound $K_{f,p}$ and the regularity of the underlying function $f(\omega)$ that is traced in the convergence properties of the decomposition coefficients a_r and a_{-r} (see Theorem 1).

It is also worth mentioning that the (polynomial) upper bound $K_0(r)$ depends on the value of $\rho_0 \in [\rho^*, 1)$ on which a minimization could be performed.

7. General Toeplitz matrix functional. In the present section, we give a proof of Theorem 4. So we consider evaluating quantities of the form

$$\tilde{\Gamma}_p^*(\sigma) M_p(f) \tilde{\Gamma}_p(\mu),$$

making use of the results obtained in the preceding sections. Therefore, we first choose the reference phase θ of the Blaschke product $\varphi_p(e^{j\omega_i})$ so that the frequency point μ belongs to $\Omega_p(\theta)$. In other words, we impose $\tilde{\Gamma}_p(\mu)$ to be one of the column of the related unitary matrix Υ_p.

Then, we immediately have that

$$\tilde{\Gamma}_p^*(\sigma) M_p(f) \tilde{\Gamma}_p(\mu) = f_p(\mu) \tilde{\Gamma}_p^*(\sigma) \tilde{\Gamma}_p(\mu) + \tilde{\Gamma}_p^*(\sigma) \Delta_p(f) \tilde{\Gamma}_p(\mu),$$

where we have used the fact that $[\Upsilon_p F_p \Upsilon_p^*] \tilde{\Gamma}_p(\mu) = f_p(\mu) \tilde{\Gamma}_p(\mu)$. With the help of the upper bound (6.3), the last term in the RHS converges to zero like $1/p$ while the first one is identical to $f(\mu)$ if $\sigma = \mu$. If $\sigma \neq \mu$, this term decreases like $1/p$ by use of the result derived in Lemma 8. More generally, we can derive the following result.

LEMMA 17. *Let Assumptions* 1 *and* 2 *hold. Then*

$$\lim_{p\to\infty} \tilde{\Gamma}_p^*(\sigma) T(M_p(f)) \tilde{\Gamma}_p(\mu) = \begin{cases} T(f(\mu)), & \text{if } \sigma = \mu, \\ 0, & \text{otherwise}, \end{cases} \tag{7.1}$$

where $T(\cdot)$ is an analytic function (i.e., having convergent Taylor expansion) on the range of $f(\omega)$. The convergence rate to the limiting value is at least as fast as $1/p$ for $\sigma = \mu$ and as $\ln p/p$ in the other cases.

Proof. First, we look for the nth term in the Taylor expansion of $T(x)$, say $T_n x^n/n!$. Therefore, we consider $M_p^n(f) := [M_p(f)]^n$ and $f_p^n(\mu) := [f_p(\mu)]^n$ and show that $\tilde{\Gamma}_p^*(\sigma) M_p^n(f) \tilde{\Gamma}_p(\mu)$ converges to $f_p^n(\mu)\tilde{\Gamma}_p^*(\sigma)\tilde{\Gamma}_p(\mu)$. Indeed, we have that

$$\begin{aligned}
&\left|\tilde{\Gamma}_p^*(\sigma)\Big(M_p^n(f) - f_p^n(\mu) I_p\Big)\tilde{\Gamma}_p(\mu)\right| \\
&\quad = \left|\tilde{\Gamma}_p^*(\sigma)\Big(\big(\Upsilon_p F_p \Upsilon_p^* + \Delta_p(f)\big)^n - \Upsilon_p F_p^n \Upsilon_p^*\Big)\tilde{\Gamma}_p(\mu)\right| \\
&\quad \le \sum_{i=0}^{p-1} |\alpha_{p,i}| \left|\tilde{\Gamma}_p^*(\omega_i)\Big(\big(\Upsilon_p F_p \Upsilon_p^* + \Delta_p(f)\big)^n - \Upsilon_p F_p^n \Upsilon_p^*\Big)\tilde{\Gamma}_p(\mu)\right| \\
&\quad \le \sum_{i=0}^{p-1} |\alpha_{p,i}| \left[n\|F_p\|_2^{n-1} \left|\tilde{\Gamma}_p^*(\omega_i)\Delta_p(f)\tilde{\Gamma}_p(\mu)\right| \right. \\
&\qquad\qquad \left. + \sum_{k=2}^{n} C_k^n \|F_p\|_2^{n-k} \|\Delta_p(f)\|_2^{k-2} \|\Delta_p^*(f)\tilde{\Gamma}_p(\omega_i)\|_2 \|\Delta_p(f)\tilde{\Gamma}_p(\mu)\|_2 \right] \\
&\quad \le (K_c^2/p)\left[(f_{p,+} + K_{f,p})^n - f_{p,+}^n\right] \sum_{i=0}^{p-1} |\alpha_{p,i}|,
\end{aligned}$$

where I_p denotes the p-dimensional identity matrix, $f_{p,+} = \max_\omega |f_p(\omega)|$ and $\alpha_{p,i} = \tilde{\Gamma}_p^*(\sigma)\tilde{\Gamma}_p(\omega_i)$ while $C_k^n = n!/k!(n-k)!$ is the binomial function. For bounding the contribution of the $\alpha_{p,i}$'s, we used the results of section 4: namely, the 1-norm of the $\alpha_{p,i}$ sequence is identical to one for $\sigma = \mu$ and diverges at most like $\ln p$ in the other cases. By use of the results in Lemma 16 together with Theorem 1 under Assumption 2, it is easily shown that the term $K_{f,p}$ has bounded limiting values for increasing p. Putting the Taylor expansion terms together, we obtain that

$$\tilde{\Gamma}_p^*(\sigma) T(M_p(f)) \tilde{\Gamma}_p(\mu) = \begin{cases} T(f_p(\mu)) + O(1/p), & \text{if } \sigma = \mu, \\ O(\ln p/p), & \text{otherwise.} \end{cases}$$

Finally, as per Assumption 2, the approximation function $f_p(\omega)$ uniformly converges to $f(\omega)$ for unbounded p (see Corollary 2), so does $T(f_p(\omega))$ to $T(f(\omega))$. This completes the proof. □

It can also be shown that

$$\lim_{p\to\infty} \tilde{\Gamma}_p^*(\sigma) M_p(f) M_p(g) \tilde{\Gamma}_p(\mu) = \begin{cases} f(\mu)g(\mu), & \text{if } \sigma = \mu, \\ 0, & \text{otherwise,} \end{cases}$$

where $g(\omega)$ is a 2π-periodic function satisfying Assumption 2 and for which we define a diagonal matrix G_p and a residual matrix $\Delta_p(g)$ having the same properties as those related to the function $f(\omega)$. This is done as follows:

$$\begin{aligned}
&\left|\tilde{\Gamma}_p^*(\sigma)\Big(M_p(f)M_p(g) - f_p(\mu)g_p(\mu)\Big)\tilde{\Gamma}_p(\mu)\right| \\
&\quad \le \left|\tilde{\Gamma}_p^*(\sigma)\Big(\big(\Upsilon_p F_p \Upsilon_p^* + \Delta_p(f)\big)\big(\Upsilon_p G_p \Upsilon_p^* + \Delta_p(g)\big) - \Upsilon_p F_p G_p \Upsilon_p^*\Big)\tilde{\Gamma}_p(\mu)\right| \\
&\quad \le (K_c^2/p)\left[K_{g,p} f_{p,+} \sum_{i=0}^{p-1} |\alpha_{p,i}| + K_{f,p} g_{p,+} + K_{f,p} K_{g,p}\right]
\end{aligned}$$

with the convergence rate at least as fast as $1/p$ when $\sigma = \mu$ and as $\ln p/p$ in the other cases. Note that this last expression is not symmetric with respect to f and g. The reason for this is that $\mu \in \Omega_p(\theta)$ while nothing similar has been assumed for σ; see also [14, section 5].

Finally, the more general result of Theorem 4 follows easily. It should now be clear that

$$\lim_{p\to\infty} \tilde{\Gamma}_p^*(\sigma) T\Big(M_p(f^{[1]}), \ldots, M_p(f^{[n]})\Big) \tilde{\Gamma}_p(\mu) = \begin{cases} T\left(f^{[1]}, \ldots, f^{[n]}\right)(\mu), & \text{if } \sigma = \mu, \\ 0, & \text{otherwise,} \end{cases}$$

where $T(\cdot)$ is a multivariable analytic function on the range of the functions $f^{[i]}(\omega)$ satisfying Assumption 2. The same remark holds for the convergence rate to the limiting values.

8. General Toeplitz matrix spectrum. For what concerns the spectrum of the general Toeplitz matrix $M_p(f)$, it is possible to derive a result that stipulates that the distribution of its eigenvalues is asymptotically (with p) given by the distribution of those in the diagonal approximation matrix F_p. Remember that these diagonal elements are eigenvalues which are values of the underlying function $f_p(\omega)$ for appropriate frequency points ω_i. Furthermore, a closed form expression can be derived for the asymptotic distribution of these particular values $f_p(\omega_i)$. Here, we assume that the function $f(\omega)$ is not only 2π-periodic but also positive real-valued. We now give the proof of Theorem 5.

Proof of Theorem 5. This result is obtained using the same arguments as for the case of equivalent Toeplitz matrices (see Grenander and Szegő [11, Chap. 5]). Let us expose it in some detail.

First, we denote the ith (in any order) eigenvalue of any Hermitian matrix X by $\lambda_i(X)$. Then, it can be written that

$$\begin{aligned} \frac{1}{p}\sum_{i=0}^{p-1} \lambda_i\left(M_p(f)\right) &= \frac{1}{p}\mathrm{tr}\left(\Upsilon_p^* M_p(f) \Upsilon_p\right) = \frac{1}{p}\left[\sum_{i=0}^{p-1} f_p(\omega_i) + \mathrm{tr}\left(\Upsilon_p^* \Delta_p(f) \Upsilon_p\right)\right] \\ &= \sum_{i=0}^{p-1} \frac{f_p(\omega_i)}{p} + O\left(\frac{1}{p}\right), \end{aligned}$$

where we have used results from the preceding section for bounding the contributions of the error matrix $\Delta_p(f)$, i.e., $p|\tilde{\Gamma}_p^*(\omega)\Delta_p(f)\tilde{\Gamma}_p(\omega)| \le K_c^2 K_{f,p}$ that is upper bounded for all p under Assumption 2 (see Lemma 16 with Theorem 1).

Furthermore, such an expression also holds for any analytic function of the eigenvalues, i.e.,

$$\frac{1}{p}\sum_{i=0}^{p-1} T\left(\lambda_i\left(M_p(f)\right)\right) = \sum_{i=0}^{p-1} \frac{T(f_p(\omega_i))}{p} + O\left(\frac{1}{p}\right), \tag{8.1}$$

where $T(\cdot)$ is analytic over the interval $[f_{p,-}, f_{p,+}] = [\min_\omega f(\omega), \max_\omega f(\omega)]$. In fact, by Weierstrass–Stone's theorem (see, e.g., Gray [10]), this result can be extended to any continuous function over the interval $[f_{p,-}, f_{p,+}]$.

Now, let us write this result in a more convenient way. Therefore, remember from section 4 that the frequency point ω_i corresponds to $\chi_p^{-1}(\theta - i2\pi)$ where $\theta \in [0, 2\pi)$ denotes the reference phase taken for the Blaschke product. This implies that

$$\sum_{i=0}^{p-1} \frac{T(f(\omega_i))}{p} = \frac{1}{2\pi}\sum_{i=0}^{p-1} T\left(f_p\left(\chi_p^{-1}\left(p\left[\frac{\theta}{p} - i\Delta\right]\right)\right)\right)\Delta,$$

where $\Delta = 2\pi/p$. This expression is obviously the discretization of the following integral:

$$\frac{1}{2\pi}\int_{\theta/p-2\pi}^{\theta/p} T\left(f_p\left(\chi_p^{-1}\left(p\omega\right)\right)\right)\mathrm{d}\omega.$$

By continuity of the functions involved and the uniform convergence of the approximation $f_p(\omega)$ to $f(\omega)$ for unbounded p (see Corollary 2 under Assumption 2), we finally end up with an asymptotic expression of the distribution of the eigenvalues of the general Toeplitz matrix $M_p(f)$. It is written as

$$\lim_{p\to\infty}\frac{1}{p}\sum_{i=0}^{p-1} T\left(\lambda_i\left(M_p(f)\right)\right) = \frac{1}{2\pi}\int_0^{2\pi} T\left(f\left(\tilde{\chi}^{-1}\left(\omega\right)\right)\right)\mathrm{d}\omega,$$

where $\tilde{\chi}(\omega)$ satisfies $\tilde{\chi}^{-1}(\omega) = \lim_{p\to\infty}\chi_p^{-1}(p\omega)$ (see section 4). □

REFERENCES

[1] N. Achieser, *Theory of Approximation*, Frederick Ungar, New York, 1956.

[2] N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68 (1950), pp. 337–404.

[3] A. Bultheel, P. González-Vera, E. Hendriksen, and O. Njåstad, *Quadrature formulas on the unit circle based on rational functions*, J. Comput. Appl. Math., 50 (1994), pp. 159–170.

[4] A. Bultheel, P. González-Vera, E. Hendriksen, and O. Njåstad, *Rates of convergence of multipoint rational approximants and quadrature formulas on the unit circle*, J. Comput. Appl. Math., 77 (1997), pp. 77–102.

[5] A. Bultheel, P. González-Vera, E. Hendriksen, and O. Njåstad, *Orthogonal rational functions*, Cambridge Monogr. Appl. Comput. Math. 5, Cambridge University Press, Cambridge, UK, 1999.

[6] A. Bultheel, P. González-Vera, E. Hendriksen, and O. Njåstad, *Quadrature and orthogonal rational functions*, J. Comput. Appl. Math., 127 (2001), pp. 67–91.

[7] E. Cheney, *Introduction to Approximation Theory*, McGraw–Hill, New York, Toronto, London, 1966.

[8] R. Edwards, *Fourier series: A modern introduction*, Vol. 1, Springer–Verlag, New York, 1979.

[9] R. Edwards, *Fourier series: A modern introduction*, Vol. 2, Springer–Verlag, New York, 1979.

[10] R. Gray, *On the asymptotic eigenvalue distribution of Toeplitz matrices*, IEEE Trans. Inform. Theory, 18 (1972), pp. 725–730.

[11] U. Grenander and G. Szegő, *Toeplitz forms and their applications*, Chelsea Publishing, New York, 1958.

[12] B. Ninness, *The Utility of Orthonormal Bases*, Technical report EE9802, Department of Electrical Engineering, University of Newcastle, Newcastle, NSW, Australia, 1998.

[13] B. Ninness and F. Gustafsson, *A unified construction of orthogonal bases for system identification*, IEEE Trans. Automat. Control, 42 (1997), pp. 515–522.

[14] B. Ninness, H. Hjalmarsson, and F. Gustafsson, *The fundamental role of general orthogonal bases in system identification*, IEEE Trans. Automat. Control, 44 (1999), pp. 1384–1407.

[15] B. Ninness, H. Hjalmarsson, and F. Gustafsson, *Generalized Fourier and Toeplitz results for rational orthonormal bases*, SIAM J. Control Optim., 37 (1998), pp. 429–460.

[16] J. Walsh, *Interpolation and Approximation*, 3rd ed., Amer. Math. Soc. Colloq. Publ. 20, AMS, Providence, RI, 1960.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 5, pp. 1440–1454

EXISTENCE FOR SHAPE OPTIMIZATION PROBLEMS IN ARBITRARY DIMENSION*

W. B. LIU†, P. NEITTAANMÄKI‡, AND D. TIBA§

Abstract. We discuss some existence results for optimal design problems governed by second order elliptic equations with the homogeneous Neumann boundary conditions or with the interior transmission conditions. We show that our continuity hypotheses for the unknown boundaries yield the compactness of the associated characteristic functions, which, in turn, guarantees convergence of any minimizing sequences for the first problem. In the second case, weaker assumptions of measurability type are shown to be sufficient for the existence of the optimal material distribution. We impose no restriction on the dimension of the underlying Euclidean space.

Key words. uniform segment property, compactness, existence of optimal shapers

AMS subject classifications. 49D37, 65K10

PII. S0363012901388142

1. Introduction.

In this paper, we study existence for two shape optimization problems. The first is the following optimal shape design problem:

$$\min_{\Omega\in\mathcal{O}} \int_\Omega (y-y_d)^2\,dx \tag{1.1}$$

(SONB)

$$-\Delta y + y = f, \quad \frac{\partial y}{\partial n}|_{\partial\Omega} = 0, \tag{1.2}$$

where $\mathcal{O}$ is a class of admissible open sets inside a fixed open set D in $\mathbf{R}^m$, and $f, y_d \in L^2(D)$.

The second problem is the following material distribution design problem:

$$\min_{\Omega\in\mathcal{O}} \int_{E\cap\Omega} |y_1 - z_d|^2\,dx + \int_{E\cap(D\setminus\Omega)} |y_2 - z_d|^2\,dx \tag{1.3}$$

(SOTB)

$$-a_1\Delta y_1 + b_1 y_1 = f \quad \text{in } \Omega, \tag{1.4}$$

$$-a_2\Delta y_2 + b_2 y_2 = f \quad \text{in } D\setminus\overline{\Omega}, \tag{1.5}$$

$$a_1\frac{\partial y_1}{\partial n} = a_2\frac{\partial y_2}{\partial n}, \quad y_1 = y_2 \quad \text{in } \partial\Omega\setminus(\partial\Omega\cap\partial D), \tag{1.6}$$

$$a_i\frac{\partial y_i}{\partial n} = 0 \quad \text{in } \Gamma_1, \qquad y_i = 0 \quad \text{in } \Gamma_2, \quad i = 1,2, \tag{1.7}$$

where $E \subset D$ are two given bounded domains in $\mathbf{R}^m$, $\Gamma_1\cup\Gamma_2 = \partial D$ with $\Gamma_1\cap\Gamma_2 = \emptyset$, $z_d \in L^2(E)$, and $\mathcal{O}$ is a class of admissible open sets inside D. The details of the above two problems will be specified in sections 3 and 4.

It is well known that, in general, such shape optimization problems have no solutions without assuming further regularity conditions on the boundaries of the

*Received by the editors April 18, 2001; accepted for publication (in revised form) March 28, 2002; published electronically January 3, 2003.
http://www.siam.org/journals/sicon/41-5/38814.html
†CBS, University of Kent, Canterbury, CT2 7PE, UK (W.B.Liu@ukc.ac.uk).
‡Department of Mathematical Information Technology, University of Jyväskylä, P.O. Box 35, FIN–40351 Jyväskylä, Finland (pn@mit.jyu.fi).
§Institute of Mathematics, Romanian Academy, P.O. Box 1–764, RO–70700 Bucharest, Romania (Dan.Tiba@imar.ro).

domain classes; see Pironneau [20] for some counterexamples. Assuming the cone property on $\mathcal{O}$ uniformly, for example, one can prove that the above optimal shape design problem indeed has solutions; see Chenais [6] and Pironneau [20] for the details. Furthermore, much effort has been devoted in the scientific literature to the relaxation of the regularity conditions required for the boundaries of the unknown domains in optimal design problems. This question is discussed in detail in the monographs by Pironneau [20], Haslinger and Neittaanmäki [8], Sokolowski and Zolesio [21], and Tiba [24], for instance. As the range of optimal design problems is very wide, including as well control into coefficients problems, optimization of certain evolution systems, some problems originating in mechanics, etc., there is a rich variety of existence results of interest. We quote here just the recent papers by Sverak [23], Bucur and Zolesio [4, 3, 5], and Henrot [9], where the question of the dependence of solutions of elliptic equations on the underlying domain of definition is discussed in a general setting and various sufficient compactness conditions are introduced. However, a complete solution of the problem seems not to be known, to our knowledge.

In this work, we first prove existence for the above optimal shape design problem governed by the Neumann boundary value problems, under the mere assumption that the unknown open sets are of class $\mathcal{C}$ (or, equivalently, they have the segment property—see Maz'ja [17] and Adams [1]) with some uniformity with respect to the parameters—see section 3 for the details. Our conditions allow cusps or certain oscillations of the boundaries, but cracks or oscillations dense in a set of positive measure (in the sense of Hausdorff–Pompeiu) are not permitted. Then, in section 4, it is shown that, for the material distribution problem, i.e., in the transmission boundary value problems, much weaker assumptions of measurability type are sufficient to obtain existence of the optimal sets. Moreover, all of our results are valid in any space dimension. This is an advantage over much of the existing literature, where very often the case of space dimension two is studied.

The approach that we are using is described in detail in section 2 and has its origin in our previous works—Liu [13], Liu and Rubio [15], Mäkinen, Neittaanmäki, and Tiba [16], and Neittaanmäki and Tiba [19]. Roughly speaking, we replace the extension technique for passing to the limit in the PDEs defined in a sequence of open sets by a local convergence analysis (see Lemma 3.2 and its proof). For set convergence, we introduce a concept of parametric convergence, which can be easily adapted to various possible representations of open sets and preserves some needed properties. As an example, the Hausdorff–Pompeiu convergence is a special case of the parametric convergence, choosing a certain distance function as the parametric representation. Notice that this is essentially different from the one used by Sverak [23]; see Proposition 2.5 and the subsequent remark.

It is recognized in the scientific literature that the a.e. convergence of the corresponding characteristic functions is an essential step in any convergence result for the PDEs defined in a sequence of open sets. Our treatment of this question, appearing mainly in sections 2 and 4, is based on a new technique using the maximal monotone extension of the Heaviside mapping in $\mathbf{R} \times \mathbf{R}$ and the closure properties of monotone operators. We also propose, in this setting, a new approximation procedure for the characteristic functions by means of the Yosida approximation and of the Friedrichs mollifiers. In this respect, we point out the constructive character of our method. Some numerical experiments together with an approximation result are reported in Mäkinen, Neittaanmäki, and Tiba [16].

Finally, we mention that, in the recent paper by Sprekels and Tiba [22], some

design problems, which are formulated as control into coefficients problems, are discussed. It is shown (by different methods) that the boundedness of the coefficients is sufficient to prove existence. An announcement of some of the results from the present work was published in Liu, Neittaanmäki, and Tiba [14] without proofs.

2. Convergence of open sets and of mappings. Let A, B be two open sets contained in the bounded domain D of $\mathbf{R}^m$, $m \in N$. The distance δ between A and B is defined by

$$\rho(A,B) = \sup_{x\in\bar{D}-A} \inf_{y\in\bar{D}-B} \|x-y\|_{\mathbf{R}^m}, \tag{2.1}$$

$$\delta(A,B) = \max\{\rho(A,B), \rho(B,A)\}, \tag{2.2}$$

and it is the Hausdorff–Pompeiu distance between the closed sets $\bar{D} \setminus A$ and $\bar{D} \setminus B$; see Pironneau [20] and Kuratowski [11]. We shall denote by Hlim the limit in the sense of Hausdorff–Pompeiu.

Another frequently used distance notion is

$$\mu(A,B) = \text{meas}[(A \setminus B) \cup (B \setminus A)], \tag{2.3}$$

defined by the Lebesgue measure of the symmetric set difference between A and B; see Hewitt and Stromberg [10, p. 144]. It should be noted that μ coincides with the well-known Ekeland metric in $L^\infty(D)$ applied to characteristic functions:

$$d_E(\chi_A, \chi_B) = \text{meas}\{x \in D \mid \chi_A(x) \neq \chi_B(x)\} = \mu(A,B). \tag{2.4}$$

Relations (2.3), (2.4) are defined up to sets of measure zero. Without supplementary regularity assumptions on the boundaries of the sets, there is no connection between δ and μ. For instance, let $\overline{S(0,1)}$ be the closed unit ball in $\mathbf{R}^m$. Add n (closed) rays of length 2, starting from the origin, into the ball such that the union of the rays is dense in $\overline{S(0,2)}$ as $n \to \infty$, and denote the resulting (closed) sets by A_n. Then

$$\text{Hlim}(A_n) = \overline{S(0,2)} \quad \text{for } n \to \infty, \tag{2.5}$$

$$\mu(A_n, \overline{S(0,1)}) \to 0 \quad \text{for } n \to \infty. \tag{2.6}$$

In Chenais [6], it was proved that, for uniformly Lipschitz domains, convergence in the metric (2.2) yields convergence in the metric (2.4) with the same limit (up to a set with zero measure).

Let us now introduce the mappings $d_\Omega : \bar{D} \to \mathbf{R}$, based on the Euclidean distance functions associated with the domain Ω and its complementary:

$$d_\Omega(x) = \begin{cases} \text{dist}(x, \bar{D} \setminus \Omega) & \text{if } x \in \Omega, \\ 0 & \text{if } x \in \partial\Omega, \\ -\text{dist}(x, \overline{\Omega}) & \text{if } x \in \bar{D} \setminus \overline{\Omega}. \end{cases} \tag{2.7}$$

The mapping d_Ω is uniformly Lipschitzian in $\bar{D}$ for any open subset $\Omega \subset \bar{D}$; see Clarke [7]. Let $\Omega_n \subset D$ be a sequence of open sets, not necessarily connected. Let $d_n = d_{\Omega_n}$ be the associated mappings via (2.7). By the Ascoli theorem, on a subsequence again denoted by n, we have $d_n \to \hat{d}$ uniformly in $\bar{D}$. However, $\hat{d}$ is not necessarily a function of the same type since, in general, the Hlim's of $\overline{\Omega}_n$ and of $\bar{D} \setminus \Omega_n$ may be not complementary to each other (see the above example with the sets A_n). Let $\widehat{\Omega} = \{x \in D \mid \hat{d}(x) > 0\}$ (possibly void).

PROPOSITION 2.1.

$$\widehat{\Omega} = D \setminus \mathrm{Hlim}(\bar{D} \setminus \Omega_n).$$

Proof. Let $x \in D \setminus \mathrm{Hlim}(\bar{D} \setminus \Omega_n)$ so that $x \notin \mathrm{Hlim}(\bar{D} \setminus \Omega_n)$. Then $\lim_{n\to\infty} \mathrm{dist}(x, \bar{D} \setminus \Omega_n) > 0$. Thus $\lim_{n\to\infty} d_n(x) > 0$; i.e., $x \in \widehat{\Omega}$.

Conversely, assume that $x \in \widehat{\Omega}$ and $x \notin D \setminus \mathrm{Hlim}(\bar{D} \setminus \Omega_n)$. Then $\hat{d}(x) > 0$, and $x \in \mathrm{Hlim}(\bar{D} \setminus \Omega_n)$. That is, $\hat{d}(x) > 0$, and there are $x_n \in \bar{D} \setminus \Omega_n$ such that $x_n \to x$. This means that $\hat{d}(x) > 0$, $d_n(x_n) \le 0$, and $x_n \to x$. By the uniform convergence, we have $\hat{d}(x) > 0$ and $\hat{d}(x) \le 0$, which lead to a contradiction. It follows that $\widehat{\Omega} = D \setminus \mathrm{Hlim}(\bar{D} \setminus \Omega_n)$, which is the desired conclusion. □

Remark. The above proposition shows that the well-known compactness property of the Hausdorff–Pompeiu distance is a direct consequence of the Ascoli compactness criteria. A variant of the mapping d_Ω (identically zero outside Ω) was considered by Sverak [23], who also proved a result similar to Proposition 2.1.

PROPOSITION 2.2. *If* $\mathrm{Hlim}(\bar{D} \setminus \Omega_n) = \bar{D} \setminus \widehat{\Omega}$, *then, for any compact* $\mathcal{K} \subset \widehat{\Omega}$, *there is an* $n_{\mathcal{K}} = n(\mathcal{K}) \in N$ *such that* $\mathcal{K} \subset \Omega_n$ *for* $n \ge n_{\mathcal{K}}$.

Proof. We use the same notation as in Proposition 2.1. Since $\hat{d}$ is continuous on $\bar{D}$ and strictly positive on $\mathcal{K}$, there is a $c_{\mathcal{K}} > 0$ such that

$$\hat{d}(x) \ge c_{\mathcal{K}} > 0 \quad \forall x \in \mathcal{K}.$$

By the uniform convergence, for $n \ge n_{\mathcal{K}}$, we obtain $d_n(x) \ge \frac{1}{2} c_{\mathcal{K}} > 0$ for all $x \in \mathcal{K}$. That is, $\mathcal{K} \subset \Omega_n$ for $n \ge n_{\mathcal{K}}$, as required. □

Remark. This property is called the Γ-*property* by Liu [13] and Liu and Rubio [15], and it plays an essential role in the local convergence theory for the solutions of PDEs defined in sequences of bounded domains. The same property is also proved in Pironneau [20], by different methods, together with other domain convergence results.

DEFINITION 2.3. *We say that the sequence of open sets* $\Omega_n \subset D$ *is parametrically convergent to the open set* $\widetilde{\Omega} \subset D$ *if there is a sequence of continuous mappings* $p_n : \bar{D} \to \mathbf{R}$ *such that* $p_n \to \tilde{p}$ *uniformly in* $\bar{D}$ *and*

$$\begin{aligned} \Omega_n &= \{x \in D \mid p_n(x) > 0\}, \\ D \setminus \overline{\Omega}_n &= \{x \in D \mid p_n(x) < 0\}, \\ \widetilde{\Omega} &= \{x \in D \mid \tilde{p}(x) > 0\}, \\ D \setminus \overline{\widetilde{\Omega}} &= \{x \in D \mid \tilde{p}(x) < 0\}. \end{aligned}$$

We denote the limit by $\widetilde{\Omega} = p - \lim \Omega_n$.

Remark. The "parametrization" p_n associated with the domain Ω_n is not unique, and the distance mapping d_n is just one example. The p-limit and the convergence properties depend on the parametrization. If it is different from the function d_Ω, then the convergence may differ from the Hausdorff–Pompeiu convergence. For instance, we choose $\check{p} : \mathbf{R} \to \mathbf{R}$ by

$$\check{p}(x) = \begin{cases} -(x-1)^2 + \frac{1}{2}, & x \ge \frac{1}{2}, \\ x^2, & |x| \le \frac{1}{2}, \\ -(x+1)^2 + \frac{1}{2}, & x \le -\frac{1}{2}, \end{cases}$$

and we rotate its graph to define a continuous mapping $p : \mathbf{R}^2 \to \mathbf{R}$. Take $p_n : \mathbf{R}^2 \to \mathbf{R}$, $p_n(x) = p(x) + \frac{1}{n}$. Then the corresponding domains are $\Omega_n = \{x \in \mathbf{R}^2 \mid |x|_{\mathbf{R}^2} < 1 + \sqrt{\frac{n+2}{2n}}\}$ and $\Omega = \{x \in \mathbf{R}^2 \mid 0 < |x|_{\mathbf{R}^2} < 1 + \frac{1}{\sqrt{2}}\}$. Notice that Ω is nonsmooth, $\Omega = p - \lim \Omega_n$, and $\bar{\Omega} \neq \operatorname{Hlim} \bar{\Omega}_n$. If $\check{p}$ is zero around $x = 0$ on some interval, then $p - \lim \Omega_n$ will be a circular crown, etc. By taking $\sup(p_n, p_k)$ or $\inf(p_n, p_k)$, one can easily "parametrize" $\Omega_n \cup \Omega_k$ or $\Omega_n \cap \Omega_k$.

PROPOSITION 2.4. *The parametric convergence has the* Γ*-property for any parametrization.*

Proof. This is similar to the proof of Proposition 2.2. □

Remark. It is possible to weaken the conditions in Definition 2.3 by replacing the uniform convergence with other types of functional convergence for the mapping p_n. This will be used in section 4 (see Theorem 4.1 and its subsequent remarks).

PROPOSITION 2.5. *If* $\Omega = p - \lim \Omega_n$ *and the closed set* $C = \{x \in \bar{D} \mid \tilde{p}(x) = 0\}$ *has zero measure, then* $\chi_{\Omega_n} \to \chi_\Omega$ *a.e. in* D.

Proof. If $x \in \Omega$, then $\tilde{p}(x) > 0$ so that $p_n(x) > 0$ for $n \geq n_x$ (depending on x). Thus $\chi_{\Omega_n}(x) = \chi_\Omega(x) = 1$ for $n \geq n_x$. If $x \in D \setminus \overline{\Omega}$, then $\tilde{p}(x) < 0$ and $p_n(x) < 0$ for $n \geq n_x$; i.e., $x \in D \setminus \overline{\Omega}_n$ for $n \geq n_x$. Consequently, $\chi_{\Omega_n}(x) = \chi_\Omega(x) = 0$ for $n \geq n_x$.

As the set C has zero measure, we get that $\chi_{\Omega_n}(x) \to \chi_\Omega(x)$ a.e. in D. □

Remark. The family of distance-type mappings used by Sverak [23] does not satisfy this property.

DEFINITION 2.6. *Assume that* $\Omega = p - \lim \Omega_n$, *and let* $y_n \in H^1(\Omega_n)$ *be such that* $\{|y_n|_{H^1(\Omega_n)}\}$ *is bounded. We say that* $\{y_n\}$ *is locally convergent to* $y \in H^1(\Omega)$, *and we write* $y = L - \lim y_n$ *if, for any* $G \subset\subset \Omega$ *(open set compactly embedded in* Ω*), we have*

$$y_n|_G \longrightarrow y|_G \quad \textit{weakly in } H^1(G). \tag{2.8}$$

Remark. This definition is motivated by Proposition 2.4. The limit mapping y is uniquely determined. The convergence in (2.8) is also valid in $L^2(G)$ strongly for any $G \subset\subset \Omega$.

THEOREM 2.7 (compactness). *Assume* $\Omega = p - \lim \Omega_n$. *Suppose that* $y_n \in H^1(\Omega_n)$ *and* $|y_n|_{H^1(\Omega_n)}$ *is uniformly bounded. Then there are a* $y \in H^1(\Omega)$ *and a subsequence still denoted by* y_n *such that* $y = L - \lim y_n$.

Proof. Take a sequence $G_j \subset\subset \Omega$ such that $G_j \subset G_{j+1}$ and $\bigcup G_j = \Omega$. For each j, we take subsequences (one after another and all denoted by n) such that $y_n|_{G_j} \longrightarrow y^j$ weakly in $H^1(G_j)$. We define y on Ω by $y(x) = y^j(x)$ a.e. $x \in G_j$, which is possible by the properties of $\{G_j\}_{j \in \mathcal{N}}$. Clearly, $y \in L^2(\Omega)$ since $|y_n|_{L^2(G_j)}$ is uniformly bounded with respect to n and j. Consider any $\varphi \in \mathcal{D}(\Omega)$. There is a j_0 such that $\varphi \in \mathcal{D}(G_j)$ for all $j \geq j_0$. Therefore,

$$\int_{G_j} \nabla y \varphi = \int_\Omega \nabla y \varphi = -\int_\Omega y \nabla \varphi = -\int_{G_j} y^j \nabla \varphi = \int_{G_j} \nabla y^j \varphi.$$

This yields that $\nabla y = \nabla y^j$ in G_j for all $j \geq j_0$. As $|y^j|_{H^1(G_j)}$ is bounded with respect to j, we obtain that $\nabla y \in L^2(\Omega)^m$; i.e., $y \in H^1(\Omega)$. Relation (2.8) then follows, and the proof is completed. □

THEOREM 2.8 (lower semicontinuity). *If* $l : \mathbf{R}^m \times \mathbf{R} \times \mathbf{R}^m \to \mathbf{R}$ *is nonnegative and measurable,* $l(x, \cdot, \cdot)$ *is continuous on* $\mathbf{R} \times \mathbf{R}^m$, $l(x, s, \cdot)$ *is convex on* $\mathbf{R}^m$, *and*

$\Omega = p - \lim \Omega_n$, *then*

$$\int_\Omega l(x, y, \nabla y)\, dx \le \liminf_{n\to\infty} \int_{\Omega_n} l(x, y_n, \nabla y_n)\, dx \tag{2.9}$$

provided that $y = L - \lim y_n$.

Proof. Let $\{G_j\}$ be selected as in the previous proof. Then we have $\chi_{G_j} \to \chi_\Omega$ a.e. in D. For any fixed G_j, we have that $y_n \to y$ weakly in $H^1(G_j)$, and we obtain

$$\int_{G_j} l(x, y, \nabla y)\, dx \le \liminf_{n\to\infty} \int_{G_j} l(x, y_n, \nabla y_n)\, dx$$

since weak lower semicontinuity is a well-known property of the convex integrals in the fixed domains. Next, Fatou's lemma gives

$$\begin{aligned}
\int_\Omega l(x, y, \nabla y)\, dx &= \int_\Omega \lim_{j\to\infty} \chi_{G_j} l(x, y, \nabla y)\, dx \\
&\le \liminf_{j\to\infty} \int_{G_j} l(x, y, \nabla y)\, dx \\
&\le \liminf_{j\to\infty} \liminf_{n\to\infty} \int_{\Omega_n} l(x, y_n, \nabla y_n)\, dx \\
&= \liminf_{n\to\infty} \int_{\Omega_n} l(x, y_n, \nabla y_n)\, dx.
\end{aligned}$$

The positivity of l is essential in the above proof. □

Remark. Theorems 2.7 and 2.8 are variants or results previously proved by Liu and Rubio [15] and Liu [13]. It should be noted that it is enough to assume the Γ-property for the open sets Ω_n and Ω to prove Theorems 2.7 and 2.8.

3. Equicontinuity. We consider the model problem (SONB). The problem is formulated as

$$\min_\Omega \int_\Omega (y - y_d)^2\, dx, \tag{3.1}$$

subject to the following variational equation with the homogeneous Neumann boundary condition:

$$\int_\Omega \nabla y \nabla v + \int_\Omega yv = \int_\Omega fv \quad \forall v \in H^1(\Omega), \tag{3.2}$$

where Ω is a variable open set such that $\Omega \subset D$ with D being a fixed bounded open set in $\mathbf{R}^m$, and $y_d \in L^2(D)$. For the admissible class of open sets denoted by $\mathcal{O}$, we require that they have the $\mathcal{C}$-property (or, equivalently, the segment property) with some uniform constants:

(H1) We consider a family $\mathcal{F}$ of equibounded and equiuniformly continuous functions $g : S(0,k) \to \mathbf{R}$, with $k > 0$ fixed and $S(0,k) \subset \mathbf{R}^{m-1}$ an open ball. For any $\Omega \in \mathcal{O}$, there is a subset $\mathcal{F}_\Omega \subset \mathcal{F}$, and, for any $g \in \mathcal{F}_\Omega$, we associate an orthogonal system of axes of center $o_g \in \partial\Omega$, "vertical" vector $l_g \in \mathbf{R}^m$ of unit length, and a rotation R_g in $\mathbf{R}^m$ such that $l_g = R_g(0, 0, \dots, 0, 1)$ and

$$\bigcup_{g\in\mathcal{F}_\Omega} \{R_g(s, 0) + o_g + g(s) l_g \mid s \in S(0,k)\} = \partial\Omega.$$

(H2) There is an $a > 0$ such that, for any $\Omega \in \mathcal{O}$ and any $g \in \mathcal{F}_\Omega$, the uniform segment property is valid:

$$\begin{aligned} &R_g(s,0) + o_g + (g(s)+t)l_g \in \mathbf{R}^m \setminus \overline{\Omega} \quad \forall s \in S(0,k),\ \forall t \in]0,a[\,, \\ &R_g(s,0) + o_g + (g(s)-t)l_g \in \Omega \qquad\quad \forall s \in S(0,k),\ \forall t \in]0,a[\,. \end{aligned}$$

These two conditions represent the usual definition of boundaries of class $\mathcal{C}$ with added uniformity assumptions. Notice that, due to the compactness of $\partial\Omega$, it can be covered by a finite number of local charts; therefore, both conditions are automatically satisfied, and the only real requirement is the uniformity with respect to the whole family $\mathcal{O}$, which does not allow the local charts to shrink.

Our specific requirement is that there is a constant $r \in]0,k[$ such that

$$\text{(H3)} \qquad \bigcup_{g\in\mathcal{F}_\Omega} \{R_g(s,0) + o_g + g(s)l_g \mid s \in \overline{S(0,r)}\} = \partial\Omega$$

for any $\Omega \in \mathcal{O}$. This is a uniform restriction (or extension) property for the whole family $\mathcal{F}$ of the local charts. It avoids, for instance, clustering of singularities (like cusps) near the boundary of any local chart. Notice that, due to the finite numbers of the local charts, a positive number $0 < r_\Omega < k$ can be found in each Ω with the above property. We just assume that it can be chosen independent of Ω.

Example. Take $g(x) = x^{\frac{1}{2}}$, $x \in]-\alpha,\alpha[$, to be a local chart with Hölder regularity for some domain in $\mathbf{R}^2$. In $x = 0$, we have a cusp, and the segment property is fulfilled only by the vertical segments. Thus the segment choice for local charts with cusps is unique, and only cusps with the same "axis" may belong to the same local chart. Hypothesis (H3) requires, in particular, that cusps with different "axes" do not cluster. In the common part (which cannot shrink) of neighboring local charts, no cusp can occur.

Remark. In the counterexample of Pironneau [20], with infinitely many oscillations of the boundary in a rectangular region, all of the conditions in (H1)–(H3) are fulfilled except the equicontinuity of the local charts. This shows the essential importance of this assumption, reflected by the title of the section. Examples of continuous oscillating boundaries which are not even of class $\mathcal{C}$ may be found in the book by Maz'ja [17]. It is also known that domains with cuts are not of class $\mathcal{C}$. Notice, however, that our assumptions allow infinitely many oscillations with vanishing amplitude (to preserve equicontinuity).

THEOREM 3.1 (compactness and existence). *Let $\{\Omega_n\}$ be a minimizing sequence of open sets for the problem* (3.1) *satisfying the assumptions* (H1)–(H3). *Then there is an open set $\widehat{\Omega}$ of class $\mathcal{C}$ which is a solution of the problem* (3.1) *and which satisfies* (H1)–(H3). *Furthermore, $\chi_{\Omega_n} \to \chi_{\hat{\Omega}}$ a.e.*

Proof. We may assume that $\bar{D}$ is large enough to include Ω_n and the segments defined in (H2). Denote by d_n the distance-type functions introduced in (2.7) corresponding to Ω_n and by $\hat{d}$ their uniform limit, $\hat{d} \in C(\bar{D})$. Let $\Lambda = \{x \in \bar{D} \mid \hat{d}(x) \geq 0\}$ be a closed set, which is clearly nonvoid. Take $\hat{x} \in \Lambda$ such that $\hat{d}(x) = 0$. Then $d_n(\hat{x}) \to 0$, by the definition of $\hat{d}$. By the definition of d_n, there are $x_n \in \partial\Omega_n$, $x_n \to \hat{x}$ (and $d_n(x_n) = 0$). By (H3), there are $g_n \in \mathcal{F}_{\Omega_n}$ such that $x_n = R_{g_n}(s_n,0) + o_{g_n} + g_n(s_n)l_{g_n}$, $d_n(o_{g_n}) = 0$, and $s_n \in \overline{S(0,r)}$. Under our conditions, we may assume that $s_n \to \hat{s} \in \overline{S(0,r)}$ and $g_n \to \hat{g}$ uniformly in $\overline{S(0,k)}$, with $\hat{g}$ being continuous in $\bar{D}$, $R_{g_n} \to \widehat{R}$, $o_{g_n} \to \hat{o}$ with $\hat{d}(\hat{o}) = 0$, $l_{g_n} \to \hat{l}$ as matrices or

vectors (since all are bounded) with $\widehat{R}(0,0,\dots,0,1)=\hat{l}$, and $|\hat{l}|=1$. We have

$$\hat{x}=\lim x_n=\lim(R_{g_n}(s_n,0)+o_{g_n}+g_n(s_n)l_{g_n})=\widehat{R}(\hat{s},0)+\hat{o}+\hat{g}(\hat{s})\hat{l}, \tag{3.3}$$

$$d_n(R_{g_n}(s,0)+o_{g_n}+g_n(s)l_{g_n})\to\hat{d}(\widehat{R}(s,0)+\hat{o}+\hat{g}(s)\hat{l})=0\quad\forall s\in S(0,k). \tag{3.4}$$

We show the segment property.

Take any $\varepsilon\in\,]0,a[$, and consider the point $\widehat{R}(s,0)+\hat{o}+(\hat{g}(s)-\varepsilon)\hat{l}\in\mathbf{R}^m$. We have that $R_{g_n}(s,0)+o_{g_n}+(g_n(s)-\varepsilon)l_{g_n}\to\widehat{R}(s,0)+\hat{o}+(\hat{g}(s)-\varepsilon)\hat{l}$; that is, $\widehat{R}(s,0)+\hat{o}+(\hat{g}(s)-\varepsilon)\hat{l}\in\bar{D}$ for $s\in S(0,k)$. As $R_{g_n}(s,0)+o_{g_n}+(g_n(s)-\varepsilon)l_{g_n}\in\Omega_n$ by (H2), we have $d_n(R_{g_n}(s,0)+o_{g_n}+(g_n(s)-\varepsilon)l_{g_n})>0$ for $s\in S(0,k)$, $\varepsilon\in\,]0,a[$, $n\geq 1$. It follows that $\hat{d}(\widehat{R}(s,0)+\hat{o}+(\hat{g}(s)-\varepsilon)\hat{l})\geq 0$ for $s\in S(0,k)$, $\varepsilon\in\,]0,a[$; i.e., $(\widehat{R}(s,0)+\hat{o}+(\hat{g}(s)-\varepsilon)\hat{l})\in\Lambda$ for such values of the parameters s, ε.

For the outside segment property, a sharper estimate is needed. By the equicontinuity of g_n, there is an $\delta>0$ (depending only on ε and independent of $s\in S(0,k)$ or $n\in N$) such that

$$|g_n(t)-g_n(s)|<\frac{\varepsilon}{2}\quad\forall n,\ \forall t\in S(s,\delta)\cap S(0,k). \tag{3.5}$$

Then, for $\varepsilon<\frac{2}{3}a$, we get

$$\begin{aligned}\operatorname{dist}[R_{g_n}(s,0)+o_{g_n}+(g_n(s)+\varepsilon)l_{g_n},\partial\Omega_n]\\ \geq\min\left\{\frac{\varepsilon}{2},\delta,a-\frac{3}{2}\varepsilon,\operatorname{dist}(s,\partial S(0,k)\right\}.\end{aligned} \tag{3.6}$$

Here we use the uniform outside segment property for Ω_n, i.e., $R_{g_n}(s,0)+o_{g_n}+(g_n(s)+\varepsilon)l_{g_n}\in D\setminus\overline{\Omega}_n$, for all $s\in S(0,k)$ and for all $\varepsilon\in\,]0,a[$. The inequality (3.6) comes from (3.5), which simply says that the cylinder $[S(0,k)\cap S(s,\delta)]\times[g_n(s)+\frac{\varepsilon}{2},g_n(s)+a-\frac{\varepsilon}{2}]$ after translation o_{g_n} and rotation R_{g_n} cannot intersect $\partial\Omega_n$ for any n. And the right-hand side in (3.6) estimates from below the distance between $(s,g_n(s)+\varepsilon)$ and the boundary of this cylinder. (This point is inside the cylinder for $\varepsilon<\frac{2}{3}a$.)

Then it yields

$$d_n(R_{g_n}(s,0)+o_{g_n}+(g_n(s)+\varepsilon)l_{g_n})\leq-\min\left\{\frac{\varepsilon}{2},\delta,a-\frac{3}{2}\varepsilon,\operatorname{dist}(s,\partial S(0,k))\right\}. \tag{3.7}$$

Inequality (3.7) is independent of n, and we can take the limit, by the uniform convergence, to obtain

$$\hat{d}(\widehat{R}(s,0)+\hat{o}+(\hat{g}(s)+\varepsilon)\hat{l})\leq-\min\left\{\frac{\varepsilon}{2},\delta,a-\frac{3}{2}\varepsilon,\operatorname{dist}(s,\partial S(0,k))\right\}; \tag{3.8}$$

that is, $\hat{d}(\widehat{R}(s,0)+\hat{o}+(\hat{g}(s)+\varepsilon)\hat{l})<0$ for all $s\in S(0,k)$ and for all $\varepsilon\in\,]0,\frac{2}{3}a[$, and, consequently, $(\widehat{R}(s,0)+\hat{o}+(\hat{g}(s)+\varepsilon)\hat{l})\notin\Lambda$ for these values of the parameters s, ε. By choosing a smaller $\delta>0$, if necessary, we can replace $\frac{\varepsilon}{2}$ by $\frac{\varepsilon}{l}$, $l\in N$, and $\frac{3}{2}\varepsilon$ by $\frac{l+1}{l}\varepsilon$ in inequalities (3.5)–(3.8). Finally, we have that $\widehat{R}(s,0)+\hat{o}+(\hat{g}(s)+\varepsilon)\hat{l}\notin\Lambda$ for $s\in S(0,k)$ and $\varepsilon\in\,]0,a[$.

Notice that estimates like (3.8) can also be obtained for $\hat{d}(\widehat{R}(s,0)+\hat{o}+(\hat{g}(s)-\varepsilon)\hat{l})$, $s\in S(0,k)$, $\varepsilon\in\,]0,a[$, with the reversed sign. Then

$$\widehat{\Omega}=\{x\in D\mid\hat{d}(x)>0\}=\operatorname{int}\Lambda \tag{3.9}$$

is a nonvoid open subset of D. The above argument shows that $\partial\widehat{\Omega} = \{x \in D \mid \hat{d}(x) = 0\}$ is of class $\mathcal{C}$ and satisfies (H1)–(H3) with the same constants a, r, k and the same modulus of continuity.

By Proposition 2.1, we see that $\widehat{\Omega} = D \setminus \mathrm{Hlim}(\bar{D} \setminus \Omega_n)$ (and $\widehat{\Omega} = p - \lim \Omega_n$), which proves the compactness of the family $\mathcal{O}$.

Let us also remark that, by Proposition 2.5, we have $\chi_{\Omega_n} \to \chi_{\hat{\Omega}}$ a.e. in D, and, by the Lebesgue theorem, this convergence is valid in any $L^q(D)$, $q \geq 1$. Here we also use the fact that $\partial\widehat{\Omega} = \{x \in D \mid \hat{d}(x) = 0\}$ has zero measure in $\mathbf{R}^m$ since it can be represented as a finite union of graphs of continuous functions.

The fact that $\widehat{\Omega}$ is a solution of the problem (3.1) follows from the subsequent lemma and Theorem 2.8. □

LEMMA 3.2. *Let y_n, $\hat{y}$ denote the unique solutions of* (3.2) *associated with Ω_n, $\widehat{\Omega}$. Then $\hat{y} = L - \lim y_n$ on a subsequence.*

Proof. Clearly, $y_n \in H^1(\Omega_n)$, $\hat{y} \in H^1(\widehat{\Omega})$, and $\{|y_n|_{H^1(\Omega_n)}\}$ is bounded. By Proposition 2.4, for any open set $G \subset\subset \widehat{\Omega}$, there are n_G such that $G \subset \Omega_n$, $n \geq n_G$. We have

$$\int_G (\nabla y_n \nabla v + y_n v) - \int_D \chi_{\Omega_n} f v = \int_{\Omega_n \setminus G} (\nabla y_n \nabla v + y_n v) \quad \forall v \in C^1(\bar{D}). \tag{3.10}$$

We can estimate

$$\left| \int_{\Omega_n \setminus G} (\nabla y_n \nabla v + y_n v) \right| \leq |v|_{C^1(\bar{D})} |y_n|_{H^1(\Omega_n)} \mu(\Omega_n - G)^{\frac{1}{2}}. \tag{3.11}$$

Taking the limit $n \to \infty$ in (3.11), we have that (3.10) yields

$$\left| \int_G \nabla\tilde{y} \nabla v + \int_G \tilde{y} v - \int_{\widehat{\Omega}} f v \right| \leq M |v|_{C^1(\bar{D})} \mu(\widehat{\Omega} - G)^{\frac{1}{2}} \tag{3.12}$$

(due to the convergence of the characteristic functions of Ω_n), where $\tilde{y}$ denotes the L-limit of y_n given by Theorem 2.7. We can take an increasing sequence of open sets $G_j \subset\subset \widehat{\Omega}$ such that $\cup G_j = \widehat{\Omega}$ and (3.12) gives

$$\int_{\widehat{\Omega}} \nabla\tilde{y} \nabla v + \int_{\widehat{\Omega}} \tilde{y} v = \int_{\widehat{\Omega}} f v \quad \forall v \in C^1(\bar{D}). \tag{3.13}$$

Since $\widehat{\Omega}$ has the segment property, $C^1(\bar{D})$ is dense in $H^1(\widehat{\Omega})$ (see Adams [1]) and (3.13) shows that $\tilde{y} = \hat{y}$. This ends the proof.

The semicontinuity result from Theorem 2.8 ensures that $\widehat{\Omega}$ is the desired minimizer for (3.1). □

Remark. The above proof does not use the uniform extension property for functions in $H^1(\Omega)$. We replace it by a density property, which is, in fact, an approximate extension result. This is one of the reasons that we can renounce the cone property for $\partial\Omega$ and use the segment property instead. A more general cost functional, as in (2.9), may be considered in the problem (3.1).

Remark. If we impose uniform Hölder conditions for the family $\mathcal{O}$, the limit domain will satisfy a similar Hölder property. In this case, trace theorems are known (see, e.g., Ladyzenskaya and Uraltseva [12], Pironneau [20]), and the result of Theorem 3.1 can be then extended to Dirichlet boundary value problems.

Example. A special simple case of the family $\mathcal{O}$ is obtained when global representations are given by

$$\mathcal{O}_1 = \{\Omega_g \mid g \in \mathcal{F}_1\},$$
$$\Omega_g = \{(s,\lambda) \in D \mid \lambda < g(s)\},$$

where $D = U \times]0, b[$, $U \subset \mathbf{R}^{m-1}$ open domain, and $\mathcal{F}_1 = \{g : U \to R_+ \mid 0 < c \le g \le b \text{ in } U\}$.

By Theorem 3.1, one can immediately prove the following corollary.

COROLLARY 3.3. *Assume that the family $\mathcal{F}_1$ is equicontinuous. Then the associated problem* (3.1) *has at least one solution in $\mathcal{O}_1$.*

Proof. We indicate a direct argument. The family $\mathcal{F}_1$ is equibounded by definition. We take a minimizing sequence of domains Ω_n associated with g_n. We may assume that $g_n \to \hat{g}$ uniformly in $\overline{U}$ and, clearly, $\hat{g} \in \mathcal{F}_1$. Here it is possible to define the continuous mappings on $\bar{D}$, $p_n(s,\lambda) = g_n(s) - \lambda$, $\hat{p}(s,\lambda) = \hat{g}(s) - \lambda$, and $p_n \to \hat{p}$ uniformly; $\Omega_n = \{(s,\lambda) \in D \mid p_n(s,\lambda) > 0)$, $\widehat{\Omega} = \{(s,\lambda) \in D \mid \hat{p}(s,\lambda) > 0\}$. Finally, $\widehat{\Omega} = p - \lim \Omega_n$, and the results of section 2 may be used directly to end the proof without using the distance functions. ∎

Remark. Another family of domains of interest may be obtained by considering a domain $D \subset \mathbf{R}^m$ such that $D \supset B(0,1)$ (the unit ball) and then defining

$$\mathcal{O}_2 = \{\Omega \subset D \mid \Omega = h(B(0,1))\},$$

where $h : D \to D$ is any homeomorphism, i.e., h and h^{-1} are continuous. This would be a generalization of the mapping method of Murat and Simon [18], where the mappings h were assumed diffeomorphisms.

If $\Omega_h = h_n(B)$ and $h_n \to h$, $h_n^{-1} \to h^{-1}$ uniformly in D, and $\Omega = h(B)$, then it is easy to see that these domains have the Γ-property, and simple representation formulae are valid for $\Omega = hh_n^{-1}(\Omega_n)$, $\partial\Omega = hh_n^{-1}(\partial\Omega_n)$. However, examples from Maz'ja [17] show that, conceptually, $\partial\Omega$ may not be of class $\mathcal{C}$ even when $\partial\Omega_n$ satisfies this assumption. Also, the compactness of $\mathcal{O}_2$ is not clear.

Remark. The examples of the families $\mathcal{O}_1$ and $\mathcal{O}_2$ also indicate that the concept of parametric convergence has enough flexibility to take into account various representation methods of open sets.

4. Measurability. We consider the material distribution problem (SOTB). In this case, it makes sense and is of interest to consider the case where the sets occupied by each material are merely measurable.

We fix $E \subset D$ to be two given bounded domains in $\mathbf{R}^m$ and $\Omega \subset D$ to be a variable measurable subset in some prescribed class $\mathcal{O}$, occupied by one material, while $D \setminus \Omega$ is occupied by another material. Then the physical properties of the two regions are different, and this is expressed by the fact that different coefficients appear in the elliptic equations describing the problem, which (formally) read as

$$-a_1 \Delta y_1 + b_1 y_1 = f \quad \text{in } \Omega, \tag{4.1}$$

$$-a_2 \Delta y_2 + b_2 y_2 = f \quad \text{in } D \setminus \overline{\Omega}, \tag{4.2}$$

$$a_1 \frac{\partial y_1}{\partial n} = a_2 \frac{\partial y_2}{\partial n}, \quad y_1 = y_2 \quad \text{in } \partial\Omega \setminus (\partial\Omega \cap \partial D), \tag{4.3}$$

$$a_i \frac{\partial y_i}{\partial n} = 0 \quad \text{in } \Gamma_1, \qquad y_i = 0 \quad \text{in } \Gamma_2, \quad i = 1, 2, \tag{4.4}$$

where $\Gamma_1 \cup \Gamma_2 = \partial D$ is assumed Lipschitz, $\Gamma_1 \cap \Gamma_2 = \emptyset$, $a_i, b_i > 0$, $i = 1, 2$, are some constants, and $f \in L^2(D)$.

Let y_Ω be defined by

$$y_\Omega(x) = \begin{cases} y_1(x) & \text{in } \Omega, \\ y_2(x) & \text{in } D \setminus \Omega. \end{cases} \tag{4.5}$$

Then the weak formulation of the transmission problem (4.1)–(4.4) reads

$$\int_D \{[a_1\chi_\Omega + a_2(1-\chi_\Omega)]\nabla y_\Omega \cdot \nabla w + [b_1\chi_\Omega + b_2(1-\chi_\Omega)]y_\Omega w\}\, dx$$

$$= \int_D fw\, dx \quad \forall w \in V, \tag{4.6}$$

$$V = \{w \in H^1(D) \mid w = 0 \quad \text{in } \Gamma_2\}. \tag{4.7}$$

For any measurable subset $\Omega \subset D$, the bilinear form governing (4.6) is bounded and coercive in V, and there exists a unique solution $y_\Omega \in V$, which formally satisfies relations (4.1)–(4.5).

In Pironneau [20], by directly interpreting the characteristic function χ_Ω as a control parameter, the following optimization problem is discussed:

$$\min_{\Omega \in \mathcal{O}} \int_E |y_\Omega - z_d|^2\, dx, \tag{4.8}$$

subject to (4.6) and with some prescribed $z_d \in L^2(E)$. However, it is very difficult to impose the constraint that the control should take only the values 0 and 1 in the whole D.

Our approach is to specify the set $\mathcal{O}$ of admissible Ω by requiring that χ_Ω is of the form $H(p_\Omega)$, where $p_\Omega \in U_{ad}$ (some admissible set of mappings will be defined later), and $H \subset \mathbf{R} \times \mathbf{R}$ is the maximal monotone extension of the Heaviside function:

$$H(p) = \begin{cases} 1, & p > 0, \\ [0,1], & p = 0, \\ 0, & p < 0. \end{cases} \tag{4.9}$$

Notice that by taking, for instance, $p_\Omega = \chi_\Omega$, we have representations via H for any measurable Ω. If $\text{meas}(\partial\Omega) = 0$, then we may take $p_\Omega = d_\Omega$, which is even Lipschitzian in D.

For the optimization problem (4.8), we define the class of admissible sets Ω by $\chi_\Omega = H(p_\Omega)$, where $p_\Omega \in U_{ad} \subset H^1_{\text{loc}}(D)$, and $p \in U_{ad}$ iff

$$|p|_{H^{1+\theta_\mathcal{K}}(\mathcal{K})} \le M_\mathcal{K} \qquad \forall \mathcal{K} \subset\subset D,\ \theta_\mathcal{K} > 0, \tag{4.10}$$

$$|p(x)| + |\nabla p(x)|_{\mathbf{R}^m} \ge \nu > 0 \quad \text{a.e. in } D. \tag{4.11}$$

If $\text{meas}(\partial\Omega) = 0$, the mappings d_Ω satisfy (4.11) (see Clarke [7, p. 66]) but do not satisfy (4.10) in general.

Conversely, the condition (4.11) ensures that the set

$$\{x \in D \mid p(x) = 0\} \tag{4.12}$$

has zero measure for $p \in H^1_{\text{loc}}(D)$; see Brezis [2, p. 195].

Remark. Under some stronger smoothness assumptions, condition (4.11) ensures the application of the implicit function theorem and a characterization of the "boundary" given by $\{x \in D \mid p(x) = 0\}$. However, our regularity hypotheses are so weak that even the implicit theorem in Clarke [7, p. 255] for Lipschitzian mappings cannot be applied. The only properties that we have are (4.12) and the measurability of Ω's, which are, in fact, defined up to sets of zero measure. Notice, as well, that the local character of (4.10) allows oscillations of $\partial\Omega$, even under the smoothness assumptions.

We reformulate the problem (4.6)–(4.8) as follows:

$$\min_{p\in U_{ad}} \int_E |y_p - z_d|^2\,dx, \tag{4.13}$$

$$\int_D \{[a_1H(p) + a_2(1 - H(p))]\nabla y_p \nabla w + [b_1H(p) + b_2(1 - H(p))]y_p w\}\,dx = \int_D fw\,dx \quad \forall w \in V. \tag{4.14}$$

THEOREM 4.1. *Under the hypotheses* (4.10)–(4.11)*, the problem* (4.13)–(4.14) *has at least one optimal pair* $[\bar{y}, \bar{p}] \in V \times U_{ad}$.

Proof. Obviously, $U_{ad} \neq \emptyset$ (as constant functions are in U_{ad}), and we may consider a minimizing sequence $[y_n, p_n] \subset V \times U_{ad}$. By taking an increasing sequence of open sets $\mathcal{K}_l \subset D$ such that $\cup\mathcal{K}_l = D$, and by the compactness of $\{p_n|_{\mathcal{K}_l}\}$ in $H^1(\mathcal{K}_l)$, we may define $p \in U_{ad}$ such that $p_n \to p$ strongly in $H^1(\mathcal{K}_l)$ for all $l \in N$, $p_n \to p$, and $\nabla p_n \to \nabla p$ a.e. in D.

Since $a_1H(p_n) + a_2(1 - H(p_n)) \geq ct > 0$ and $b_1H(p_n) + b_2(1 - H(p_n)) \geq ct > 0$, it is easy to infer that $\{y_n\}$ is bounded in $V \subset H^1(D)$, and we may assume that $y_n \to y$ weakly in $H^1(D)$ on a subsequence.

Obviously, $H(p_n)$ is bounded in $L^\infty(D)$; therefore, we may also assume that $H(p_n) \to H(p)$ weakly star in $L^\infty(D)$. The identification of the limit is a consequence of the demiclosedness of the maximal monotone operator H, applied in $L^2(\mathcal{K}) \times L^2(\mathcal{K})$ for any $\mathcal{K} \subset\subset D$. Notice that, by $p \in U_{ad}$ and by (4.12), it follows that $H(p)$ is a characteristic function in D.

We also have that $H(p_n) \to H(p)$ a.e. in D. We know that $p_n(x) \to p(x) \neq 0$ a.e. in D. If $p(x) > 0$, then $p_n(x) > 0$ for $n \geq n_x$ and $H(p_n(x)) = H(p(x)) = 1$ for $n \geq n_x$. If $p(x) < 0$, similarly, we obtain $H(p_n(x)) = H(p(x)) = 0$ for $n \geq n_x$. Consequently, we get that $H(p_n) \to H(p)$ strongly in $L^s(D)$ for all $s \geq 1$ by the Lebesgue dominated convergence theorem.

As $[y_n, p_n]$ satisfies (4.14), the above convergence allows to take the limit and to see that $[y, p]$ also satisfies (4.14) and $p \in U_{ad}$; i.e., the pair $[y, p]$ is admissible. Moreover, for the minimizing sequence, we have

$$\lim_{n\to\infty} \int_E |y_n - z_d|^2\,dx = \int_E |y - z_d|^2\,dx,$$

which shows that the pair $[y, p]$ is optimal for the problem (4.13), (4.14), and we redenote it by $[\bar{y}, \bar{p}]$. □

Remark. The above argument remains valid for any weakly lower semicontinuous cost functional on $H^1(D)$—for instance, for boundary cost functionals. Other boundary conditions may be imposed on ∂D as well.

It is possible to impose (4.10) only for $\mathcal{K} \subset\subset D \setminus C$, where $C \subset D$ is a given closed set of zero measure. This allows cracks in the corresponding $\partial\Omega$'s; see Bucur and Zolesio [3].

Remark. Considering the measurable sets $\Omega_n = \{x \in D \mid p_n(x) > 0\}$, the proof of Theorem 4.1 uses a property of parametric convergence for Ω_n based on a.e. convergence in D of p_n and ∇p_n, and a variant of Definition 2.3.

We continue by describing an approximation procedure which is suggested by our approach to characteristic functions. We denote by H_ε the Yosida approximation of H, given in this case by

$$H_\varepsilon(p) = \begin{cases} 1, & p > \varepsilon, \\ \frac{p}{\varepsilon}, & p \in [0, \varepsilon], \\ 0, & p < 0. \end{cases} \tag{4.15}$$

Notice that H_ε is Lipschitzian with Lipschitz constant $\frac{1}{\varepsilon}$. The approximation of the optimization problem (4.13), (4.14) is obtained by replacing (4.14) with

$$\int_D \{[a_1 H_\varepsilon(p) + a_2(1 - H_\varepsilon(p))]\nabla y_p^\varepsilon \nabla w + [b_1 H_\varepsilon(p) + b_2(1 - H_\varepsilon(p))] y_p^\varepsilon w\}\, dx = \int_D f w\, dx \quad \forall w \in V. \tag{4.14'}$$

By a variant of Theorem 4.1, we have existence of at least one optimal pair $[y_\varepsilon, p_\varepsilon] \in H^1(D) \times U_{ad}$ for the problem (4.13), (4.14').

THEOREM 4.2. *For any open set $\mathcal{K} \subset\subset D$, we have $[y_\varepsilon, p_\varepsilon] \to [\hat{y}, \hat{p}]$ in the weak-strong topology of $H^1(D) \times H^1(\mathcal{K})$ on a subsequence, and $[\hat{y}, \hat{p}]$ is an optimal pair for the problem* (4.13), (4.14).

Proof. By (4.10), we have $p_\varepsilon \to \hat{p}$ strongly in $H^1(\mathcal{K})$ on a subsequence, and $p_\varepsilon \to \hat{p}$, $\nabla p_\varepsilon \to \nabla \hat{p}$ a.e. in D for some $\hat{p} \in U_{ad}$. As $1 \geq H_\varepsilon(p) \geq 0$ a.e. in D, we also obtain that $\{y_\varepsilon\}$ is bounded in $H^1(D)$ by fixing $w = y_\varepsilon$ in (4.14'). We may assume that $y_\varepsilon \to \hat{y}$ weakly in $H^1(D)$ on a subsequence. Moreover, it is known from the theory of maximal monotone operators that $H_\varepsilon(p_\varepsilon) \to H(\hat{p})$ weakly in $L^2(\mathcal{K})$ for any $\mathcal{K} \subset\subset D$ since $\{H_\varepsilon(p_\varepsilon)\}$ is bounded in $L^\infty(D)$ and $p_\varepsilon \to \hat{p}$ strongly in $L^2(\mathcal{K})$, for instance. By the fact that $\hat{p} \in U_{ad}$ and by (4.12), (4.9), we know that $H(\hat{p})$ is a characteristic function in D. We can also prove the pointwise convergence of $H_\varepsilon(p_\varepsilon)$ a.e. in D. If $\hat{p}(x) > 0$, then $p_\varepsilon(x) > \frac{1}{2}\hat{p}(x)$ for $\varepsilon < \varepsilon_x$; that is, $p_\varepsilon(x) > \varepsilon$ for $\varepsilon < \varepsilon_x$ and $H_\varepsilon(p_\varepsilon(x)) = H(\hat{p}(x)) = 1$ by (4.15). If $\hat{p}(x) < 0$, then $p_\varepsilon(x) < 0$ for $\varepsilon < \varepsilon_x$ and $H_\varepsilon(p_\varepsilon(x)) = H(\hat{p}(x)) = 0$ for $\varepsilon < \varepsilon_x$. These two situations are valid a.e. in D by $\hat{p} \in U_{ad}$ and (4.12). Combining these with the Lebesgue theorem, we obtain that $H_\varepsilon(p_\varepsilon) \to H(\hat{p})$ strongly in $L^s(D)$ for all $s \geq 1$. Then we can take the limit in (4.14') and infer that the pair $[\hat{y}, \hat{p}]$ is admissible for the problem (4.13), (4.14). To show that it is optimal, we note that

$$\int_E |y_\varepsilon - z_d|^2\, dx \leq \int_E |y_p^\varepsilon - z_d|^2\, dx, \tag{4.16}$$

where y_p^ε denotes the solution of (4.14') corresponding to some $p \in U_{ad}$. By an argument of the same type as above, we can prove that $y_p^\varepsilon \to y_p$ weakly in $H^1(D)$ on a subsequence, where y_p is the solution of (4.14) associated with p. Taking the limit in (4.16) yields the optimality of $[\hat{y}, \hat{p}]$ in the problem (4.13)–(4.14) and completes the proof. □

COROLLARY 4.3. *On a subsequence, we have*

$$y_\varepsilon \to \hat{y} \quad \textit{strongly in } H^1(D), \tag{4.17}$$

where $\hat{y}$ is some optimal state for the problem (4.13)–(4.14).

Proof. There is a constant $c > 0$ such that

$$
\begin{aligned}
c\,|y_\varepsilon - \hat{y}|^2_{H^1(D)} &\le \int_D [a_1 H_\varepsilon(p_\varepsilon) + a_2(1 - H_\varepsilon(p_\varepsilon))]|\nabla(y_\varepsilon - \hat{y})|^2_{\mathbf{R}^m}\,dx \\
&\quad + \int_D [b_1 H_\varepsilon(p_\varepsilon) + b_2(1 - H_\varepsilon(p_\varepsilon))](y_\varepsilon - \hat{y})^2\,dx \\
&= \int_D \{[a_1 H_\varepsilon(p_\varepsilon) + a_2(1 - H_\varepsilon(p_\varepsilon))]\nabla y_\varepsilon \nabla(y_\varepsilon - \hat{y}) \\
&\qquad + [b_1 H_\varepsilon(p_\varepsilon) + b_2(1 - H_\varepsilon(p_\varepsilon))] y_\varepsilon (y_\varepsilon - \hat{y})\}\,dx \\
&\quad - \int_D \{[a_1 H_\varepsilon(p_\varepsilon) + a_2(1 - H_\varepsilon(p_\varepsilon))]\nabla \hat{y} \nabla(y_\varepsilon - \hat{y})) \\
&\qquad + [b_1 H_\varepsilon(p_\varepsilon) + b_2(1 - H_\varepsilon(p_\varepsilon))]\hat{y}(y_\varepsilon - \hat{y})\}\,dx \\
&= \int_D f(y_\varepsilon - \hat{y})\,dx - \int_D \{[a_1 H_\varepsilon(p_\varepsilon) \\
&\qquad + a_2(1 - H_\varepsilon(p_\varepsilon))]\nabla \hat{y} \nabla(y_\varepsilon - \hat{y})) \\
&\qquad + [b_1 H_\varepsilon(p_\varepsilon) + b_2(1 - H_\varepsilon(p_\varepsilon))]\hat{y}(y_\varepsilon - \hat{y})\}\,dx \\
&= I_1 + I_2.
\end{aligned} \tag{4.18}
$$

By Theorem 4.2, we may assume that $I_1 \to 0$ on a sequence as $\varepsilon \to 0$. For I_2, we first estimate the term

$$
\int_D [a_1 H_\varepsilon(p_\varepsilon) + a_2(1 - H_\varepsilon(p_\varepsilon))]\nabla \hat{y} \nabla(y_\varepsilon - \hat{y})\,dx, \tag{4.19}
$$

which is the most difficult. We know that $\nabla \hat{y} \nabla(y_\varepsilon - \hat{y})$ is weakly convergent in $L^1(D)$ and the coefficients $a_1 H_\varepsilon(p_\varepsilon) + a_2(1 - H_\varepsilon(p_\varepsilon))$ are bounded in $L^\infty(D)$ and strongly convergent in $L^s(D)$ for all $s \ge 1$. On a subsequence, we may assume that $[a_1 H_\varepsilon(p_\varepsilon) + a_2(1 - H_\varepsilon(p_\varepsilon))]\nabla(y_\varepsilon - \hat{y}) \to u$ weakly in $L^2(D)^m$. Egorov's theorem gives that $a_1 H_\varepsilon(p_\varepsilon) + a_2(1 - H_\varepsilon(p_\varepsilon)) \to a_1 H(\hat{p}) + a_2(1 - H(\hat{p}))$ uniformly in D_δ for any $\delta > 0$ and some measurable subset $D_\delta \subset D$ with $\mathrm{meas}(D - D_\delta) < \delta$. Combining this with $\nabla(y_\varepsilon - \hat{y}) \to 0$ weakly in $L^2(D)$, we have that $u = 0$ a.e. in D. Then the limit of the integral (4.19) is zero, and the same is true, by a similar reasoning, for I_2. Finally, (4.18) has limit zero, and this achieves the proof. □

Remark. It is possible to further smooth H_ε by means of a Friedrichs mollifier and to compute the gradient of the smoothed cost functional (4.13) with respect to $p \in U_{ad}$. This shows the constructive character of our approach presented in this paper. Numerical tests for the problem in this section were reported in Mäkinen, Neittaanmäki, and Tiba [16] together with an approximation result.

Remark. In Pironneau [20, p. 134], it is mentioned that, by taking $a_2 \to 0$, $b_2 \to 0$ in (4.14), the Neumann boundary value is approximated. Therefore, the results of this section may open a way to relax the continuity assumptions from section 3. A similar idea is possible for Dirichlet boundary value problems, which could be obtained by taking $a_2 \to \infty$, $b_2 \to \infty$. The hypotheses under which such passages to the limit can be performed are not clear yet.

REFERENCES

[1] R. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] H. Brezis, *Analyse fonctionelle. Théorie et applications*, Masson, Paris, 1983.
[3] D. Bucur and J. P. Zolesio, *Continuité par rapport an domaine dans le probleme de Neumann*, C. R. Acad. Sci. Paris Sér. I Math., 319 (1994), pp. 57–60.
[4] D. Bucur and J. P. Zolesio, *Optimization de forme sous contrainte capacitaire*, C. R. Acad. Sci. Paris Sér. I Math., 318 (1994), pp. 795–800.
[5] D. Bucur and J. P. Zolesio, *Free boundary problems and density perimeter*, J. Differential Equations, 126 (1996), pp. 224–243.
[6] D. Chenais, *On the existence of a solution in a domain identification problem*, J. Math. Anal. Appl., 52 (1975), pp. 189–219.
[7] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley and Sons, New York, 1983.
[8] J. Haslinger and P. Neittaanmäki, *Finite Element Approximation of Optimal Shape, Material and Topology Design*, John Wiley and Sons, Chichester, UK, 1996.
[9] A. Henrot, *Continuity with respect to the domain for the Laplacian: A survey*, Control Cybernet., 23 (1994), pp. 427–443.
[10] E. Hewitt and K. Stromberg, *Real and Abstract Analysis*, Springer-Verlag, Berlin, 1965.
[11] K. Kuratowski, *Introduction to Set Theory and Topology*, Pergamon Press, Oxford, UK, 1961.
[12] O. Ladyzenskaya and N. Uraltseva, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.
[13] W. B. Liu, *Optimal Shape Design for Variational Inequalities*, Ph.D. thesis, University of Leeds, Leeds, UK, 1991.
[14] W. B. Liu, P. Neittaanmäki, and D. Tiba, *Sur les problèmes d'optimisation structurelle*, C. R. Acad. Sci. Paris Sér. I Math., 331 (2000), pp. 101–106.
[15] W. B. Liu and J. E. Rubio, *Local convergences and optimal shape design*, SIAM J. Control Optim., 30 (1992), pp. 49–62.
[16] R. Mäkinen, P. Neittaanmäki, and D. Tiba, *On a fixed domain approach for a shape optimization problem*, in Computational and Applied Mathematics II, W. F. Ames and P. J. van der Houwen, eds., North–Holland, Amsterdam, 1992, pp. 317–326.
[17] V. Maz'ja, *Sobolev Spaces*, Springer-Verlag, Berlin, 1985.
[18] F. Murat and J. Simon, *Studies on Optimal Shape Design Problems*, Lecture Notes in Control and Inform. Sci. 41, Springer-Verlag, Berlin, 1976.
[19] P. Neittaanmäkiand D. Tiba, *Existence and Approximation in Optimal Shape Design Problems*, Tech. report 6, Department of Mathematics, Laboratory of Scientific Computing, University of Jyväskylä, Jyväskylä, Finland, 1998.
[20] O. Pironneau, *Optimal Shape Design for Elliptic Systems*, Springer-Verlag, Berlin, 1984.
[21] J. Sokolowski and J. P. Zolesio, *Introduction to Shape Optimization*, Springer-Verlag, Berlin, 1991.
[22] J. Sprekels and D. Tiba, *A duality approach in the optimization of beams and plates*, SIAM J. Control Optim., 37 (1998), pp. 486–501.
[23] V. Sverak, *On optimal shape design*, J. Math. Pures Appl. (9), 72 (1993), pp. 537–551.
[24] D. Tiba, *Lectures on the Control of Elliptic Systems*, Lecture Notes 32, Department of Mathematics, University of Jyväskylä, Jyväskylä, Finland, 1995.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 5, pp. 1455–1476

FLOW STABILITY OF PATCHY VECTOR FIELDS AND ROBUST FEEDBACK STABILIZATION*

FABIO ANCONA† AND ALBERTO BRESSAN‡

Abstract. The paper is concerned with *patchy vector fields*, a class of discontinuous, piecewise smooth vector fields that were introduced by the authors to study feedback stabilization problems. We prove the stability of the corresponding solution set w.r.t. a wide class of impulsive perturbations. These results yield the robustness of *patchy feedback controls* in the presence of measurement errors and external disturbances.

Key words. patchy vector field, impulsive perturbation, feedback stabilization, discontinuous feedback, robustness

AMS subject classifications. 34A, 34D, 49E, 93D

PII. S0363012901391676

1. Introduction and basic notation. The aim of this paper is to establish the stability of the set of trajectories of a patchy vector field w.r.t. various types of perturbations and the robustness of patchy feedback controls.

Patchy vector fields were introduced in [A-B] in order to study feedback stabilization problems. The underlying motivation is the following: The analysis of stabilization problems by means of Lyapunov functions usually leads to stabilizing feedbacks with a wild set of discontinuities. On the other hand, as shown in [A-B], by patching together open-loop controls one can always construct a piecewise smooth stabilizing feedback whose discontinuities have a very simple structure. In particular, one can develop the whole theory by studying the corresponding discontinuous ODEs within the classical framework of Carathéodory solutions. We recall here the main definitions.

DEFINITION 1.1. *By a patch we mean a pair* $\big(\Omega,\, g\big)$*, where* $\Omega \subset \mathbb{R}^n$ *is an open domain with smooth boundary* $\partial\Omega$ *and* g *is a smooth vector field defined on a neighborhood of the closure* $\overline{\Omega}$*, which points strictly inward at each boundary point* $x \in \partial\Omega$*.*

Calling $\mathbf{n}(x)$ the outer normal at the boundary point x, we thus require

$$\big\langle g(x),\ \mathbf{n}(x)\big\rangle < 0 \qquad \text{for all} \quad x \in \partial\Omega. \tag{1.1}$$

DEFINITION 1.2. *We say that* $g : \Omega \mapsto \mathbb{R}^n$ *is a patchy vector field on the open domain* Ω *if there exists a family of patches* $\big\{(\Omega_\alpha,\ g_\alpha);\quad \alpha \in \mathcal{A}\big\}$ *such that*

- $\mathcal{A}$ *is a totally ordered set of indices;*
- *the open sets* Ω_α *form a locally finite covering of* Ω*, i.e.,* $\Omega = \cup_{\alpha\in\mathcal{A}}\Omega_\alpha$ *and every compact set* $K \subset \mathbb{R}^n$ *intersect only a finite number of domains* Ω_α*,* $\alpha \in \mathcal{A}$*;*
- *the vector field* g *can be written in the form*

$$g(x) = g_\alpha(x) \qquad \text{if} \qquad x \in \Omega_\alpha \setminus \bigcup_{\beta>\alpha} \Omega_\beta. \tag{1.2}$$

*Received by the editors June 29, 2001; accepted for publication (in revised form) April 2, 2002; published electronically January 3, 2003.

http://www.siam.org/journals/sicon/41-5/39167.html

†Dipartimento di Matematica and C.I.R.A.M., Università di Bologna, Piazza Porta S. Donato 5, Bologna 40127, Italy (ancona@ciram3.ing.unibo.it).

‡S.I.S.S.A., Via Beirut 4, Trieste 34014, Italy (bressan@sissa.it).

By setting

$$\alpha^*(x) \doteq \max\big\{\alpha \in \mathcal{A}\,;\quad x \in \Omega_\alpha\big\}, \tag{1.3}$$

we can write (1.2) in the equivalent form

$$g(x) = g_{\alpha^*(x)}(x) \qquad \text{for all} \quad x \in \Omega. \tag{1.4}$$

Remark 1.1. The patches $(\Omega_\alpha,\, g_\alpha)$ are not uniquely determined by the patchy vector field g. Indeed, whenever $\alpha < \beta$, by (1.2) the values of g_α on the set $\Omega_\alpha \cap \Omega_\beta$ are irrelevant. Therefore, if the open sets Ω_α form a locally finite covering of Ω and we assume that, for each $\alpha \in \mathcal{A}$, the vector field g_α satisfies (1.1) at every point $x \in \partial\Omega_\alpha \setminus \bigcup_{\beta>\alpha} \Omega_\beta$, then the vector field g defined according with (1.2) is again a patchy vector field. To see this, it suffices to construct vector fields $\tilde g_\alpha$ which satisfy the inward-pointing property (1.1) at every point $x \in \partial\Omega_\alpha$ and such that $\tilde g_\alpha = g_\alpha$ on $\Omega_\alpha \setminus \bigcup_{\beta>\alpha} \Omega_\beta$. To accomplish this, for each α we first consider a smooth vector field v_α such that $v_\alpha(x) = -\mathbf{n}(x)$ on $\partial\,\Omega_\alpha$. Then we construct a smooth scalar function $\psi_\alpha : \Omega \mapsto [0,1]$ such that

$$\psi_\alpha(x) = \begin{cases} 1 & \text{if} \quad x \in \Omega_\alpha \setminus \bigcup_{\beta>\alpha} \Omega_\beta, \\ 0 & \text{if} \quad x \in \partial\Omega_\alpha, \quad \big\langle g(x),\ \mathbf{n}(x)\big\rangle \geq 0. \end{cases}$$

Finally, for each $\alpha \in \mathcal{A}$ we define the interpolation

$$\tilde g_\alpha(x) \doteq \psi_\alpha(x) g_\alpha(x) + \big(1 - \psi_\alpha(x)\big) v_\alpha(x).$$

The vector fields $\tilde g_\alpha$ thus defined satisfy our requirements.

We shall occasionally adopt the longer notation $\big(\Omega,\ g,\ (\Omega_\alpha,\, g_\alpha)_{\alpha\in\mathcal{A}}\big)$ to indicate a patchy vector field, specifying both the domain and the single patches. If g is a patchy vector field, the differential equation

$$\dot x = g(x) \tag{1.5}$$

has many interesting properties. In particular, in [A-B] it was proved that the set of Carathéodory solutions of (1.5) is closed in the topology of uniform convergence but possibly not connected. Moreover, given an initial condition

$$x(t_0) = x_0, \tag{1.6}$$

the Cauchy problem (1.5)–(1.6) has at least one forward solution and at most one backward solution. For every Carathéodory solution $x = x(t)$ of (1.5), the map $t \mapsto \alpha^*(x(t))$ is left continuous and nondecreasing.

In this paper we study the stability of the solution set for (1.5) w.r.t. various perturbations. Most of our analysis will be concerned with impulsive perturbations, described by

$$\dot y = g(y) + \dot w. \tag{1.7}$$

Here $w = w(t)$ is any left continuous function with bounded variation. By a solution of the perturbed system (1.7) with an initial condition

$$y(t_0) = y_0, \tag{1.8}$$

we mean a measurable function $t \mapsto y(t)$ such that

$$y(t) = y_0 + \int_{t_0}^{t} g\big(y(s)\big)\, ds + \big[w(t) - w(t_0)\big] \tag{1.9}$$

(see [B1]). If $w(\cdot)$ is discontinuous, the system (1.7) has impulsive behavior and the solution $y(\cdot)$ will be discontinuous as well. We choose to work with (1.7) because it provides a simple and general framework to study robustness properties. Indeed, consider a system with both inner and outer perturbations of the form

$$\dot{x} = g\big(x + e_1(t)\big) + e_2(t). \tag{1.10}$$

The map $t \mapsto y(t) \doteq x(t) + e_1(t)$ then satisfies the impulsive equation

$$\dot{y} = g(y) + e_2(t) + \dot{e}_1(t) = g(y) + \dot{w},$$

where

$$w(t) = e_1(t) + \int_{t_0}^{t} e_2(s)\, ds.$$

Therefore, from the stability of solutions of (1.7) w.r.t. perturbations w that have small total variation, one can immediately deduce a result on the stability of solutions of (1.10), when Tot.Var.$\{e_1\}$ and $\|e_2\|_{\mathbf{L}^1}$ are suitably small. Here, Tot.Var.$\{e_1\}$ denotes the total variation of the function e_1 over the whole interval where it is defined, while Tot.Var.$\big\{e_1\,;\,J\,\big\}$ denotes the total variation of e_1 over a subset J. We shall also denote by BV the space of all functions of bounded variation. Any function of bounded variation $w = w(t)$ can be redefined up to $\mathbf{L}^1$-equivalence. For the sake of definiteness, throughout the paper we shall always consider left continuous representatives, so that $w(t) = w(t^-) \doteq \lim_{s\to t-} w(s)$ for every t. The Lebesgue measure of a Borel set $J \subset \mathbb{R}$ will be denoted by meas(J).

We observe that since the Cauchy problem for (1.5) does not have forward uniqueness and continuous dependence, one clearly cannot expect that a single solution of (1.5) be stable under small perturbations. What we establish here is a different stability property, involving not a single trajectory but the whole solution set: If the perturbation w is small in the BV norm, then every solution of (1.7) is close to some solution of (1.5). This is essentially an upper semicontinuity property of the solution set. Namely, we will prove in section 2 the following results.

THEOREM 1.3. *Let g be a patchy vector field on an open domain $\Omega \subset \mathbb{R}^n$. Consider a sequence of solutions $y_\nu(\cdot)$ of the perturbed system*

$$\dot{y}_\nu = g(y_\nu) + \dot{w}_\nu, \qquad\qquad t \in [0,T], \tag{1.11}$$

with Tot.Var.$\{w_\nu\} \to 0$ as $\nu \to \infty$. If the $y_\nu : [0,T] \mapsto \Omega$ converge to a function $y : [0,T] \mapsto \Omega$, uniformly on $[0,T]$, then $y(\cdot)$ is a Carathéodory solution of (1.5) *on* $[0, T]$.

COROLLARY 1.4. *Let g be a patchy vector field on an open domain $\Omega \subset \mathbb{R}^n$. Given any closed subset $A \subset \Omega$, any compact set $K \subset A$, and any $T, \varepsilon > 0$, there exists $\delta = \delta(A, K, T, \varepsilon) > 0$ such that the following holds. If $y : [0,T] \mapsto A$ is a solution of the perturbed system* (1.7), *with $y(0) \in K$ and Tot.Var.$\{w\} < \delta$, then there exists a solution $x : [0,T] \mapsto \Omega$ of the unperturbed equation* (1.5) *with*

$$\big\|x - y\big\|_{\mathbf{L}^\infty([0,T])} < \varepsilon\,. \tag{1.12}$$

We remark that the type of stability described above is precisely what is needed in many feedback control applications. As an example, consider the problem of stabilizing to the origin the control system

$$\dot{x} = f(x,u). \tag{1.13}$$

Given a compact set K and $\varepsilon > 0$, assume that there exists a piecewise constant feedback $u = U(x)$ such that $g(x) \doteq f\big(x, U(x)\big)$ is a patchy vector field and such that every solution of (1.5) starting from a point $x(0) \in K$ is steered inside the ball B_ε centered at the origin with radius ε, within a time $T > 0$. By Corollary 1.4, if the perturbation w is sufficiently small (in the BV norm), every solution of the perturbed system (1.7) will be steered inside the ball $B_{2\varepsilon}$ within time T. In other words, the feedback still performs well in the presence of small perturbations. Applications to feedback control will be discussed in more detail in section 3.

Throughout the paper, by $B(x,r)$ we denote the closed ball centered at x with radius r and, for every given set A, we let $B(A,r) \doteq \cup_{x\in A} B(x,r)$. The closure, the interior, and the boundary of a set Ω are written as $\overline{\Omega}$, $\overset{\circ}{\Omega}$ and $\partial\Omega$, respectively.

2. Stability of patchy vector fields. We begin by proving a local existence result for solutions of the perturbed system (1.7).

PROPOSITION 2.1. *Let g be a patchy vector field on an open domain $\Omega \subset \mathbb{R}^n$. Given any compact set $K \subset \Omega$, there exists $\overline{\chi} = \overline{\chi}_K > 0$ such that, for each $y_0 \in K, t_0 \in \mathbb{R}$, and for every Lipschitz continuous function $w = w(t)$, with Lipschitz constant $\|\dot{w}\|_{\mathbf{L}^\infty} < \overline{\chi}$, the Cauchy problem* (1.7)–(1.8) *has at least one local forward solution.*

Proof. Fix some compact subset $K' \subset \Omega$ whose interior contains K. To prove the local existence of a forward solution to (1.7), first observe that because of the inward-pointing condition (1.1) and the smoothness assumptions on the vector fields g_α, one can find for any $\alpha \in \mathcal{A}$ some constant $\chi_\alpha > 0$ such that

$$\sup_{\substack{x\in\partial\Omega_\alpha\cap K'\\ |v|\leq\chi_\alpha}} \big\langle g_\alpha(x)+v,\, \mathbf{n}_{\boldsymbol{\alpha}}(x)\big\rangle < 0, \tag{2.1}$$

where $\mathbf{n}_{\boldsymbol{\alpha}}(x)$ is the outer normal to $\partial\Omega_\alpha$ at the boundary point x. Since K' is a compact set and $\{\Omega_\alpha\}_\alpha$ is a locally finite covering of Ω, there will be only finitely many elements of $\{\Omega_\alpha\}_\alpha$ that intersect K'. Let

$$\big\{\alpha_1,\ldots,\alpha_{_N}\big\} = \big\{\alpha\in\mathcal{A}\ :\ \Omega_\alpha\cap K'\neq\emptyset\big\}, \tag{2.2}$$

and, by possibly renaming the indices α_i, assume that

$$\alpha_1 < \cdots < \alpha_{_N}. \tag{2.3}$$

Choose a constant $\overline{\chi} > 0$ such that

$$\overline{\chi} \leq \inf\big\{\chi_{\alpha_i}\ :\ i=1,\ldots,N\big\}. \tag{2.4}$$

For any fixed $y_0 \in K$, consider the index

$$\widehat{\alpha}(y_0) \doteq \max\big\{\alpha\ :\ y_0\in\overline{\Omega}_\alpha\big\}.$$

By the definition of $\overline{\chi}$, any solution $y = y(\cdot)$ to the Cauchy problem

$$\dot{y} = g_{\widehat{\alpha}}(y) + \dot{w}, \qquad y(t_0) = y_0,$$

associated to a piecewise Lipschitz map $w = w(t)$ with $\|\dot{w}\|_{\mathbf{L}^\infty} < \overline{\chi}$, remains inside $\Omega_{\widehat{\alpha}}$ for all $t \in [t_0, t_0 + \delta]$ for some $\delta > 0$. Hence, it provides also a solution to (1.6) on some interval $[t_0, t_0 + \delta']$, $0 < \delta' \leq \delta$. □

Toward a proof of Theorem 1.3, we first derive an intermediate result. By the basic properties of a patchy vector field, for every solution $t \mapsto x(t)$ of (1.5) the corresponding map $t \mapsto \alpha^*\big(x(t)\big)$ in (1.3) is nondecreasing. Roughly speaking, a trajectory can move from a patch Ω_α to another patch Ω_β only if $\alpha < \beta$. This property no longer holds in the presence of an impulsive perturbation. However, the next proposition shows that for a solution y of (1.7) the corresponding map $t \mapsto \alpha^*\big(y(t)\big)$ is still nondecreasing, after a possible modification on a small set of times. Alternatively, one can slightly modify the impulsive perturbation w, say replacing it by another perturbation $w^\diamondsuit$, such that the map $t \mapsto \alpha^*\big(y^\diamondsuit(t)\big)$ is monotone along the corresponding trajectory $t \mapsto y^\diamondsuit(t)$.

PROPOSITION 2.2. *Let g be a patchy vector field on an open domain $\Omega \subset \mathbb{R}^n$ determined by the family of patches $\big\{(\Omega_\alpha,\ g_\alpha);\ \alpha \in \mathcal{A}\big\}$. For any $T > 0$ and any compact set $K \subset \Omega$, there exist constants $C, \delta > 0$ and an integer N such that the following hold.*

(i) *For every $w \in \mathrm{BV}$ with Tot.Var.$\{w\} < \delta$, and for every solution $y : [0, T] \mapsto \Omega$ of the Cauchy problem (1.7)–(1.8) with $y_0 \in K$, there is a partition of $[0,\ T]$, $0 = \tau_1 \leq \tau_2 \leq \cdots \leq \tau_{N+1} = T$, and indices*

$$\alpha_1 < \alpha_2 < \cdots < \alpha_N \tag{2.5}$$

such that

$$\alpha^*(y(t)) \geq \alpha_i \qquad \text{for all } t \in \,]\tau_i,\ \tau_{i+1}], \qquad i = 0, \ldots, N, \tag{2.6}$$

$$\text{meas}\Big(\bigcup_{i \geq 0} \big\{t \in [\tau_i,\ \tau_{i+1}]\ :\ \alpha^*(y(t)) > \alpha_i\big\}\Big) < C \cdot \textit{Tot.Var.}\{w\}. \tag{2.7}$$

(ii) *For every BV function $w = w(t)$ with Tot.Var.$\{w\} < \delta$, and for every solution $y : [0, T] \mapsto \Omega$ of the Cauchy problem (1.7)–(1.8) with $y_0 \in K$, there is a BV function $w^\diamondsuit = w^\diamondsuit(t)$ and a solution $y^\diamondsuit : [0, T] \mapsto \Omega$ of*

$$\dot{y}^\diamondsuit = g(y^\diamondsuit) + \dot{w}^\diamondsuit \tag{2.8}$$

so that the map $t \mapsto \alpha^*(y^\diamondsuit(t))$ is nondecreasing and left continuous, and there holds

$$\begin{aligned} \textit{Tot.Var.}\{w^\diamondsuit\} &\leq C \cdot \textit{Tot.Var.}\{w\}, \\ \big\|y^\diamondsuit - y\big\|_{\mathbf{L}^\infty([0,T])} &\leq C \cdot \textit{Tot.Var.}\{w\}. \end{aligned} \tag{2.9}$$

Proof. (i) The proof of (i) will be given in three steps.

Step 1. Since each g_α is a smooth vector field and we are assuming a uniform bound on the total variation of every perturbation $w = w(t)$, there will be some compact subset $K' \subset \overline{\Omega}$ that contains every solution $y : [0, T] \mapsto \Omega$ of (1.7) starting at a point $y_0 \in K$. We will assume without loss of generality that every domain Ω_α is bounded since, otherwise, one can replace Ω_α with its intersection $\Omega_\alpha \cap \Omega'$ with a

bounded domain $\Omega' \subset \Omega$ that contains K', preserving the inward-pointing condition (1.1) (cf. Remark 1.1). For each $\alpha \in \mathcal{A}$, define the map $\varphi_\alpha : \Omega \mapsto \mathbb{R}$ by setting

$$\varphi_\alpha(x) \doteq \begin{cases} d(x,\, \partial\Omega_\alpha) & \text{if} \quad x \in \Omega_\alpha, \\ -d(x,\, \partial\Omega_\alpha) & \text{otherwise,} \end{cases} \tag{2.10}$$

and let

$$\varphi^+_\alpha(x) \doteq \max\{\varphi_\alpha(x),\, 0\}$$

denote the positive part of $\varphi_\alpha(x)$. The regularity assumptions on the patch $(\Omega_\alpha, g_\alpha)$ guarantee that φ_α is smooth if restricted to a sufficiently small neighborhood of the boundary $\partial\Omega_\alpha$. Thus, if $\{\Omega_{\alpha_i} \,:\, i = 1, \dots, N\}$ denotes the finite collection of domains that intersect the compact set K' as in (2.2)–(2.3), there will be some constant $\overline{\rho} > 0$ so that, setting

$$\Omega^{\overline{\rho}}_\alpha \doteq \big\{x \in \Omega \;:\; d(x,\, \partial\Omega_\alpha) \geq \overline{\rho}\big\}, \tag{2.11}$$

the restriction of any map φ_{α_i} to the domain $\Omega \setminus \Omega^{\overline{\rho}}_{\alpha_i}$ will be smooth. In particular, for any $i = 1, \dots, N$, we will have

$$\nabla\varphi_{\alpha_i}(x) = -\mathbf{n}_{\boldsymbol{\alpha_i}}\big(\pi_{\alpha_i}(x)\big) \qquad \text{for all} \quad x \in \Omega \setminus \Omega^{\overline{\rho}}_{\alpha_i}, \tag{2.12}$$

where $\mathbf{n}_{\boldsymbol{\alpha_i}}$ represents as usual the outer normal to $\partial\Omega_{\alpha_i}$, while $\pi_{\alpha_i}(x)$ denotes the projection of the point x onto the set $\partial\Omega_{\alpha_i}$. On the other hand, thanks to the inward-pointing condition (1.1), we can choose the constant $\overline{\rho}$ so that

$$\sup_{\substack{i=1,\dots,N \\ x\in\Omega_{\alpha_i}\setminus\Omega^{\overline{\rho}}_{\alpha_i}}} \big\langle g_{\alpha_i}(x),\, \mathbf{n}_{\boldsymbol{\alpha_i}}(\pi_{\alpha_i}(x))\big\rangle \leq -c' \tag{2.13}$$

for some $c' > 0$. Moreover, the smoothness of the fields g_α on $\overline{\Omega}$ implies the existence of some $c'' > 0$ such that

$$\sup_{\substack{i=1,\dots,N, j>i \\ x\in\Omega_{\alpha_i}}} \Big|\big\langle g_{\alpha_j}(x),\, \mathbf{n}_{\boldsymbol{\alpha_i}}(\pi_{\alpha_i}(x))\big\rangle\Big| \leq c''. \tag{2.14}$$

Step 2. Consider now a left continuous BV function $w = w(t)$ and let $y : [0, T] \mapsto \Omega$ be a solution of the corresponding Cauchy problem (1.7)–(1.8), with $y_0 \in K$. Observe that, for any $i = 1, \dots, N$, and for any interval $J \subset [0, T]$ such that

$$y(t) \in \Omega \setminus \Omega^{\overline{\rho}}_{\alpha_i} \qquad \text{for all} \quad t \in J,$$

the composed map $\varphi^+_{\alpha_i} \circ y : J \mapsto \mathbb{R}$ is also a left continuous BV function whose distributional derivative $\mu_i \doteq D(\varphi^+_{\alpha_i} \circ y)$ is a Radon measure, which can be decomposed into an absolutely continuous μ^{ac}_i and a singular part μ^s_i w.r.t. the Lebesgue measure dt. One can easily verify that for any Borel set $E \subset J$, the absolutely continuous part of μ_i is given by

$$\mu^{ac}_i(E) = \int_{E^+} \big\langle \nabla\varphi_{\alpha_i}(y(t)),\; g(y(t)) + \dot{w}(t)\big\rangle\, dt, \qquad E^+ \doteq \{t \in E \,;\, y(t) \in \Omega_{\alpha_i}\}. \tag{2.15}$$

Moreover, calling μ_w^{ac} and μ_w^s, respectively, the absolutely continuous and the singular part of $\mu_w \doteq \dot{w}$, the following bounds hold:

$$\begin{aligned}\left| \int_{E^+} \big\langle \nabla \varphi_{\alpha_i}(y(t)),\ \dot{w}(t) \big\rangle\ dt + \mu_i^s(E) \right| &\leq c''' \cdot \Big\{ \big|\mu_w^{ac}(E)\big| + \big|\mu_w^s(E)\big| \Big\} \\ &\leq c''' \cdot \text{Tot.Var.}\{w\} \end{aligned} \tag{2.16}$$

for some constant $c''' > 0$ that depends only on the compact set K' and on the time interval $[0, T]$. Let $C_i, \ell_i,\ i = N, N-1, \dots, 1$, be the constants recursively defined by

$$C_N \doteq 1 + c''', \qquad \ell_N \doteq \frac{2C_N}{c'}, \tag{2.17}$$

$$C_i \doteq c'' \cdot \ell_{i+1} + \sum_{j=i+1}^{N} C_j, \qquad \ell_i \doteq \frac{1}{c'} \left(2C_i + c'' \cdot \sum_{j=i+1}^{N} \ell_j \right) \quad \text{if} \quad i < N. \tag{2.18}$$

□

LEMMA 2.3. *Assume that*

$$\textit{Tot. Var.}\{w\} < \delta \doteq \frac{\overline{\rho}}{2C_1}, \tag{2.19}$$

and assume that there exists some interval $[t_1, t_2] \subset [0, T]$ *and some index* $i \in \{1, \dots, N\}$ *such that*

$$\text{meas}\big\{ t \in [t_1, t_2]\ :\ \alpha^*(y(t)) = \alpha_j \big\} \leq \ell_j \cdot \textit{Tot. Var.}\{w\} \qquad \textit{for all}\ \ j > i \tag{2.20$_i$}$$

together with one of the following two conditions:

(a)$_i$

$$\varphi_{\alpha_i}(y(t)) < 2C_i \cdot \textit{Tot. Var.}\{w\} \qquad \textit{for all}\ \ t \in [t_1, t_2], \tag{2.21$_i$}$$

$$\text{meas}\big\{ t \in [t_1, t_2]\ :\ \alpha^*(y(t)) = \alpha_i \big\} > \ell_i \cdot \textit{Tot. Var.}\{w\}. \tag{2.22$_i$}$$

(b)$_i$ *There exists* $\tau \in [t_1, t_2]$ *such that*

$$\varphi_{\alpha_i}(y(\tau)) \geq 2C_i \cdot \textit{Tot. Var.}\{w\}. \tag{2.23$_i$}$$

Then one has

$$\varphi_{\alpha_i}(y(t_2)) \geq C_i \cdot \textit{Tot. Var.}\{w\}. \tag{2.24$_i$}$$

Proof of Lemma 2.3. Observe first that the recursive definition (2.17)–(2.18) of the constants C_i, ℓ_i and the bound (2.19) clearly imply

$$C_i \geq 1 + c''' + c'' \cdot \sum_{j=i+1}^{N} \ell_j, \tag{2.25}$$

$$2C_i \cdot \text{Tot.Var.}\{w\} < \overline{\rho}. \tag{2.26}$$

Assume now that (2.20_i)–(2.22_i) hold. Then, using (2.13)–(2.16) and recalling (2.25)–(2.26), we obtain

$$\begin{aligned}
\varphi^+_{\alpha_i}(y(t_2)) &\geq \varphi^+_{\alpha_i}(y(t_1)) + \int_{\{t\in[t_1,t_2]\,:\,\alpha^*(t)=\alpha_i\}} \big\langle \nabla\varphi_{\alpha_i}(y(t)),\ g_{\alpha_i}(y(t))\big\rangle\, dt \\
&\quad - \sum_{j=i+1}^{N} \int_{\{t\in[t_1,t_2]\,:\,\alpha^*(t)=\alpha_j\}} \Big|\big\langle \nabla\varphi_{\alpha_i}(y(t)),\ g_{\alpha_j}(y(t))\big\rangle\Big|\, dt - c'''\cdot \text{Tot.Var.}\{w\} \\
&\geq \int_{\{t\in[t_1,t_2]\,:\,\alpha^*(t)=\alpha_i\}} -\big\langle \mathbf{n_{\alpha_i}}(\pi_{\alpha_i}(y(t))),\ g_{\alpha_i}(y(t))\big\rangle\, dt \\
&\quad - \Big(c''\cdot \sum_{j=i+1}^{N} \ell_j + c'''\Big)\cdot \text{Tot.Var.}\{w\} \\
&\geq \Big(\ell_i\cdot c' - c''\cdot \sum_{j=i+1}^{N} \ell_j - c'''\Big)\cdot \text{Tot.Var.}\{w\} \\
&\geq C_i\cdot \text{Tot.Var.}\{w\}, \qquad (2.27)
\end{aligned}$$

proving (2.24_i). Next, assume that (2.20_i) and (2.23_i) hold, and let

$$\tau' \doteq \sup\big\{t\in[t_1,\,t_2]\ :\ \varphi_{\alpha_i}(y(t)) > 2C_i\cdot \text{Tot.Var.}\{w\}\big\}. \qquad (2.28)$$

Clearly, the bound (2.24_i) is satisfied if $\tau' = t_2$ since the map φ_{α_i} is left continuous. Next, consider the case $\tau' < t_2$. By computations similar to those in (2.27), using (2.13)–(2.16), and thanks to (2.20_i), (2.25)–(2.26), we get

$$\begin{aligned}
\varphi^+_{\alpha_i}(y(t_2)) &\geq \varphi^+_{\alpha_i}(y(\tau')) + \int_{\{t\in[\tau',t_2]\,:\,\alpha^*(t)=\alpha_i\}} \big\langle \nabla\varphi_{\alpha_i}(y(t)),\ g_{\alpha_i}(y(t))\big\rangle\, dt \\
&\quad - \sum_{j=i+1}^{N} \int_{\{t\in[\tau',t_2]\,:\,\alpha^*(t)=\alpha_j\}} \Big|\big\langle \nabla\varphi_{\alpha_i}(y(t)),\ g_{\alpha_j}(y(t))\big\rangle\Big|\, dt - c'''\cdot \text{Tot.Var.}\{w\} \\
&\geq \left(2C_i - 1 - c''' - c''\cdot \sum_{j=i+1}^{N} \ell_j\right)\cdot \text{Tot.Var.}\{w\} \\
&\geq C_i\cdot \text{Tot.Var.}\{w\}, \qquad (2.29)
\end{aligned}$$

thus concluding the proof of Lemma 2.3. □

Step 3. Assume that the bound (2.19) on the total variation of $w = w(t)$ holds. Set $\tau_1 = 0$, $\tau_{N+1} \doteq T$, and define recursively the points $\tau_N, \tau_{N-1}, \dots, \tau_2$ by setting, for every $1 < i \leq N$,

(2.30_i)

$$\mathcal{T}_i \doteq \Big\{t\in[0,\,\tau_{i+1}]\ :\ \varphi_{\alpha_i}(y(s)) \geq C_i\cdot \text{Tot.Var.}\{w\} \qquad \text{for all}\ \ s\in[t,\,\tau_{i+1}]\Big\},$$

$$\tau_i \doteq \begin{cases} \inf \mathcal{T}_i & \text{if}\quad \mathcal{T}_i \neq \emptyset, \\ \tau_{i+1} & \text{if}\quad \mathcal{T}_i = \emptyset. \end{cases}$$

Applying Lemma 2.3 and proceeding by backward induction on $i = N, N-1, \dots, 2$,

we show now that for any $t < \tau_i$, $i = 2, \dots, N$, one has

$$\begin{aligned} &\text{meas}\big\{s \in [0,\, t] \ :\ \alpha^*(y(s)) = \alpha_i\big\} \leq \ell_i \cdot \text{Tot.Var.}\{w\}, \\ &\varphi_{\alpha_i}(y(t)) < 2C_i \cdot \text{Tot.Var.}\{w\}. \end{aligned} \tag{2.31$_i$}$$

Indeed, if (2.31$_i$) is not satisfied, one of the two conditions (a)$_{\rm N}$ or (b)$_{\rm N}$ must be true on some interval $[0,\, \bar t\,]$, $\bar t < \tau_N$. But then, by (2.24)$_N$, we have

$$\varphi_{\alpha_N}(y(s)) \geq C_N \cdot \text{Tot.Var.}\{w\} \qquad \text{for all} \ \ s \in [\,\bar t,\, T],$$

which contradicts the definition (2.30$_i$). On the other hand, if we assume that (2.31)$_j$ holds for $j = i+1, \dots, N$ but not for $j = i$, then one of the two conditions (a)$_i$ or (b)$_i$ must be true on some interval $[0,\, \bar t\,]$, $\bar t < \tau_i$. Moreover, the inductive assumptions (2.31)$_j$, $j > i$, imply (2.20$_i$) and hence, as above, thanks to (2.24$_i$), we get

$$\varphi_{\alpha_i}(y(s)) \geq C_i \cdot \text{Tot.Var.}\{w\} \qquad \text{for all} \ \ s \in [\bar t,\, T],$$

reaching a contradiction with the definition (2.30$_i$).

To conclude the proof of property (i) stated in Proposition 2.2, observe that, thanks to (2.31$_i$), $i = 2, \dots, N$, we have

$$\text{meas}\big\{s \in [\tau_i,\, \tau_{i+1}] \ :\ \alpha^*(y(s)) > \alpha_i\big\} \leq \left(\sum_{j>i} \ell_j\right) \cdot \text{Tot.Var.}\{w\} \qquad \text{for all} \ \ i \geq 1. \tag{2.32}$$

Therefore, recalling the definitions of the map φ_{α_i} at (2.10), taking δ as in (2.19), and

$$C > (N+1) \cdot \sum_{j=1}^{N} \ell_j, \tag{2.33}$$

from (2.31$_i$) and (2.32) we deduce that the partition $\tau_1 = 0 \leq \tau_2 \leq \cdots \leq \tau_{N+1} = T$ of $[0,\, T]$, defined at (2.30$_i$), satisfies the properties (2.5)–(2.7).

(ii) Concerning (ii), let $C, \delta > 0$ be the constants defined according to (i) and, given a BV function $w = w(t)$ with Tot.Var.$\{w\} < \delta$, and a solution $y : [0, T] \mapsto \Omega$ of the Cauchy problem (1.7)–(1.8) with $y_0 \in K$, consider the partition $0 = \tau_1 \leq \tau_2 \leq \cdots \leq \tau_{N+1} = T$, of $[0,\, T]$, with the properties in (i). Setting

$$\tau_i' \doteq \inf\Big\{t \in [\tau_i,\, \tau_{i+1}] \ :\ \alpha^*(y(t)) = \alpha_i\Big\}, \qquad i = 1, \dots, N,$$

define the map

$$\tau(t) \doteq \begin{cases} \tau_i' & \text{if} \quad t \in \,]\tau_i,\, \tau_i'], \\ \sup\big\{s \in [\tau',\, t] \ :\ \alpha^*(y(s)) = \alpha_i\big\} & \text{if} \quad t \in \,]\tau_i',\, \tau_{i+1}] \end{cases} \tag{2.34}$$

over any interval $]\tau_i,\, \tau_{i+1}]$, $i = 1, \dots, N$. Notice that in the particular case where $\alpha^*(y(t)) > \alpha_i$ for all $t \in \,]\tau_i,\, \tau_{i+1}]$, by the above definitions one has $\tau(t) = \tau_i' = \tau_{i+1}$ for any $t \in \,]\tau_i,\, \tau_{i+1}]$. Then let $y^\diamondsuit : [0, T] \mapsto \Omega$ be the map recursively defined by setting

$$y^\diamondsuit(t) \doteq y(t) \qquad \text{for all} \ \ t \in \,]\tau_N,\, T], \tag{2.35}$$

(2.36)
$$y^\diamond(t) \doteq \begin{cases} y^\diamond(\tau_{i+1}+) & \text{if} \quad \tau_i' = \tau_{i+1}, \\ y(\tau(t)+) & \text{if} \quad \tau_i' < \tau_{i+1}, \quad \alpha^*(y(\tau(t))) > \alpha_i, \\ y(\tau(t)) & \text{if} \quad \tau_i' < \tau_{i+1}, \quad \alpha^*(y(\tau(t))) = \alpha_i, \end{cases} \qquad \text{for all} \quad t \in]\tau_i,\, \tau_{i+1}], \quad i < N,$$

(2.37)
$$y^\diamond(0) \doteq y^\diamond(0^+),$$

and let $w^\diamond = w^\diamond(t)$ be the function defined as

(2.38)
$$w^\diamond(t) \doteq y^\diamond(t) - \int_0^t g\big(y^\diamond(s)\big)\, ds \qquad \text{for all} \quad t \in [0,\, T].$$

Clearly $y^\diamond, w^\diamond$ are both BV functions as well as y, w. Moreover, $y^\diamond$ is a solution of the perturbed equation (2.8). By construction, for every $1 \leq i \leq N$ there holds

(2.39)
$$\alpha^*(y^\diamond(t)) = \begin{cases} \alpha_i & \text{if} \quad \tau_i' < \tau_{i+1}, \\ \alpha^*(y^\diamond({\tau_{i+1}}^+)) & \text{if} \quad \tau_i' = \tau_{i+1}, \end{cases} \qquad \text{for all} \quad t \in]\tau_i,\, \tau_{i+1}].$$

Hence the map $t \mapsto \alpha^*(y^\diamond(t))$ is nonincreasing and left continuous. Next, recalling (2.6) and observing that

(2.40)
$$\alpha^*(y(t)) = \alpha_i \qquad \Longrightarrow \qquad \begin{matrix} \tau(t) = t, \\ y^\diamond(t) = y(t), \end{matrix} \qquad \text{for all} \quad t \in]\tau_i,\, \tau_{i+1}]$$

and defining

(2.41)
$$\mathcal{I} \doteq \bigcup_i \big\{t \in]\tau_i,\, \tau_{i+1}] \ :\ \alpha^*(y(t)) > \alpha_i\big\},$$

we have

(2.42)
$$y(t) = y^\diamond(t) \qquad \text{for all} \quad t \in (0,\, T) \setminus \mathcal{I}.$$

On the other hand, by the above definitions, calling $M \doteq \sup_{y \in \Omega} |g(y)|$, we derive

(2.43)
$$\big|\tau(t) - t\big| \leq \text{meas}(\mathcal{I}) \qquad \text{for all} \quad t \in \mathcal{I},$$

$$\begin{aligned} \big|y^\diamond(t) - y(t)\big| &\leq \int_{\tau(t)}^t \big|g(y(s))\big|\, ds + \text{Tot.Var.}\big\{w\,;\,[0,t]\big\} \\ &\leq M \cdot \text{meas}(\mathcal{I}) + \text{Tot.Var.}\{w\} \qquad \text{for all} \quad t \in \mathcal{I}, \end{aligned} \tag{2.44}$$

and

$$\begin{aligned} \Big|\text{Tot.Var.}\{y^\diamond\} - \text{Tot.Var.}\{y\}\Big| &\leq \text{Tot.Var.}\big\{y\,;\,\mathcal{I}\big\} \\ &\leq M \cdot \text{meas}(\mathcal{I}) + \text{Tot.Var.}\{w\}. \end{aligned} \tag{2.45}$$

Then, using (2.44)–(2.45), we obtain

$$\begin{aligned}\Big|\text{Tot.Var.}\{w^\diamondsuit\} - \text{Tot.Var.}\{w\}\Big| &\leq \int_{\mathcal{I}} \Big|\big|g\big(y^\diamondsuit(s)\big)\big| - \big|g\big(y(s)\big)\big|\Big|\, ds \\ &\quad + \Big|\text{Tot.Var.}\{y^\diamondsuit\} - \text{Tot.Var.}\{y\}\Big| \\ &\leq M' \cdot \Big\{\text{meas}(\mathcal{I}) + \text{Tot.Var.}\{w\}\Big\} \end{aligned} \tag{2.46}$$

for some constant $M' > 0$, depending only on the field g. Hence, from (2.42), (2.44), (2.46), and applying (2.7), it follows that $y^\diamondsuit(\cdot)$ satisfies the estimates in (2.9) for some constant $C' > 0$, which concludes the proof of (ii).

We can now take δ as in (2.19) and choose $C > C'$ according to (2.33). Both properties (i) and (ii) are then satisfied, completing the proof of Proposition 2.2. ∎

Proof of Theorem 1.3. For a given sequence of solutions $y_\nu : [0,T] \mapsto \Omega$ of the perturbed system (1.11) with $\text{Tot.Var.}\{w_\nu\} \leq \delta_\nu$, $\delta_\nu \to 0$ as $\nu \to \infty$, assume that the $y_\nu(\cdot)$ converge to a function $y : [0,T] \mapsto \Omega$ uniformly on $[0,T]$ and that $y_\nu(0)$ belongs to some compact set $K \subset \Omega$ for every ν. Thanks to property (ii) of Proposition 2.2, in connection with any pair $w_\nu(\cdot)$, $y_\nu(\cdot)$, there will be a BV function $w_\nu^\diamondsuit(\cdot)$ and a solution $y_\nu^\diamondsuit(\cdot)$ of (2.8) that satisfy

$$\text{Tot.Var.}\{w_\nu^\diamondsuit\} \leq C' \cdot \delta_\nu, \qquad \big\|y_\nu^\diamondsuit - y_\nu\big\|_{\mathbf{L}^\infty([0,T])} \leq C' \cdot \delta_\nu \tag{2.47}$$

for some constant $C' > 0$ independent of ν. Moreover there exists a partition $0 = \tau_{1,\nu} \leq \tau_{2,\nu} \leq \cdots \leq \tau_{N+1,\nu} = T$ of $[0, T]$ such that

$$\alpha^*(y_\nu^\diamondsuit(t)) = \alpha_i \qquad \text{for all } \ t \in \,]\tau_{i,\nu},\, \tau_{i+1,\nu}], \qquad i = 1, \ldots, N. \tag{2.48}$$

Recalling (1.4) and (1.9), because of (2.48) we have

$$\begin{aligned} y_\nu^\diamondsuit(t) = y_\nu^\diamondsuit(0) \ + \ \sum_{\ell=1}^{i-1} \int_{\tau_{\ell,\nu}}^{\tau_{\ell+1,\nu}} g_{\alpha_\ell}(y_\nu^\diamondsuit(s))\, ds \ + \int_{\tau_{i,\nu}}^{t} g_{\alpha_i}(y_\nu^\diamondsuit(s))\, ds \ + \ [w_\nu^\diamondsuit(t) - w_\nu^\diamondsuit(0)] \\ \text{for all } \ t \in [\tau_{i,\nu},\, \tau_{i+1,\nu}], \qquad i = 1, \ldots, N. \end{aligned} \tag{2.49}$$

By possibly taking a subsequence, we can assume that every sequence $(\tau_{i,\nu})_{\nu \geq 1}$ converges to some limit point, say

$$\overline{\tau}_i \doteq \lim_{\nu \to \infty} \tau_{i,\nu}, \qquad i = 1, \ldots, N+1.$$

We now observe that

$$\,]\overline{\tau}_i,\, \overline{\tau}_{i+1}[\ \subseteq \bigcup_{\mu=1}^{\infty} \bigcap_{\nu=\mu}^{\infty}]\tau_{i,\nu},\, \tau_{i+1,\nu}\,] \qquad \text{for all } \ i.$$

Moreover, the second inequality in (2.47) and the uniform convergence $y_\nu(\cdot) \to y(\cdot)$ yield

$$\lim_{\nu \to \infty} \big\|y_\nu^\diamondsuit - y\big\|_{\mathbf{L}^\infty([0,T])} = 0. \tag{2.50}$$

From the first inequality in (2.47), and from (2.48)–(2.49), we now deduce

(2.51)
$$y(t) \in \overline{\Omega}_{\alpha_i} \setminus \bigcup_{\beta>\alpha_i} \Omega_\beta,$$
$$\text{for all } \ t \in]\overline{\tau}_i,\, \overline{\tau}_{i+1}], \ \text{for all } \ i.$$
$$y(t) = y(0) + \sum_{\ell=1}^{i-1} \int_{\overline{\tau}_\ell}^{\overline{\tau}_{\ell+1}} g_{\alpha_\ell}(y(s))\, ds + \int_{\overline{\tau}_i}^{t} g_{\alpha_i}(y(s))\, ds$$

In particular, on each interval $[\overline{\tau}_i,\, \overline{\tau}_{i+1}]$, the function $y(\cdot)$ is a classical solution of $\dot y = g_{\alpha_i}(y)$ and satisfies

$$\dot y(s^-) = g_{\alpha_i}(y(s)) \qquad \text{for all } \ s \in]\overline{\tau}_i,\, \overline{\tau}_{i+1}].$$

Moreover, observe that because of the inward-pointing condition (1.1), the set $\big\{t \in [\overline{\tau}_i,\, \overline{\tau}_{i+1}] : y(t) \in \partial\,\Omega_{\alpha_i}\big\}$ is nowhere dense in $[\overline{\tau}_i,\, \overline{\tau}_{i+1}]$. Thus, if s is any point in $]\overline{\tau}_i,\, \overline{\tau}_{i+1}]$ such that $y(s) \in \partial\,\Omega_{\alpha_i}$, there will be some increasing sequence $(s_n)_n \subset]\overline{\tau}_i,\, \overline{\tau}_{i+1}[$ converging to s and such that $y(s_n) \in \Omega_{\alpha_i}$ for any n. But this yields a contradiction with (1.1), because

$$0 \leq \lim_{n\to\infty} \Big\langle \frac{y(s) - y(s_n)}{s - s_n},\ \mathbf{n_{\alpha_i}}\big(y(s)\big) \Big\rangle = \Big\langle \dot y(s-),\ \mathbf{n}\big(y(s)\big) \Big\rangle = \Big\langle g_{\alpha_i}\big(y(s)\big),\ \mathbf{n_{\alpha_i}}\big(y(s)\big) \Big\rangle.$$

Hence, recalling the definition (1.2), from (2.51) we conclude

$$y(t) \in \Omega_{\alpha_i} \setminus \bigcup_{\beta>\alpha_i} \Omega_\beta \qquad \text{for all } \ t \in]\overline{\tau}_i,\, \overline{\tau}_{i+1}], \qquad i = 1, \ldots, N,$$

$$y(t) = y(0) + \int_0^t g\big(y(s)\big)\, ds \qquad \text{for all } \ t \in [0,\, T],$$

proving that $y : [0,\, T] \mapsto \Omega$ is a Carathéodory solution of (1.5) on $[0,\, T]$. □

Proof of Corollary 1.4. Assuming that statement is false, we shall reach a contradiction. Fix any closed subset $A \subset \Omega$, any compact set $K \subset A$, and assume that, for some $T, \varepsilon > 0$, there exists a sequence of solutions $y_\nu : [0, T] \mapsto A$ of the perturbed system (1.7), with $y_\nu(0) \in K$, Tot.Var.$\{w_\nu\} \leq \delta_\nu$, $\delta_\nu \to 0$ as $\nu \to \infty$, such that the following property holds.

(P) Every solution $x : [0, T] \mapsto \Omega$ of the unperturbed equation (1.5) satisfies

$$\big\|x - y_\nu\big\|_{\mathbf{L}^\infty([0,T])} \geq \varepsilon \qquad \text{for all } \ \nu. \tag{2.52}$$

For each ν, call $y_\nu^\diamondsuit : [0, T] \mapsto \mathbb{R}^n$ the polygonal curve with vertices at the points $y_\nu(\ell\delta_\nu), \ell \geq 0$, defined by setting

$$y_\nu^\diamondsuit(t) \doteq y_\nu\big(\ell\delta_\nu\big) + \frac{t - \ell\delta_\nu}{\delta_\nu} \cdot \Big(y_\nu\big((\ell+1)\delta_\nu\big) - y_\nu\big(\ell\delta_\nu\big)\Big)$$
$$\text{for all } \ t \in [\ell\delta_\nu,\, (\ell+1)\delta_\nu] \cap [0, T], \qquad 0 \leq \ell \leq \lfloor T/\delta_\nu \rfloor, \tag{2.53}$$

where $\lfloor T/\delta_\nu \rfloor$ denotes the integer part of T/δ_ν. Since every $y_\nu(\cdot)$ is a BV function that solves (1.7), it follows that there will be some constant $C > 0$, independent of ν, such that

$$\text{Tot.Var.}\{y_\nu\ ;\ J\} \leq C \cdot \text{meas}(J) + \text{Tot.Var.}\{w_\nu\ ;\ J\} \tag{2.54}$$

for any interval $J \subset [0,\, T]$. Then, using (2.54), we derive for any fixed $0 \leq \ell < \ell' \leq \lfloor T/\delta_\nu \rfloor$ the bound

$$\begin{aligned}\Big|y_\nu^\diamondsuit(\ell'\delta_\nu) - y_\nu^\diamondsuit(\ell\delta_\nu)\Big| &= \Big|y_\nu(\ell'\delta_\nu) - y_\nu(\ell\delta_\nu)\Big| \\ &\leq \text{Tot.Var.}\big\{y_\nu\,;\ [\ell\delta_\nu,\, \ell'\delta_\nu]\big\} \\ &\leq (1+C)\cdot(\ell'-\ell)\delta_\nu. \end{aligned} \tag{2.55}$$

Therefore $y_\nu^\diamondsuit(\cdot)$ is a uniformly bounded sequence of Lipschitz maps, having Lipschitz constant $\text{Lip}(y_\nu^\diamondsuit) \leq (1+C)$. Hence, applying the Ascoli–Arzelà theorem, we can find a subsequence, which we still denote $y_\nu^\diamondsuit(\cdot)$, that converges to some function $y : [0,\, T] \mapsto \mathbb{R}^n$, uniformly on $[0,\, T]$. On the other hand, by construction, and thanks to (2.54), for any fixed $0 \leq t \leq T$, with $\ell\delta_\nu \leq t < (\ell+1)\delta_\nu$, the following holds:

$$\begin{aligned}\Big|y_\nu(t) - y_\nu^\diamondsuit(t)\Big| &\leq \Big|y_\nu(t) - y_\nu\big(\ell\delta_\nu\big)\Big| + \Big|y_\nu\big(\ell\delta_\nu\big) - y_\nu^\diamondsuit(t)\Big| \\ &\leq \Big|y_\nu(t) - y_\nu(\ell\delta_\nu)\Big| + \Big|y_\nu\big((\ell+1)\delta_\nu\big) - y_\nu\big(\ell\delta_\nu\big)\Big| \\ &\leq 2\cdot\text{Tot.Var.}\big\{y_\nu\,;\ [\ell\delta_\nu,\, (\ell+1)\delta_\nu]\big\} \\ &\leq 2(1+C)\cdot\delta_\nu. \end{aligned} \tag{2.56}$$

Thus, since $\delta_\nu \to 0$ as $\nu \to \infty$, the uniform convergence of $y_\nu^\diamondsuit(\cdot)$ to $y(\cdot)$ implies

$$\lim_{\nu\to\infty} \big\|y_\nu - y\big\|_{\mathbf{L}^\infty([0,T])} = 0. \tag{2.57}$$

By assumption, $\text{Range}(y_\nu) \subset A \subset \Omega$ for every ν, and hence from (2.57) we deduce also that the limit function $y(\cdot)$ takes values inside Ω. We can thus apply Theorem 1.3 to the sequence $y_\nu(\cdot)$ and conclude that the function $y : [0,\, T] \mapsto \Omega$ is a Carathéodory solution of the unperturbed equation (1.5) with

$$\big\|y - y_\nu\big\|_{\mathbf{L}^\infty([0,T])} < \varepsilon$$

for all ν sufficiently large. We thus obtain a contradiction with (2.52), concluding the proof. □

3. Robustness of patchy feedbacks. In this section we apply the previous results on patchy vector fields with impulsive perturbations and construct (discontinuous) stabilizing feedback controls that enjoy robustness properties in the presence of measurement errors and external disturbances. Consider the nonlinear control system on $\mathbb{R}^n$

$$\dot{x} = f(x,u), \qquad\qquad u(t) \in \mathcal{K}, \tag{3.1}$$

assuming that the control set $\mathcal{K} \subset \mathbb{R}^m$ is compact and that the map $f : \mathbb{R}^n \times \mathbb{R}^m \mapsto \mathbb{R}^n$ is smooth. We seek a feedback control $u = U(x) \in \mathcal{K}$ that stabilizes the trajectories of the closed-loop system

$$\dot{x} = f\big(x,\, U(x)\big) \tag{3.2}$$

at the origin. It is well known that even if every initial state $\overline{x} \in \mathbb{R}^n$ can be steered to the origin by an open-loop control $u = u^{\overline{x}}(t)$, a topological obstruction can prevent the existence of a continuous feedback control $u = U(x)$ which (locally) stabilizes the

system (3.1). This fact was first pointed out by Sussmann [Su] for a two-dimensional system ($n = 2,\ \mathcal{K} = \mathbb{R}^2$) and by Sontag and Sussmann [SS] for one-dimensional systems ($n = 1,\ \mathcal{K} = \mathbb{R}$). For general nonlinear systems, it was further analyzed by Brockett [Bro] and Coron [Cor1], [Cor2]. It is thus natural to look for a stabilizing control within a class of discontinuous functions. However, this leads to a theoretical difficulty, because when the function U is discontinuous, the differential equation (3.2) may not admit any Carathéodory solution. To cope with this problem, two different approaches have been pursued.

1. An algorithm is defined which constructs approximate trajectories in connection with an arbitrary (discontinuous) feedback control function. For example, one can sample the feedback control at a discrete set of times. The resulting trajectory, called a sampling solution, was first studied by Krasovskii and Subbotin in the context of positional differential games (see [KS]). In this case, one is not concerned with the existence of exact solutions but only in the asymptotic stabilization properties of all approximate solutions.

2. Alternatively, by the asymptotic controllability to the origin of system (3.1) by means of open-loop controls, one proves the existence of a stabilizing feedback $u = U(x)$ having only a particular type of discontinuities. This feedback thus generates a *patchy vector field*, and the corresponding system (3.2) always admits Carathéodory solutions.

The first approach was initiated in [CLSS] and further developed in [Ri1], [Ri2], [Ri3], [CLRS], [So2]. The second was introduced in [A-B], defining the following class of piecewise constant feedback controls:

DEFINITION 3.1. *Let* $\big(\Omega,\ g,\ (\Omega_\alpha,\ g_\alpha)_{\alpha\in\mathcal{A}}\big)$ *be a patchy vector field. Assume that there exist control values* $k_\alpha \in \mathcal{K}$ *such that for each* $\alpha \in \mathcal{A}$, *there holds*

$$g_\alpha(x) \doteq f(x,\ k_\alpha) \qquad \text{for all } x \in D_\alpha \doteq \Omega_\alpha \setminus \bigcup_{\beta>\alpha} \Omega_\beta. \tag{3.3}$$

Then the piecewise constant map

$$U(x) \doteq k_\alpha \qquad \text{if} \qquad x \in D_\alpha \tag{3.4}$$

is called a patchy feedback control on Ω *and referred to as* $\big(\Omega,\ U,\ (\Omega_\alpha,\ k_\alpha)_{\alpha\in\mathcal{A}}\big)$.

Remark 3.1. From Definitions 1.2 and 3.1, it is clear that the field

$$g(x) = f\big(x,\ U(x)\big)$$

defined in connection with a given patchy feedback $\big(\Omega,\ U,\ (\Omega_\alpha,\ k_\alpha)_{\alpha\in\mathcal{A}}\big)$ is precisely the patchy vector field $\big(\Omega,\ g,\ (\Omega_\alpha,\ g_\alpha)_{\alpha\in\mathcal{A}}\big)$ associated with a family of fields $\big\{g_\alpha : \alpha \in \mathcal{A}\big\}$ satisfying (1.1) Clearly, the patches $(\Omega_\alpha,\ g_\alpha)$ are not uniquely determined by the patchy feedback U. Indeed, whenever $\alpha < \beta$, by (3.3) the values of g_α on the set $\Omega_\alpha \setminus \Omega_\beta$ are irrelevant. Moreover, recalling the notation (1.3) we have

$$U(x) = k_{\alpha^*(x)} \qquad \text{for all } \ x \in \Omega. \tag{3.5}$$

Here, we address the issue of robustness of a stabilizing feedback law $u = U(x)$ w.r.t. small internal and external perturbations

$$\dot{x} = f\big(x,\ U(x + \zeta(t))\big) + d(t), \tag{3.6}$$

where $\zeta = \zeta(t)$ represents a state measurement error and $d = d(t)$ represents an external disturbance of the system dynamics (3.2). Since we are dealing with a discontinuous ODE, one cannot expect the full robustness of the feedback $U(x)$ w.r.t. measurement errors because of possible chattering behavior that may arise at discontinuity points (see [He1], [Ry], [So1], [L-S1], [L-S2], [CR]). Therefore, we shall consider state measurement errors which are small in BV norm, avoiding such phenomena.

Before stating our main result in this direction, we recall here some basic definitions and Proposition 4.2 in [A-B]. This provides the semiglobal practical stabilization (steering all states from a given compact set of initial data into a prescribed neighborhood of zero) of an asymptotically controllable system by means of a patchy feedback control which is robust w.r.t. external disturbances. We consider as (open-loop) *admissible controls* all the measurable functions $u : [0, \infty) \to \mathbb{R}^m$ such that $u(t) \in \mathcal{K}$ for a.e. $t \geq 0$.

DEFINITION 3.2. *The system* (3.1) *is globally asymptotically controllable to the origin if the following holds.*

1. Attractiveness. *For each* $\overline{x} \in \mathbb{R}^n$ *there exists some admissible (open-loop) control* $u = u^{\overline{x}}(t)$ *such that the corresponding trajectory of*

$$\dot{x}(t) = f\big(x(t),\, u^{\overline{x}}(t)\big), \qquad x(0) = \overline{x}, \tag{3.7}$$

either reaches the origin in finite time or tends to the origin as $t \to \infty$.

2. Lyapunov stability. *For each* $\varepsilon > 0$ *there exists* $\delta > 0$ *such that the following holds. For every* $\overline{x} \in \mathbb{R}^n$ *with* $|\overline{x}| < \delta$, *there is an admissible control* $u^{\overline{x}}$ *as in* 1 *steering the system from* $\overline{x}$ *to the origin, so that the corresponding trajectory of* (3.7) *satisfies* $|x(t)| < \varepsilon$ *for all* $t \geq 0$.

PROPOSITION 3.3 (see [A-B, Proposition 4.1]). *Let system* (3.1) *be globally asymptotically controllable to the origin. Then, for every* $0 < r < s$, *one can find* $T > 0$, $\chi > 0$, *and a patchy feedback control* $U : D \mapsto \mathcal{K}$ *defined on some domain*

$$D \supset \big\{x \in \mathbb{R}^n \ ;\ r \leq |x| \leq s\big\} \tag{3.8}$$

so that the following holds. For any measurable map $d : [0, T] \mapsto \mathbb{R}^n$ *such that*

$$\big\|d\big\|_{\mathbf{L}^\infty([0,\,T])} \leq \chi,$$

and for any initial state x_0 *with* $r \leq |x_0| \leq s$, *the perturbed system*

$$\dot{x} = f\big(x,\, U(x)\big) + d(t) \tag{3.9}$$

admits a (forward) Carathéodory trajectory starting from x_0. *Moreover, for any such trajectory* $t \mapsto \gamma(t)$, $t \geq 0$, *one has*

$$\gamma(t) \in D \qquad \text{for all} \quad t \geq 0, \tag{3.10}$$

and there exists $\bar{t}_\gamma < T$ *such that*

$$\big|\gamma(\bar{t}_\gamma)\big| < r. \tag{3.11}$$

Relying on Corollary 1.4 of Theorem 1.3 and on Proposition 3.3, we obtain here the following result concerning robustness of a stabilizing feedback w.r.t. both internal and external perturbations.

THEOREM 3.4. *Let system* (3.1) *be globally asymptotically controllable to the origin. Then, for every* $0 < r < s$, *one can find* $T' > 0$, $\chi' > 0$, *and a patchy*

feedback control $U' : D' \mapsto \mathcal{K}$ defined on some domain D' satisfying (3.8) *so that the following holds. Given any pair of maps $\zeta \in BV([0,\, T'])$, $d \in \mathbf{L}^\infty([0,\, T'])$ such that*

$$\text{Tot.Var.}\{\zeta\} \leq \chi', \qquad \big\|d\big\|_{\mathbf{L}^\infty([0,\, T'])} \leq \chi', \tag{3.12}$$

and any initial state x_0 with $r \leq |x_0| \leq s$, for every solution $t \mapsto x(t)$, $t \geq 0$, of the perturbed system (3.6) *starting from x_0, one has*

$$x(t) \in D' \qquad \text{for all} \quad t \in [0,\, T'], \tag{3.13}$$

and there exists $\bar{t}_x < T'$ such that

$$\big|x(\bar{t}_x)\big| < r. \tag{3.14}$$

Proof. 1. Fix $0 < r < s$. Then, according to Proposition 3.3, we can find $T' > 0$ and a patchy feedback control $U' : D' \mapsto \mathcal{K}$ defined on some domain

$$D' \supset \big\{x \in \mathbb{R}^n \,;\; r/3 \leq |x| \leq s\big\} \tag{3.15}$$

so that the following holds. For every Carathéodory solution $t \mapsto x(t)$, $t \geq 0$, of the unperturbed system (3.2) (with $U = U'$) starting from a point x_0 in the compact set

$$K \doteq \big\{x \in \mathbb{R}^n \,;\; r \leq |x| \leq s\big\}, \tag{3.16}$$

one has

$$x(t) \in D_\rho \doteq \big\{x \in D' \;:\; d\big(x,\, \partial D'\big) > \rho\big\} \qquad \text{for all} \quad t \geq 0 \tag{3.17}$$

for some constant $\rho > 0$. Moreover, there exists $\bar{t}_x < T'$ such that

$$\big|x(\bar{t}_x)\big| < \frac{r}{3}. \tag{3.18}$$

According with Definition 3.1, the field

$$g(x) \doteq f\big(x,\, U'(x)\big) \tag{3.19}$$

is a patchy vector field associated to the family of fields $\big\{g_\alpha : \alpha \in \mathcal{A}\big\}$ defined as in (3.3). The smoothness of f guarantees that for BV perturbations $w = w(t)$ having some uniform bound Tot.Var.$\{w\} \leq \widehat{\chi}$ on the total variation, every (left continuous) solution $y : [0, T'] \mapsto \mathbb{R}^2$ of the impulsive equation (1.7), starting at a point $x_0 \in K$, takes values in the closed set

$$A \doteq B(D_\rho,\, \rho/2). \tag{3.20}$$

Therefore, thanks to Corollary 1.4 of Theorem 1.3, there exists some constant

$$0 < \widehat{\chi}' = \widehat{\chi}'(A,\, K,\, T',\, r/3) < \widehat{\chi} \tag{3.21}$$

such that the following holds. If $y : [0, T'] \mapsto \mathbb{R}^2$ is a (left continuous) solution of the impulsive equation (1.7), with $y(0) \in K$ and Tot.Var.$(w) < \widehat{\chi}'$, then one has

$$y(t) \in A \qquad \text{for all} \quad t \in [0,\, T'], \tag{3.22}$$

and there exists $\bar{t}_y < T'$ such that

$$\big|y(\bar{t}_y)\big| < \frac{2r}{3}. \tag{3.23}$$

2. In connection with the patchy feedback U' introduced above, define the map

$$h(y,z) \doteq f\big(y-z,\, U'(y)\big) - f\big(y,\, U'(y)\big) \tag{3.24}$$

and observe that, by the smoothness of f, there will be some constant $\bar{c} > 0$ such that

$$\big|h(y,z)\big| \leq \bar{c}\cdot|z| \qquad \text{for all } \ y\in A, \quad |z|\leq \widehat{\chi}'. \tag{3.25}$$

Consider now a pair of maps $\zeta \in BV([0,\,T'])$, $d\in \mathbf{L}^\infty([0,\,T'])$ satisfying (3.12) with

$$\chi' < \min\left\{\frac{\widehat{\chi}'}{2(1+T'\bar{c}')},\ \frac{r}{3},\ \frac{\rho}{2}\right\}, \tag{3.26}$$

and let $x = x(t)$ be any Carathéodory solution of the perturbed system (3.6), with an initial condition $x(0) = x_0 \in K$. Then, as observed in the introduction, the map

$$t\mapsto y(t) \doteq x(t) + \zeta(t) \tag{3.27}$$

satisfies the impulsive equation (1.7), where

$$w(t) \doteq \zeta(t) + \int_0^t \big(h\big(y(s),\zeta(s)\big) + d(s)\big)\, ds. \tag{3.28}$$

But then, since (3.12), (3.25), (3.26) together imply

$$\begin{aligned}\text{Tot.Var.}\big\{w\ ;\ [0,\,T']\big\} &\leq \text{Tot.Var.}\big\{\zeta\ ;\ [0,\,T']\big\} + T'\bar{c}\cdot\big\|\zeta\big\|_{\mathbf{L}^\infty([0,\,T'])} + T'\cdot\big\|d\big\|_{\mathbf{L}^\infty([0,\,T'])}\\ &\leq \big(1+T'\bar{c}\big)\cdot\text{Tot.Var.}\big\{\zeta\ ;\ [0,\,T']\big\} + \big\|d\big\|_{\mathbf{L}^\infty([0,\,T'])}\\ &< \widehat{\chi}',\end{aligned}$$

from (3.22)–(3.23) and (3.12), (3.20), (3.26), (3.27) it follows that

$$x(t)\in B(A,\,\rho/2)\subset D' \qquad \text{for all } \ t\in[0,\,T'], \tag{3.29}$$

$$\big|x(\bar{t}_y)\big| < r \qquad \text{for some} \qquad \bar{t}_y < T',$$

which completes the proof of the theorem, taking χ' as in (3.26). □

Remark 3.2. For discontinuous stabilizing feedbacks constructed in terms of sampling solutions, an alternative concept of robustness was introduced in [CLRS], [So1], [So2]. In this case, one considers a partition of the time interval and applies a constant control between two consecutive sampling times. To preserve stability, the measurement error should be sufficiently small compared to the maximum step size. Moreover, each step size should be big enough to prevent possible chattering phenomena. The next result shows that the feedback provided by [A-B, Proposition 4.2] also enjoys this type of robustness. Before stating this result we describe now the concept of a sampling trajectory associated to the perturbed system (3.6) that was introduced in [CLRS], [So2].

Let an initial condition x_0 and a partition $\pi = \{0 = \tau_0 < \tau_1 < \cdots < \tau_{m+1} = T\}$ of the interval $[0, T]$ be given. A sampling trajectory x_π of the perturbed system (3.6), corresponding to a set of measurement errors $\{e_i\}_{i=0}^m$ and an external disturbance $d \in \mathbf{L}^\infty([0, T])$, is defined in a step-by-step fashion as follows. Between τ_0 and τ_1, let $x_\pi(\cdot)$ be a Carathéodory solution of

$$\dot{x} = f\big(x,\ U(x_0 + e_0)\big) + d(t), \qquad t \in [\tau_0,\, \tau_1], \tag{3.30}$$

with initial condition $x_\pi(0) = x_0$. Then, $x_\pi(\cdot)$ is recursively obtained by solving the system

$$\dot{x} = f\big(x,\ U(x_\pi(\tau_i) + e_i)\big) + d(t), \qquad t \in [\tau_i,\, \tau_{i+1}], \quad i > 0. \tag{3.31}$$

The sequence $\{x_\pi(\tau_i) + e_i\}_{i=0}^m$ corresponds to the nonexact measurements used to select control values.

THEOREM 3.5. *Let system* (3.1) *be globally asymptotically controllable to the origin. Then, for every* $0 < r < s$, *one can find* $T'' > 0$, $\chi'' > 0$, $\overline{\delta} > 0$, $\overline{k} > 0$, *and a patchy feedback control* $U'' : D'' \mapsto \mathcal{K}$ *defined on some domain* D'' *satisfying* (3.8) *so that the following holds. Given an initial state* x_0 *with* $r \leq |x_0| \leq s$, *a partition* $\pi = \{\tau_0 = 0,\, \tau_1,\, \ldots,\, \tau_{m+1} = T''\}$ *of the interval* $[0,\, T'']$ *having the property*

$$\frac{\delta}{2} \leq \tau_{i+1} - \tau_i \leq \delta \qquad \textit{for all } \ i \qquad \textit{for some} \qquad \delta \in\ ]0,\, \overline{\delta}\,], \tag{3.32}$$

a set of measurement errors $\{e_i\}_{i=0}^m$, *and an external disturbance* $d \in \mathbf{L}^\infty([0,\, T''])$ *that satisfy*

$$\max_i |e_i| \leq \overline{k} \cdot \delta, \tag{3.33}$$

$$\big\|d\big\|_{\mathbf{L}^\infty} \leq \chi'', \tag{3.34}$$

the resulting sampling solution $x_\pi(\cdot)$ *starting from* x_0 *has the property*

$$x_\pi(t) \in D'' \qquad \textit{for all } \ t \in [0,\, T'']. \tag{3.35}$$

Moreover, there exists $\overline{t}_{x_\pi} < T''$ *such that*

$$\big|x_\pi(\overline{t}_{x_\pi})\big| < r. \tag{3.36}$$

Proof. 1. Fix $0 < r < s$. Then, according with Proposition 3.4, we can find $T' > 0, \chi' > 0$, and a patchy feedback control $U'' : D'' \longmapsto \mathcal{K}$ defined on a domain

$$D'' \supset \big\{x \in \mathbb{R}^n\ ;\ r/3 \leq |x| \leq 2s\big\}$$

so that the following holds. For every external disturbance $d \in \mathbf{L}^\infty$ satisfying (3.34) with $\chi'' \leq \chi'$, and for any Carathéodory solution $t \mapsto x(t)$, $t \geq 0$, of the perturbed system (3.9) (with $U = U''$), starting from a point x_0 with $r \leq |x_0| \leq s$, one has

$$x(t) \in D_{\rho_1} \doteq \big\{x \in D''\ :\ d\big(x,\, \partial D''\big) > \rho_1\big\} \qquad \text{for all } \ t \geq 0 \tag{3.37}$$

for some constant $\rho_1 > 0$. Moreover, there exists $\overline{t}_x < T'$ such that

$$\big|x(\overline{t}_x)\big| < \frac{r}{3}. \tag{3.38}$$

Let

$$\big\{(\Omega_\alpha,\ g_\alpha)\ :\ \alpha = 1, \dots, N\big\}, \qquad g_\alpha(x) = f(x, k_\alpha), \qquad k_\alpha \in \mathcal{K}, \tag{3.39}$$

be the collection of patches associated with the patchy vector field

$$g(x) = f\big(x,\, U''(x)\big). \tag{3.40}$$

We may assume that every vector field g_α is defined on a neighborhood $B(\Omega_\alpha,\, \rho_2)$, $0 < \rho_2 \leq \rho_1$, of the domain Ω_α so that, setting

$$\Omega_\alpha^\rho \doteq \big\{x \in \Omega_\alpha\ ;\ d\big(x,\ \partial\Omega_\alpha\big) > \rho\big\}, \tag{3.41}$$

one has

$$\Omega_\alpha^{\rho_2} \neq \emptyset,$$

and that every g_α is uniformly nonzero on the domain D_α defined in (3.3). Moreover, thanks to the inward-pointing condition (1.1), we may choose the constants $0 < \rho_2 < r/3$ and $\chi'' \leq \chi'$ so that the following hold:

$$\big|g_\alpha(x)\big| \geq 2\chi'' \qquad \text{for all}\ \ x \in B(D_\alpha,\, \rho_2) \tag{3.42}$$

and

$$\big\langle g_\alpha(x) + v,\ \mathbf{n}(x)\big\rangle < 0 \qquad \text{for all}\ \ x \in B(\partial\Omega_\alpha,\, \rho_2), \qquad |v| \leq \chi''. \tag{3.43}$$

For every $d \in \mathbf{L}^\infty$, we denote by $t \mapsto x^\alpha\big(t;\, t_0,\, x_0,\ d\big)$ the solution of the Cauchy problem

$$\dot{x} = g_\alpha(x) + d(t), \qquad x(t_0) = x_0, \tag{3.44}$$

and let $[t_0,\, t^{\max}]$ be the domain of definition of the maximal (forward) solution of (3.44) that is contained in $B(D_\alpha,\, \rho_2)$.

Observe that since every Carathéodory solution of the perturbed system (3.9) (with $U = U''$), starting from a point $x_0 \in B(0, s)\backslash\circ \to B(0,\, r)$, reaches the interior of the ball $B(0,\, r/3)$ in finite time, and because of (3.42), for any $\alpha = 1, \dots, N$ one can find $T_\alpha > 0$ with the following property.

(P)$_1$ For every $x_0 \in B(D_\alpha,\, \rho/2)$, $0 < \rho < \rho_2$, and for any $d \in \mathbf{L}^\infty$ satisfying (3.34), there exists some time $t_\rho \doteq t_\rho(x_0,\, d) < T_\alpha$ such that either one has

$$\big|x^\alpha\big(t_0 + t_\rho;\, t_0,\, x_0,\, d\big)\big| < \frac{2r}{3} \tag{3.45}$$

or else the following holds:

$$x^\alpha\big(t;\, t_0,\, x_0,\, d\big) \in B(D_\alpha,\, \rho_2) \setminus B(D_\alpha,\, \rho) \qquad \text{for all}\ \ t \in [t_0 + t_\rho,\ t^{\max}]. \tag{3.46}$$

On the other hand, relying on the inward-pointing condition (3.43), we deduce two further properties of the solutions of (3.44).

(P)$_2$ The sets Ω_α^ρ, $0 < \rho \leq \rho_2$, defined in (3.41) are positive invariant regions for trajectories of (3.44), i.e., for every $x_0 \in \Omega_\alpha^\rho$, and for any $d \in \mathbf{L}^\infty$ satisfying (3.34), one has

$$x^\alpha\big(t;\, t_0,\, x_0,\, d\big) \in \Omega_\alpha^\rho \qquad \text{for all}\ \ t \geq t_0. \tag{3.47}$$

(P)$_3$ There exists some constant $\overline{c} > 0$ so that, for every $x_0 \in B(\Omega_\alpha, \rho)$, $0 < \rho \leq \rho_2$, such that $d\big(x_0,\ \partial\Omega_\alpha\big) \leq \rho$, and for any $d \in \mathbf{L}^\infty$ satisfying (3.34), one has

$$x^\alpha\big(t;\, t_0,\, x_0,\, d\big) \in \Omega_\alpha^{2\rho} \qquad \text{for all} \quad t \geq t_0 + \overline{c}\cdot\rho\,. \tag{3.48}$$

2. Consider an initial state $x_0 \in B(0,s)\backslash\circ \to B\big(0,\, r\big)$ and a partition $\pi = \{\tau_i\}_{i\geq 0}$ of $[0,\, \infty[$ having the property (3.32), with

$$0 < \delta \leq \overline{\delta} \doteq \min\Big\{\overline{c}\cdot\rho_2,\ \frac{\rho_1}{M}\Big\}, \tag{3.49}$$

$$M \doteq \sup\big\{|g_\alpha(x)| \ :\ x \in B(\Omega_\alpha,\, \rho_2), \quad \alpha = 1, \ldots, N\big\}.$$

Let $x_\pi : [0,\, \infty[\ \mapsto \mathbb{R}^n$ be a sampling solution starting from x_0 and corresponding to a set of measurement errors $\{e_i\}_{i=0}^m$ and to an external disturbance $d(\cdot) \in \mathbf{L}^\infty$ that satisfy (3.33)–(3.34) with

$$\overline{k} \doteq \frac{1}{2\overline{c}}. \tag{3.50}$$

We will first show the following.

LEMMA 3.6. *The map*

$$i \longmapsto \alpha^*(\tau_i) \doteq \alpha^*\big(x_\pi(\tau_i) + e_i\big), \qquad i \geq 0, \tag{3.51}$$

is nondecreasing.

Indeed, assume that $\alpha^*(\tau_i) = \widehat{\alpha}$, which by definitions (1.3), (3.3), (3.5) implies

$$x_\pi(\tau_i) + e_i \in D_{\widehat{\alpha}}, \tag{3.52}$$

$$x_\pi(\tau_{i+1}) = x^{\widehat{\alpha}}\big(\tau_{i+1};\ \tau_i,\, x_\pi(\tau_i),\, d\restriction_{[\tau_i,\,\tau_{i+1}]}\big). \tag{3.53}$$

Then, because of (3.33), (3.49)–(3.50), one has

$$x_i \doteq x_\pi(\tau_i) \in B(D_{\widehat{\alpha}},\ \overline{k}\delta) \subset B(\Omega_{\widehat{\alpha}},\ \rho_2). \tag{3.54}$$

We shall consider separately the case in which

$$x_i \in D_{\widehat{\alpha}}^{\overline{k}\delta} \subset \Omega_{\widehat{\alpha}}^{\overline{k}\delta}, \qquad \overline{k}\delta \leq \rho_2, \tag{3.55}$$

and the case where

$$x_i \in B(D_{\widehat{\alpha}},\ \overline{k}\delta), \qquad d\big(x_i\,,\, \partial\Omega_{\widehat{\alpha}}\big) \leq \overline{k}\delta \leq \rho_2. \tag{3.56}$$

In the first case, using (3.53) and applying (P)$_2$ we deduce that $x_\pi(\tau_{i+1}) \in \Omega_{\widehat{\alpha}}^{\overline{k}\delta}$, which, in turn, because of (3.33), (3.49)–(3.50), implies

$$x_\pi(\tau_{i+1}) + e_{i+1} \in \Omega_{\widehat{\alpha}}. \tag{3.57}$$

From (3.57), by definition (1.3) we derive

$$\alpha^*(\tau_{i+1}) \geq \widehat{\alpha}, \tag{3.58}$$

proving the lemma whenever (3.55) holds. On the other hand, when (3.56) is verified, since by (3.32), (3.50) one has

$$\tau_{i+1} - \tau_i \geq \frac{\delta}{2} = \overline{c}\overline{k} \cdot \delta,$$

applying $(\mathrm{P})_3$, we deduce $x_\pi(\tau_{i+1}) \in \Omega^{2\overline{k}\delta}_{\widehat{\alpha}}$. This again implies (3.57)–(3.58), completing the proof of Lemma 3.6.

Next, relying on $(\mathrm{P})_1$ and setting

$$\begin{aligned} (3.59)\quad i'_\alpha &\doteq \min\big\{\, i \geq 0 \;\;;\;\; \alpha^*(\tau_i) = \alpha, \quad x_\pi(\tau_i) \notin B(0,\, 2r/3)\,\big\}, \\ i''_\alpha &\doteq \max\big\{\, i \geq 0 \;\;;\;\; \alpha^*(\tau_i) = \alpha\,, \quad x_\pi(\tau_i) \notin B(0,\, 2r/3)\,\big\}, \qquad \alpha \in \mathrm{Range}(\alpha^*), \end{aligned}$$

we deduce

$$\tau_{i''_\alpha} - \tau_{i'_\alpha} \leq T_\alpha \qquad \text{for all} \;\; \alpha \in \mathrm{Range}(\alpha^*). \tag{3.60}$$

Indeed, if (3.60) does not hold, by definitions (3.3), (3.5) one has

$$x_{i'_\alpha} \doteq x_\pi(\tau_{i'_\alpha}) \in B(D_\alpha,\, \overline{k}\delta) \subset B(\Omega_\alpha,\, \rho_2/2), \tag{3.61}$$

$$x_\pi(t) = x^\alpha\big(t;\; \tau_{i'_\alpha}\,,\, x_{i'_\alpha},\, d\restriction_{[\tau_{i'_\alpha},\, \tau_{i''_\alpha+1}]}\big) \qquad \text{for all} \;\; t \in [\tau_{i'_\alpha}\,,\; \tau_{i''_\alpha+1}]. \tag{3.62}$$

But then, applying $(\mathrm{P})_1$, one could find some $\widehat{i} \leq i''_\alpha$ such that

$$x_\pi(t) \in B(D_\alpha,\, \rho_2) \setminus B(D_\alpha,\, 2\overline{k}\delta) \qquad \text{for all} \;\; t \in [\tau_{\widehat{i}},\, \tau_{i''_\alpha+1}].$$

By definitions (1.3), (3.51) and because of (3.33), this implies

$$\alpha^*(\tau_i) > \alpha \qquad \text{for all } \widehat{i} \leq i \leq i''_\alpha,$$

providing a contradiction with (3.59).

To conclude the proof of Theorem 3.5, we observe that the monotonicity of the map (3.51), together with the estimate (3.60), implies that there exists some time $\overline{t}_{x_\pi} < T'' \doteq \sum_{\alpha=1}^N T_\alpha$ such that (3.36) is verified. Moreover, (3.35) clearly follows from (3.37) and (3.49). □

Remark 3.3. Consider a partition $\pi = \{\tau_0 = 0,\, \tau_1,\, \ldots,\, \tau_{m+1} = T\}$ of the interval $[0,\, T\,]$ having the property (3.32). If we associate to a set of measurement errors $\{e_i\}_{i=1}^m$ satisfying (3.33) the piecewise constant function $\zeta \;: [0,\, T] \mapsto \mathbb{R}^n$ defined as

$$\zeta(t) = e_i \qquad \text{for all } t \in \;]\tau_i,\, \tau_{i+1}],$$

then

$$\mathrm{Tot.Var.}\{\zeta\} \leq 4\overline{k} \cdot T.$$

Thus, taking the constant $\overline{k}$ sufficiently small we may reinterpret the *discrete* internal disturbance allowed for a sampling solution in Theorem 3 as a particular case of the measurement errors with small total variation considered in Theorem 3.4.

REFERENCES

[A-B] F. Ancona and A. Bressan, *Patchy vector fields and asymptotic stabilization,* ESAIM Control Optim. Calc. Var., 4 (1999), pp. 445–471.

[B-W] J. Behrens and F. Wirth, *A globalization procedure for locally stabilizing controllers,* in Nonlinear Control in the Year 2000, Lecture Notes in Control and Inform. Sci. 258, A. Isidori, F. Lamnabi-Lagarrige, and W. Respondek, eds., Springer-Verlag, London, 2001, pp. 171–184.

[B1] A. Bressan, *On differential systems with impulsive controls,* Rend. Sem. Mat. Univ. Padova, 78 (1987), pp. 227–235.

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

[CLRS] F.H. Clarke, Yu.S. Ledyaev, L. Rifford, and R.J. Stern, *Feedback stabilization and Lyapunov functions,* SIAM J. Control Optim., 39 (2000), pp. 25–48.

[CLSS] F.H. Clarke, Yu.S. Ledyaev, E.D. Sontag, and A.I. Subbotin, *Asymptotic controllability implies feedback stabilization,* IEEE Trans. Automat. Control, 42 (1997), pp. 1394–1407.

[Cor1] J.-M. Coron, *A necessary condition for feedback stabilization,* Systems Control Lett., 14 (1998), pp. 227–232.

[Cor2] J.-M. Coron, *On the stabilization in finite time of locally controllable systems by means of continuous time-varying feedback law,* SIAM J. Control Optim., 33 (1995), pp. 804–833.

[CR] J.-M. Coron and L. Rosier, *A relation between continuous time-varying and discontinuous feedback stabilization,* J. Math. Systems Estim. Control, 4 (1994), pp. 67–84.

[He1] H. Hermes, *Discontinuous vector fields and feedback control,* in Differential Equations and Dynamical Systems, J.K. Hale and J.P. La Salle, eds., Academic Press, New York, 1967, pp. 155–165.

[KS] N.N. Krasovskii and A.I. Subbotin, *Game-Theoretical Control Problems,* Springer-Verlag, New York, 1988.

[L-S1] Yu.S. Ledyaev and E.D. Sontag, *A remark on robust stabilization of general asymptotically controllable systems,* in Proceedings of the Conference on Information Sciences and Systems (CISS 97), Johns Hopkins University Press, Baltimore, MD, 1997 pp. 246–251.

[L-S2] Yu.S. Ledyaev and E.D. Sontag, *A Lyapunov characterization of robust stabilization,* J. Nonlinear Anal., 37 (1999), pp. 813–840.

[Ri1] L. Rifford, *Existence of Lipschitz and semiconcave control-Lyapunov functions,* SIAM J. Control Optim., 39 (2000), pp. 1043–1064.

[Ri2] L. Rifford, *Semiconcave control-Lyapunov functions and stabilizing feedbacks,* SIAM J. Control Optim., 41 (2002), pp. 659–681.

[Ri3] L. Rifford, *On the existence of nonsmooth control-Lyapunov functions in the sense of generalized gradient,* ESAIM Control Optim. Calc. Var., 6 (2001), pp. 593–611.

[Ry] E.P. Ryan, *On Brockett's condition for smooth stabilizability and its necessity in a context of nonsmooth feedback,* SIAM J. Control Optim., 32 (1994), pp. 1597–1604.

[So1] E.D. Sontag, *Stability and stabilization: Discontinuities and the effect of disturbances,* in Nonlinear Analysis, Differential Equations, and Control, NATO Sci. Ser. C Math. Phys. Sci. 528, F.H. Clarke and R.J. Stern, eds., Kluwer Academic, Dordrecht, The Netherlands, 1999, pp. 551–598.

[So2] E.D. Sontag, *Clocks and insensitivity to small measurement errors,* ESAIM Control Optim. Calc. Var., 4 (1999), pp. 537–557.

[SS] E.D. Sontag and H.J. Sussmann, *Remarks on continuous feedback,* in Proceedings of the IEEE Conference on Decision and Control, IEEE, Piscataway, NJ, 1980, pp. 916–921.

[Su] H.J. Sussmann, *Subanalytic sets and feedback control,* J. Differential Equations, 31 (1979), pp. 31–52.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 5, pp. 1477–1497

ACCURACY AND CONVERGENCE PROPERTIES OF THE FINITE DIFFERENCE MULTIGRID SOLUTION OF AN OPTIMAL CONTROL OPTIMALITY SYSTEM*

ALFIO BORZÌ†, KARL KUNISCH†, AND DO Y. KWAK‡

Abstract. The finite difference multigrid solution of an optimal control problem associated with an elliptic equation is considered. Stability of the finite difference optimality system and optimal-order error estimates in the discrete L^2 norm and in the discrete H^1 norm under minimum smoothness requirements on the exact solution are proved. Sharp convergence factor estimates of the two grid method for the optimality system are obtained by means of local Fourier analysis. A multigrid convergence theory is provided which guarantees convergence of the multigrid process towards weak solutions of the optimality system.

Key words. optimal control problem, Poisson equation, finite differences, accuracy estimate, convergence theory, multigrid method

AMS subject classifications. 49K20, 65N06, 65N12, 65N55

PII. S0363012901393432

1. Introduction.

Optimal control problems involving partial differential equations [17, 18] are nowadays receiving much attention because of their importance in the industrial design process. Especially, the need for accurate and efficient solution methods for these problems has become an important issue.

We consider a finite difference framework and multigrid methods for the case of distributed optimal control of an elliptic problem and provide for this case optimal estimates for the accuracy of the solution and for the convergence factor of the multigrid process. The present work is characterized by the fact that we extend known analytic tools for scalar elliptic problems to the case of a (nonsymmetric) system of elliptic partial differential equations, called an optimality system.

In our finite difference analysis, based on results stated in [14, 20], we prove stability of the finite difference optimality system and prove optimal-order error estimates in the discrete L^2 norm and in the discrete H^1 norm under minimum smoothness requirements on the analytic solution.

It is known that multigrid methods [5, 13, 21] solve elliptic problems with optimal computational order, i.e., the number of computer operations required scales linearly with respect to the number of unknowns. This fact has been demonstrated in the case of multigrid applied to a singular optimal control problem associated with a nonlinear elliptic equation [2]. In particular, results in [2] show that the convergence properties of the multigrid method do not deteriorate as the weight of the cost of the control tends to zero, demonstrating the robustness of this method.

We prove convergence of the multigrid method applied to the optimality system within two analytic frameworks which have complementary features. We use two grid

*Received by the editors August 7, 2001; accepted for publication (in revised form) June 20, 2002; published electronically January 14, 2003. This research was supported in part by the SFB 03 "Optimization and Control" and by grant 2000-2-10300-001-5 from Basic Research Program of KOSEF.

http://www.siam.org/journals/sicon/41-5/39343.html.

†Institut für Mathematik, Karl-Franzens-Universität Graz, Heinrichstr. 36, A-8010 Graz, Austria (alfio.borzi@uni-graz.at, karl.kunisch@uni-graz.at).

‡Department of Mathematics, Korea Advanced Institute of Science and Technology, Taejon, Korea 305-701 (dykwak@math.kaist.ac.kr).

local Fourier analysis [21, 10] with simplifying assumptions on the boundary conditions to obtain sharp convergence estimates of the multigrid method. These convergence estimates agree very well with results of numerical experiments and appear to be independent of the mesh size and of the value of the control parameter.

While the extension of two grid local Fourier analysis to systems of partial differential equations may be considered straightforward, the extension of the multigrid theory provided in [6, 8, 9, 15] to the case of optimality systems requires additional analysis, which is presented in this paper. The resulting multigrid theory does not require special assumptions on the boundary, it applies to polygonal domains, and guarantees convergence of the multigrid method to weak solutions of the optimality system.

In the following section we introduce and analyze our model problem. The finite difference discretization of this model problem and the corresponding stability and accuracy analysis are considered in section 3. In section 4 we describe the multigrid method and define and analyze its components. Two grid local Fourier analysis is presented in section 5. In section 6, a general convergence theory for multigrid applied to the optimality system is provided.

2. Optimal control problem. We consider the optimal control problem

$$\min J(y,u) = \frac{1}{2}||y-z||^2_{L^2(\Omega)} + \frac{\nu}{2}||u||^2_{L^2(\Omega)}, \tag{2.1}$$

subject to $u \in L^2(\Omega)$ and

$$\begin{aligned} -\Delta y &= u + g \text{ in } \Omega, \\ y &= 0 \text{ on } \partial\Omega, \end{aligned} \tag{2.2}$$

where $\Omega = (0,1)\times(0,1)$, $g \in L^2(\Omega)$, $z \in L^2(\Omega)$ is the objective function, and $\nu > 0$ is the weight of the cost of the control. Existence of a unique solution to (2.1) and its characterization are well known. Let us, for the sake of completeness, give a short derivation and denote by $\hat{J}(u) = J(y(u),u)$, where $y(u)$ denotes the solution of (2.2) as a function of u. Recall that $\Delta : H^1_0(\Omega)\cap H^2(\Omega) \to L^2(\Omega)$ is a homeomorphism. Here we use the fact that Ω is convex. The mapping $u \to y(u)$ from $L^2(\Omega)$ to $H^1_0(\Omega)\cap H^2(\Omega)$ is affine and continuous. Let us denote its first derivative at u in the direction δu by $y'(u,\delta u)$. It is characterized as the solution to

$$\begin{aligned} -\Delta y'(u,\delta u) &= \delta u \text{ in } \Omega, \\ y'(u,\delta u) &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{2.3}$$

The second derivative of $u \to y(u)$ is zero. Hence we find for the second derivative of $u \to \hat{J}(u)$

$$\hat{J}''(u)(\delta u,\delta u) = ||y'(u,\delta u)||^2_{L^2(\Omega)} + \nu||\delta u||^2_{L^2(\Omega)},$$

and thus $u \to \hat{J}(u)$ is uniformly convex. This implies existence of a unique solution u^* to (2.1). Moreover, the solution is characterized by $\hat{J}'(u)(u^*;\delta u) = 0$ for all δu and consequently

$$\hat{J}'(u^*,\delta u) = (y^*-z, y'(u^*,\delta u))_{L^2(\Omega)} + \nu(u^*,\delta u)_{L^2(\Omega)} = 0 \qquad \text{for all } \delta u \in L^2(\Omega),$$

where $y^* = y(u^*)$. Introduce $\lambda^* \in H^1_0(\Omega)\cap H^2(\Omega)$ as the unique solution to

$$\begin{aligned} -\Delta\lambda^* &= -(y^*-z) \text{ in } \Omega, \\ \lambda^* &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{2.4}$$

Then by (2.3) and (2.4) we have

$$\hat{J}'(u^*, \delta u) = -(\lambda^*, \delta u)_{L^2(\Omega)} + \nu(u^*, \delta u)_{L^2(\Omega)} = 0 \qquad \text{for all } \delta u \in L^2(\Omega), \tag{2.5}$$

which constitutes the necessary and sufficient optimality condition for (2.1). In (2.5), λ^* is defined via (2.2) and (2.4).

For later reference let us summarize (2.2), (2.4), and (2.5):

$$\begin{aligned} -\Delta y &= \frac{1}{\nu}\lambda + g \text{ in } \Omega, \\ y &= 0 \text{ on } \partial\Omega, \\ -\Delta\lambda &= -(y - z) \text{ in } \Omega, \\ \lambda &= 0 \text{ on } \partial\Omega, \\ \nu u - \lambda &= 0 \text{ in } \Omega. \end{aligned} \tag{2.6}$$

Here, for convenience, we dropped the $*$-notation. System (2.6) is referred to as the optimality system for (2.1). From the optimality system one concludes the following regularity property.

COROLLARY 2.1. *If z, $g \in L^2(\Omega)$, then $(y^*, u^*, \lambda^*) \in (H_0^1(\Omega) \cap H^2(\Omega))^3$.*

In the following section we address finite difference approximations to (2.1) and (2.6).

3. Finite difference approximation of the optimality system. While finite element approximations to (2.1) are rather well investigated, see [18] and the references given there, much less rigorous analysis is available for finite difference methods. Thus, before addressing multigrid methods in the remainder of the paper, we investigate convergence of finite difference approximations to (2.1). We consider a sequence of grids $\{\Omega_h\}_{h>0}$ defined by

$$\Omega_h = \{\mathbf{x} \in \mathbf{R}^2 : x_i = s_i\, h, \quad s_i \in \mathbb{Z}\} \cap \Omega.$$

Here and below we follow the notation and terminology of [14], especially section 9. To avoid certain technicalities we assume also in this section that Ω is a square and that the values of h are chosen such that the boundaries of Ω coincide with grid lines. The case of general convex domains is addressed in Remark 1 below. The negative Laplacian with homogeneous Dirichlet boundary conditions is approximated by the common five-point stencil as in [14, section 4] and denoted by $-\Delta_h$.

For grid functions v_h and w_h defined on Ω_h we introduce the discrete L^2-scalar product

$$(v_h, w_h)_{L_h^2} = h^2 \sum_{\mathbf{x} \in \Omega_h} v_h(\mathbf{x})\, w_h(\mathbf{x}),$$

with associated norm $|v_h|_0 = (v_h, v_h)_{L_h^2}^{1/2}$. We require as well the discrete H^1-product given by

$$|v_h|_1 = \left(|v_h|_0^2 + \sum_{i=1}^{2} |\partial_i^- v_h|_0^2 \right)^{1/2},$$

where ∂_i^- denotes the backward difference quotient in the x_i direction and v_h is extended by 0 on grid points outside of Ω. The spaces L_h^2 and H_h^1 consist of the sets

of grid functions v_h endowed with $|v_h|_0$, respectively, $|v_h|_1$, as norm. Further denote with M_h the vector space of nodal functions v_h defined on Ω_h which are zero on the boundary. The system of nodal functions (v_h, w_h) is denoted by $\mathcal{M}_h = M_h \times M_h$.

We need the following lemma.

LEMMA 3.1 (Poincaré–Friedrichs inequality for finite differences). *For any grid function $v_h \in M_h$, there exists a constant c_*, independent of v_h and h, such that*

$$|v_h|_0^2 \leq c_* \sum_{i=1}^{2} |\partial_i^- v_h|_0^2, \tag{3.1}$$

where $c_* = \frac{1}{4}$.

Proof. For the proof see [20]. □

Functions in $L^2(\Omega)$ and $H^2(\Omega)$ are approximated by grid functions defined through their mean values with respect to elementary cells $[x_1-\frac{h}{2}, x_1+\frac{h}{2}]\times[x_2-\frac{h}{2}, x_2+\frac{h}{2}]$. This gives rise to the restriction operators $\tilde{R}_h : L^2(\Omega) \to L_h^2$ and $R_h : H_0^1(\Omega)\cap H^2(\Omega) \to L_h^2$ defined in [14, p. 232]. For the definition of H_h^2 we refer to [14], as well. Further, we define $\tilde{R}_h^2 : L^2(\Omega)\times L^2(\Omega) \to L_h^2 \times L_h^2$ by $\tilde{R}_h^2 = (\tilde{R}_h, \tilde{R}_h)$ and analogously $R_h^2 = (R_h, R_h)$.

The discrete optimal control problems are specified next:

$$\begin{cases} \min \frac{1}{2}|y_h - \tilde{R}_h z|_0^2 + \frac{\nu}{2}|u_h|_0^2, \\ \quad -\Delta_h y_h = u_h + \tilde{R}_h g, \qquad u_h \in L_h^2. \end{cases} \tag{3.2}$$

Let u_h^* denote the unique solution to (3.2) and set $y_h^* = y_h(u_h^*)$. The optimality system related to (3.2) is found to be

$$\begin{aligned} -\Delta_h y_h^* &= u_h^* + \tilde{R}_h g, \\ -\Delta_h \lambda_h^* &= -(y_h^* - \tilde{R}_h z), \\ \nu u_h^* - \lambda_h^* &= 0. \end{aligned} \tag{3.3}$$

We can eliminate u_h^* from this system and obtain, dropping the superscript $*$,

$$\begin{cases} -\nu\,\Delta_h y_h - \lambda_h \;=\; \nu\,\tilde{R}_h g, \\ \quad -\Delta_h \lambda_h + y_h \;=\; \tilde{R}_h z. \end{cases} \tag{3.4}$$

To investigate the convergence of the solution of (3.4) to the solution of (2.6) as $h \to 0^+$, we introduce the family of operators

$$\mathcal{A}_h = \begin{pmatrix} -\nu\,\Delta_h & -I_h \\ I_h & -\Delta_h \end{pmatrix}, \tag{3.5}$$

where I_h is the identity operator on grid functions v_h. The operators $\mathcal{A}_h$ are defined between product spaces of grid functions. For us the cases $\mathcal{A}_h : H_h^1 \times H_h^1 \to H_h^{-1} \times H_h^{-1}$ and $\mathcal{A}_h : H_h^2 \times H_h^2 \to L_h^2 \times L_h^2$ are important. Here H_h^{-1} denotes the dual space of H_h^1 with L_h^2 as pivot space.

The family $\{\mathcal{A}_h\}_{h>0}$ is called H_h^1-regular if A_h is invertible and there exists a constant C_1 independent of h such that

$$||\mathcal{A}_h^{-1}||_{\mathcal{L}(H_h^{-1}\times H_h^{-1}, H_h^1 \times H_h^1)} \leq C_1,$$

and analogously it is called H_h^2-regular if

$$||\mathcal{A}_h^{-1}||_{\mathcal{L}(L_h^2\times L_h^2, H_h^2 \times H_h^2)} \leq C_2,$$

for C_2 independent of h.

LEMMA 3.2. *The family of operators $\{\mathcal{A}_h\}_{h>0}$, with h such that the boundaries of Ω are grid lines, is H_h^1-regular.*

Proof. Let $(v_h, w_h) \in \mathcal{M}_h$ be a pair of grid functions. Then

$$\begin{aligned}(\mathcal{A}_h(v_h,w_h),(v_h,w_h))_{L_h^2\times L_h^2} &= \nu(-\Delta_h v_h, v_h)_{L_h^2} + (-\Delta_h w_h, w_h)_{L_h^2}\\ &\geq \min(\nu,1)\, C \sum_{i=1}^{2}(|\partial_i^- v_h|_0^2 + |\partial_i^- w_h|_0^2),\end{aligned} \tag{3.6}$$

where C is independent of h and arises from the coercivity estimate for $-\Delta_h$, i.e.,

$$(-\Delta_h v_h, v_h)_{L_h^2} \geq C \sum_{i=1}^{2} |\partial_i^- v_h|_0^2 \qquad \text{for all } v_h; \tag{3.7}$$

see, e.g., [14, p. 231]. Using Poincaré inequality in (3.6) results in

$$(\mathcal{A}_h(v_h,w_h),(v_h,w_h))_{L_h^2\times L_h^2} \geq C_1^{-2}\,|(v_h,w_h)|^2_{H_h^1\times H_h^1} \qquad \text{for all } (v_h,w_h)\in L_h^2\times L_h^2,$$

with $C_1^{-2} = \min(\nu,1)\, C\, c_0$. Due to the Lax–Milgram lemma $\mathcal{A}_h$ is invertible. Moreover,

$$||\mathcal{A}_h^{-1}||_{\mathcal{L}(H_h^{-1}\times H_h^{-1}, H_h^1\times H_h^1)} \leq C_1 \qquad \text{for all } h. \qquad \square$$

The infinite dimensional analogue of $\mathcal{A}_h$ is the operator

$$\mathcal{A} = \begin{pmatrix} -\nu\,\Delta & -I \\ I & -\Delta \end{pmatrix}, \tag{3.8}$$

where Δ is understood with homogeneous Dirichlet boundary conditions. It is well defined from $H_0^1(\Omega)\times H_0^1(\Omega)$ to $H^{-1}(\Omega)\times H^{-1}(\Omega)$ as well as from $(H^2(\Omega)\cap H_0^1(\Omega))\times (H^2(\Omega)\cap H_0^1(\Omega))$ to $L^2(\Omega)\times L^2(\Omega)$. We have the following consistency result.

LEMMA 3.3. *There exists a constant C_K independent of h such that*

$$||\mathcal{A}_h R_h^2 - \tilde{R}_h^2 \mathcal{A}||_{\mathcal{L}((H^2\cap H_0^1)^2, (H_h^{-1}\times H_h^{-1}))} \leq C_K\, h.$$

Proof. Let $(v,w) \in (H^2(\Omega)\cap H_0^1(\Omega))^2$ and note that, due to the consistency property of $-\Delta_h$ as discretization of $-\Delta$, we have

$$\begin{aligned}&|\mathcal{A}_h R_h^2(v,w) - \tilde{R}_h^2\mathcal{A}(v,w)|^2_{H_h^{-1}\times H_h^{-1}}\\ &\quad\leq \nu\,|(-\Delta_h)R_h v - \tilde{R}_h(-\Delta)v|^2_{H_h^{-1}} + |(-\Delta_h)R_h w - \tilde{R}_h(-\Delta)w|^2_{H_h^{-1}}\\ &\qquad + |R_h v - \tilde{R}_h v|^2_{H_h^{-1}} + |R_h w - \tilde{R}_h w|^2_{H_h^{-1}}\\ &\qquad\leq C_K^2\, h^2\,|(v,w)|_{H^2(\Omega)^2};\end{aligned}$$

see [14, p. 232]. $\square$

THEOREM 3.4. *There exists a constant K_1, depending on Ω, g, z, and independent of h, such that*

$$|y_h^* - R_h y^*|_1 + |u_h^* - R_h u^*|_1 + |\lambda_h^* - R_h\lambda^*|_1 \leq K_1\, h.$$

Proof. From (2.6) and (3.4) we have

$$(y_h^*, \lambda_h^*) - R_h^2\,(y^*, \lambda^*) = \mathcal{A}_h^{-1}(\tilde{R}_h^2 \mathcal{A} - \mathcal{A}_h R_h^2)\,(y^*, \lambda^*). \tag{3.9}$$

Lemmas 3.2 and 3.3 imply the existence of $\bar{K}_1$ such that

$$|y_h^* - R_h y^*|_1 + |\lambda_h^* - R_h \lambda^*|_1 \le \bar{K}_1 h.$$

Using $\nu\, u^* = \lambda^*$ and its discrete analogue, we have the following claim. ∎

Remark 1. In the case of a general convex domain attention must be paid to the discretization of $-\Delta$ along the boundary. The literature offers several options. For the Shortley–Weller discretization, as described in [14, p. 78], $-\Delta_h$ is H_h^1-regular and consistent with $-\Delta$ from $H^2(\Omega)$ to H_h^{-1}. Using these facts the generalization of Theorem 3.4 to convex domains is straightforward.

In the following result the assumption that the boundaries of Ω coincide with grid lines is used.

THEOREM 3.5. *There exists a constant K_2, depending on Ω, g, z, and independent of h, such that*

$$|y_h^* - R_h y^*|_0 + |u_h^* - R_h u^*|_0 + |\lambda_h^* - R_h \lambda^*|_0 \le K_2\, h^2.$$

Proof. We start by showing that $\mathcal{A}_h^T$ is H_h^2-regular. For this purpose it suffices to show the existence of a constant C_2 independent of h such that for all $(f_h, g_h) \in L_h^2 \times L_h^2$

$$|(v_h, w_h)|_{H_h^2 \times H_h^2} \le C_2\, |(f_h, g_h)|_{L_h^2 \times L_h^2}, \tag{3.10}$$

where $\mathcal{A}_h^T(v_h, w_h) = (f_h, g_h)$. Proceeding as in Lemma 3.2 one shows that $\mathcal{A}_h^T$ is H_h^1-regular. In particular, there exists $\tilde{C}_2 \ge 1$, independent of h, such that

$$|(v_h, w_h)|_{H_h^1 \times H_h^1} \le \tilde{C}_2\, |(f_h, g_h)|_{L_h^2 \times L_h^2}. \tag{3.11}$$

Since $-\Delta_h$ is H_h^2-regular [14, p. 242], $\tilde{C}_2$ can also be chosen such that

$$||(-\Delta_h)^{-1}||_{\mathcal{L}(L_h^2, H_h^2)} \le \tilde{C}_2. \tag{3.12}$$

Note that v_h satisfies $\nu\, v_h = (-\Delta_h)^{-1}(f_h - w_h)$. Hence by (3.11) and (3.12)

$$|v_h|_{H_h^2} \le \frac{2}{\nu}\, \tilde{C}_2^2\, |(f_h, g_h)|_{L_h^2 \times L_h^2}. \tag{3.13}$$

Similarly $w_h = (-\Delta_h)^{-1}(g_h + v_h)$ and hence

$$|w_h|_{H_h^2} \le 2\, \tilde{C}_2^2\, |(f_h, g_h)|_{L_h^2 \times L_h^2}. \tag{3.14}$$

Combining (3.13) and (3.14) we have (3.10). From (3.10) it follows by duality that

$$||\mathcal{A}_h^{-1}||_{\mathcal{L}(H_h^{-2} \times H_h^{-2}, L_h^2 \times L_h^2)} \le C_2. \tag{3.15}$$

Turning to consistency, due to the assumption that the boundary of Ω coincides with grid lines, we have

$$||(-\Delta_h) R_h - \tilde{R}_h(-\Delta)||_{\mathcal{L}(H^2, H_h^{-2})} \le K\, h^2, \tag{3.16}$$

$$||R_h - \tilde{R}_h||_{\mathcal{L}(H^2, L_h^2)} \le K\, h^2, \tag{3.17}$$

for a constant K independent of h. Estimate (3.16) is given in [14, p. 239] and (3.17) follows from a direct computation. From (3.9) and (3.16) we have

$$|(y_h^*, \lambda_h^*) - R_h^2 (y^*, \lambda^*)|_{L_h^2 \times L_h^2} \leq C_2 \, |(\tilde{R}_h^2 \mathcal{A} - \mathcal{A}_h R_h^2)(y^*, \lambda^*)|_{H_h^{-2} \times H_h^{-2}}. \tag{3.18}$$

Using (3.16) and (3.17), we proceed as in the proof of Lemma 3.3 to obtain the desired result. □

The approximation results stated in Theorem 3.5 are demonstrated in numerical experiments with global mesh refinement (multigrid); see [2].

4. The multigrid method. Multigrid methods have been extensively used to solve discretized partial differential equations; see, e.g., [21] and the references given there. This fact has motivated intensive research towards the determination of convergence properties of multigrid schemes; see [5, 6, 9, 10, 11, 14, 15]. Multigrid methods have also been used to solve optimal control problems. Most of these contributions except for [12] are rather recent, e.g., [1, 2, 3, 4, 19]. Concerning the convergence theory of multigrid applied to systems of partial differential equations and, in particular, to optimality systems, the theory is far from being complete.

The purpose of the present work is to analyze a multigrid algorithm that solves the optimality system (2.6) with typical multigrid efficiency. We briefly describe the multigrid framework to keep this paper self-contained. Let us index the operators and variables defined on the grid with mesh size $h = h_k = 1/2^k$, $k = 1, \ldots, L$, with the index k, and for simplicity of presentation let us introduce vector notation: we let $\mathbf{w} = (u, v)$ and $|\mathbf{w}|_0 = |(u, v)|_0$, etc.

Consider the discrete problem (3.4) expressed as

$$\mathcal{A}_k \boldsymbol{\phi}_k = \mathbf{f}_k \text{ on } \Omega_k, \tag{4.1}$$

where $\boldsymbol{\phi}_k = (y_k, \lambda_k)$ and $\mathbf{f}_k = (g_k, z_k)$ are defined on the mesh Ω_{h_k}.

For the purpose of multigrid methods it is important to utilize the fact that the solution of (4.1) is equivalent to solve $\mathcal{A}_k \boldsymbol{\phi}_k^e = \mathbf{r}_k$, where $\boldsymbol{\phi}_k^e = \bar{\boldsymbol{\phi}}_k - \boldsymbol{\phi}_k$ is the error grid function between the solution $\bar{\boldsymbol{\phi}}_k$ to (4.1) and its current approximation $\boldsymbol{\phi}_k$, and $\mathbf{r}_k$ is the residual defined by

$$\mathbf{r}_k = \mathbf{f}_k - \mathcal{A}_k \boldsymbol{\phi}_k. \tag{4.2}$$

In fact, the multigrid strategy is to solve for all frequency components of the error using multiple grids.

On the grid of level k, a smoothing procedure is applied in order to solve for the high-frequency components of the error. This is an iterative scheme denoted by $\boldsymbol{\phi}_k^{(m)} = (\mathcal{S}_k)^m (\boldsymbol{\phi}_k, \mathbf{f}_k)$, where $(\mathcal{S}_k)^m$ is a linear smoothing operator applied m times. One sweep of this iteration is written in the form $\boldsymbol{\phi}_k^{(m)} = \boldsymbol{\phi}_k^{(m-1)} + \mathcal{R}_k (\mathbf{f}_k - \mathcal{A}_k \boldsymbol{\phi}_k^{(m-1)})$, where the operator $\mathcal{R}_k$ applies to the residual.

To correct for the smooth components of the error, a coarse grid correction (CGC) is defined. For this purpose a coarse grid problem for the error function is constructed on the grid with mesh size h_{k-1}:

$$\mathcal{A}_{k-1} \boldsymbol{\phi}_{k-1} = \mathcal{I}_k^{k-1} \mathbf{r}_k, \tag{4.3}$$

where $\boldsymbol{\phi}_{k-1}$ aims to represent, on the coarse grid Ω_{k-1}, the error $\boldsymbol{\phi}_k^e$ on the next finer grid. Because of Dirichlet boundary conditions, we have $\boldsymbol{\phi}_{k-1} = 0$ at the boundary.

The operator $\mathcal{I}_k^{k-1} : \mathcal{M}_k \to \mathcal{M}_{k-1}$ restricts the residual computed at level k to the grid with level $k-1$.

Once the coarse grid problem is solved, the CGC follows:

$$\boldsymbol{\phi}_k^{new} = \boldsymbol{\phi}_k + \mathcal{I}_{k-1}^k \, \boldsymbol{\phi}_{k-1}, \tag{4.4}$$

where $\mathcal{I}_{k-1}^k : \mathcal{M}_{k-1} \to \mathcal{M}_k$ is an interpolation operator. Here $\boldsymbol{\phi}_k$ represents the current approximation at level k as it was obtained by the smoothing process and before coarsening. If the high-frequency components of the error on the finer grid k were well damped, then the solution at level $\boldsymbol{\phi}_{k-1}$ should provide enough resolution for the error of $\boldsymbol{\phi}_k$ through $\mathcal{I}_{k-1}^k \boldsymbol{\phi}_{k-1}$.

The idea of transferring to a coarser grid can be applied along the set of nested meshes. One starts at level k with a given initial approximation (zero) and applies the smoothing iteration m_1 times. The residual is then computed and transferred to the next coarser grid while $\boldsymbol{\phi}_k$ obtained by smoothing is left unchanged. On the coarse grid with index $k-1$ the smoothing process is again applied. This procedure is repeated until the coarsest grid is reached.

On the coarsest grid, one solves the problem exactly and the result is used to improve $\boldsymbol{\phi}_k$ via (4.4). The CGC is then followed by m_2 postsmoothing steps at level k before the CGC procedure followed by postsmoothing is repeated for the next (if any) finer level. This entire process represents one multigrid $V(m_1, m_2)$-cycle.

A compact description of the multigrid method is given in section 6.

In the following sections we specify and analyze the multigrid components introduced here.

4.1. Smoothing iterations. Numerical experience [2] has shown that in order to obtain a multigrid algorithm which is robust with respect to changes of ν, care must be taken in the choice of the smoother. For example, when using the Picard–Gauss–Seidel iteration [1, 2], difficulties arise when the value of the weight of the cost of the control is smaller than h^2, which may easily occur when coarse grids are used. On the other hand, the collective Gauss–Seidel (CGS) scheme appears to be a reasonable choice [2]. Notice that this iterative method belongs to the class of Vanka smoothers [22].

To analyze the CGS scheme, let us introduce some notation:

$$\mathcal{A}_h^+ = \begin{bmatrix} \nu\,\Sigma_h^+ & 0 \\ 0 & \Sigma_h^+ \end{bmatrix}, \quad \mathcal{A}_h^- = \begin{bmatrix} \nu\,\Sigma_h^- & 0 \\ 0 & \Sigma_h^- \end{bmatrix}, \quad \mathcal{D}_h = \begin{bmatrix} \nu\,\frac{4}{h^2} I_h & -I_h \\ I_h & \frac{4}{h^2} I_h \end{bmatrix}, \tag{4.5}$$

where I_h is the identity operator on Ω_h and the operators Σ_h^+ and Σ_h^- are given in stencil form by

$$\Sigma_h^+ = \frac{1}{h^2} \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \qquad \Sigma_h^- = \frac{1}{h^2} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}. \tag{4.6}$$

Thus a sweep of the forward CGS scheme and of a backward CGS scheme are expressed by

$$(\mathcal{D}_h - \mathcal{A}_h^+)\,\boldsymbol{\phi}^{(1)} - \mathcal{A}_h^-\,\boldsymbol{\phi}^{(0)} = \mathbf{f} \quad \text{and} \quad (\mathcal{D}_h - \mathcal{A}_h^-)\,\boldsymbol{\phi}^{(2)} - \mathcal{A}_h^+\,\boldsymbol{\phi}^{(1)} = \mathbf{f}, \tag{4.7}$$

respectively. In the symmetric version of the CGS smoother, the forward CGS step is followed by a backward CGS step. The resulting iteration can be written in the

linear form

$$\boldsymbol{\phi}^{(2)} = \boldsymbol{\phi}^{(0)} + \mathcal{R}_h\,[f_h - \mathcal{A}_h \boldsymbol{\phi}^{(0)}], \qquad \text{where } \mathcal{R}_h = (\mathcal{D}_h - \mathcal{A}_h^-)^{-1}\,\mathcal{D}_h\,(\mathcal{D}_h - \mathcal{A}_h^+)^{-1},$$

which gives the smoothing operator $\mathcal{S}_h = I_h - \mathcal{R}_h\,\mathcal{A}_h$.

We analyze this iteration by local Fourier analysis [10, 21]. Consider the Fourier space spanned by the functions

$$\boldsymbol{\phi}(\boldsymbol{\theta}, \mathbf{x}) = \mathbf{a}\, e^{i\theta_1 x/h}\, e^{i\theta_2 y/h}, \qquad \boldsymbol{\theta} = (\theta_1, \theta_2),$$

where $\mathbf{a} = (1,1)^T$. One defines

$$\begin{aligned} \boldsymbol{\phi} \text{ low-frequency component } &\iff \boldsymbol{\theta} \in [-\frac{\pi}{2}, \frac{\pi}{2})^2, \\ \boldsymbol{\phi} \text{ high-frequency component } &\iff \boldsymbol{\theta} \in [-\pi, \pi)^2 \setminus [-\frac{\pi}{2}, \frac{\pi}{2})^2. \end{aligned}$$

In the Fourier space consider the symbols of the discrete operators $(\mathcal{D}_h - \mathcal{A}_h^+)$ and $\mathcal{A}_h^-$ for the forward CGS iteration. We have

$$\overline{(\mathcal{D}_h - \mathcal{A}_h^+)}(\boldsymbol{\theta}) = -\frac{1}{h^2}\begin{bmatrix} \nu\,(e^{-i\theta_1} + e^{-i\theta_2} - 4) & h^2 \\ -h^2 & (e^{-i\theta_1} + e^{-i\theta_2} - 4) \end{bmatrix}, \tag{4.8}$$

$$\overline{(\mathcal{A}_h^-)}(\boldsymbol{\theta}) = -\frac{1}{h^2}\begin{bmatrix} \nu\,(e^{i\theta_1} + e^{i\theta_2}) & 0 \\ 0 & (e^{i\theta_1} + e^{i\theta_2}) \end{bmatrix}. \tag{4.9}$$

Thus, the symbol of the forward CGS scheme is given by

$$\overline{\mathcal{S}_h^+}(\boldsymbol{\theta}) = (\overline{(\mathcal{D}_h - \mathcal{A}_h^+)}(\boldsymbol{\theta}))^{-1}\,\overline{(\mathcal{A}_h^-)}(\boldsymbol{\theta}).$$

In this framework, the smoothing factor of the forward CGS scheme for the optimality system is defined by

$$\mu = \mu(\mathcal{S}_h^+) = \sup\{|\rho(\overline{\mathcal{S}_h^+}(\boldsymbol{\theta}))| : \boldsymbol{\theta} \text{ high frequency }\}, \tag{4.10}$$

where ρ denotes the spectral radius. In the same way one defines the smoothing factor for the backward CGS step: $\mathcal{S}_h^- = (\mathcal{D}_h - \mathcal{A}_h^-)^{-1}\,(\mathcal{A}_h^+)$. The symmetric CGS scheme is then given by $\mathcal{S}_h^s = \mathcal{S}_h^-\,\mathcal{S}_h^+$. Since the symbols associated with the CGS iterative schemes considered here are 2×2 operators with entries being functions of (θ_1, θ_2), it is possible, by any symbolical package, to obtain the eigenvalues of the symbols. Thus we have the following.

Remark 2. By inspection in the range of high frequencies for $h \in [0.01, 0.25]$ and ν ranging in the interval $[10^{-6}, 1]$, the following upper bounds for the smoothing factor are found:

$$\mu(\mathcal{S}_h^+) \le 0.5, \quad \mu(\mathcal{S}_h^-) \le 0.5, \quad \text{and } \mu(\mathcal{S}_h^s) \le 0.25.$$

Therefore, we can conclude that the forward CGS, the backward CGS, and the symmetric CGS are all good smoothers for the purpose of the multigrid scheme.

4.2. Intergrid transfer operators. Among two grids $\overline{\Omega}_k$ and $\overline{\Omega}_{k-1}$, corresponding to mesh sizes h_k and h_{k-1}, we define a prolongation operator, $I_{k-1}^k : M_{k-1} \to M_k$, given in stencil form by

$$I_{k-1}^k = \frac{1}{4}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}. \tag{4.11}$$

This choice is consistent with the assumption of *bilinear* finite elements on each square partition of the discretization. That is, on each square partition $[x_i, x_{i+1}] \times [y_j, y_{j+1}]$ of $\overline{\Omega}_{k-1}$, the piecewise bilinear function which interpolates U at the nodes is given by

$$\tilde{u}(\tilde{x}, \tilde{y}) = (1-\tilde{x})(1-\tilde{y})\, u_{ij} + \tilde{x}(1-\tilde{y})\, u_{i+1\,j} + \tilde{y}(1-\tilde{x})\, u_{ij+1} + \tilde{x}\tilde{y}\, u_{i+1\,j+1}.$$

Here, $0 \le \tilde{x}, \tilde{y} \le 1$ are local coordinates such that $x = x_i + \tilde{x}\, h_{k-1}$ and $y = y_j + \tilde{y}\, h_{k-1}$. Thus the prolongation of u on a grid point of $\overline{\Omega}_k$ is the value of $\tilde{u}$ corresponding to that grid point.

Next, we define the full-weighting restriction operator, $I_k^{k-1} : M_k \to M_{k-1}$, given in stencil form by

$$I_k^{k-1} = \frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}, \tag{4.12}$$

with the inner product

$$(v, w)_k = \sum_{i=2}^{N_k} \sum_{j=2}^{N_k} h_k^2\, v_{kij}\, w_{kij}, \tag{4.13}$$

where $N_k = 2^k$. We have that the restriction operator is the adjoint of the prolongation operator [13], in the sense that

$$(I_k^{k-1} v_k, w_{k-1})_{k-1} = (v_k, I_{k-1}^k w_{k-1})_k \quad \text{for all } v_k \in M_k,\ w_{k-1} \in M_{k-1}.$$

The action of I_{k-1}^k (resp., I_k^{k-1}) on pairs of grid functions is denoted by $\mathcal{I}_{k-1}^k$ (resp., $\mathcal{I}_k^{k-1}$).

In order to extend the multigrid convergence theory formulated in [9, 15, 6] to the present multigrid method for optimality systems, we need the following lemma [9].

LEMMA 4.1. *Let us introduce the bilinear form* $a_k(u, v) = (-\Delta_k u, v)_k$, $u, v \in M_k$. *The prolongation operator* (4.11) *satisfies the following conditions:*

$$a_k(I_{k-1}^k u_{k-1}, I_{k-1}^k u_{k-1}) \le a_{k-1}(u_{k-1}, u_{k-1}) \qquad \textit{for all } u_{k-1} \in M_{k-1}, \tag{4.14}$$

$$(I_{k-1}^k u_{k-1}, I_{k-1}^k u_{k-1})_k \le (u_{k-1}, u_{k-1})_{k-1} \qquad \textit{for all } u_{k-1} \in M_{k-1}. \tag{4.15}$$

In particular, the result of Lemma 4.1 applied to the operator (3.5) results in the following:

$$(\mathcal{A}_k\, \mathcal{I}_{k-1}^k \mathbf{w}_{k-1}, \mathcal{I}_{k-1}^k \mathbf{w}_{k-1})_k \le (\mathcal{A}_{k-1} \mathbf{w}_{k-1}, \mathbf{w}_{k-1})_{k-1} \tag{4.16}$$

for all $\mathbf{w}_{k-1} = (u_{k-1}, v_{k-1}) \in \mathcal{M}_{k-1}$.

5. Two grid local Fourier analysis. In this section we perform local Fourier analysis [10, 21] of the two grid solution process for the optimal control optimality system. That is, we apply the local Fourier analysis to the two grid operator given by

$$TG_k^{k-1} = \mathcal{S}_k^{m_2} \left[\mathcal{I}_k - \mathcal{I}_{k-1}^k \left(\mathcal{A}_{k-1}\right)^{-1} \mathcal{I}_k^{k-1} \mathcal{A}_k\right] \mathcal{S}_k^{m_1}. \tag{5.1}$$

Here, the coarse grid operator is $CG_k^{k-1} = [\mathcal{I}_k - \mathcal{I}_{k-1}^k \left(\mathcal{A}_{k-1}\right)^{-1} \mathcal{I}_k^{k-1} \mathcal{A}_k]$.

The local Fourier analysis considers infinite grids, $G_k = \{(ih_k, jh_k), i, j \in \mathbb{Z}\}$, and therefore the influence of boundary conditions is not taken into account. Nevertheless, experience shows that local Fourier analysis provides predictions of multigrid convergence which are very sharp. This analysis is based on the quadruples of Fourier components

$$\phi_k(\boldsymbol{\theta}, \mathbf{x}) = e^{i\theta_1 x/h_k}\, e^{i\theta_2 y/h_k}$$

that coincide on G_{k-1}. For any low frequency $\boldsymbol{\theta} = (\theta_1, \theta_2) \in [-\pi/2, \pi/2)^2$, we consider

$$\begin{aligned} \boldsymbol{\theta}^{(0,0)} &:= (\theta_1, \theta_2), \qquad & \boldsymbol{\theta}^{(1,1)} &:= (\overline{\theta_1}, \overline{\theta_2}), \\ \boldsymbol{\theta}^{(1,0)} &:= (\overline{\theta_1}, \theta_2), \qquad & \boldsymbol{\theta}^{(0,1)} &:= (\theta_1, \overline{\theta_2}), \end{aligned}$$

where

$$\overline{\theta_i} = \begin{cases} \theta_i + \pi \text{ if } \theta_i < 0, \\ \theta_i - \pi \text{ if } \theta_i \geq 0. \end{cases}$$

We have $\phi(\boldsymbol{\theta}^{(0,0)}, \cdot) = \phi(\boldsymbol{\theta}^{(1,1)}, \cdot) = \phi(\boldsymbol{\theta}^{(1,0)}, \cdot) = \phi(\boldsymbol{\theta}^{(0,1)}, \cdot)$ for $\boldsymbol{\theta}^{(0,0)} \in [-\pi/2, \pi/2)^2$ and $(x, y) \in G_{k-1}$. Denote with $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ and consider $\boldsymbol{\alpha} \in \{(0,0), (1,1), (1,0), (0,1)\}$; then on G_{k-1} we have $\phi_k(\boldsymbol{\theta}^{\boldsymbol{\alpha}}, \mathbf{x}) = \phi_{k-1}(2\boldsymbol{\theta}^{(0,0)}, \mathbf{x})$. The four components $\phi_k(\boldsymbol{\theta}^{\boldsymbol{\alpha}}, \cdot)$ are called harmonics. For a given $\boldsymbol{\theta} = \boldsymbol{\theta}^{(0,0)} \in [-\pi/2, \pi/2)^2$, the four dimensional space of harmonics is defined by

$$E_k^\theta = span[\phi_k(\boldsymbol{\theta}^{\boldsymbol{\alpha}}, \cdot) : \boldsymbol{\alpha} \in \{(0,0), (1,1), (1,0), (0,1)\}].$$

For each θ, the spaces $E_k^\theta \times E_k^\theta$ are invariant under the action of TG_k^{k-1}; see [21]. We now study the action of TG_k^{k-1} on an arbitrary couple $(\psi_y, \psi_\lambda) \in E_k^\theta \times E_k^\theta$, where

$$\psi_y = \sum_{\boldsymbol{\alpha}} A^{\boldsymbol{\alpha}}\, \phi_k(\boldsymbol{\theta}^{\boldsymbol{\alpha}}, \cdot) \quad \text{ and } \quad \psi_\lambda = \sum_{\boldsymbol{\alpha}} B^{\boldsymbol{\alpha}}\, \phi_k(\boldsymbol{\theta}^{\boldsymbol{\alpha}}, \cdot).$$

We analyze how the vector of coefficients $(A^{(0,0)}, \dots, B^{(0,0)}, \dots)$ is transformed if the two grid iteration (5.1) is applied to (ψ_y, ψ_λ). We use the following theorem, which is an extension of Theorem 4.4.1 of [21] to our system of equations.

THEOREM 5.1. *Under the assumption that all multigrid components in* (5.1) *are linear and that* $(\mathcal{A}_{k-1})^{-1}$ *exists, the coarse grid operator* CG_k^{k-1} *is represented on* E_k^θ *by the* 8×8 *matrix* $\widehat{CG}_k^{k-1}(\boldsymbol{\theta})$,

$$\widehat{CG}_k^{k-1}(\boldsymbol{\theta}) = [\hat{\mathcal{I}}_k - \hat{\mathcal{I}}_{k-1}^k(\boldsymbol{\theta}) \left(\widehat{\mathcal{A}}_{k-1}(2\boldsymbol{\theta})\right)^{-1} \hat{\mathcal{I}}_k^{k-1}(\boldsymbol{\theta})\, \widehat{\mathcal{A}}_k(\boldsymbol{\theta})],$$

for each $\boldsymbol{\theta} \in [-\pi/2, \pi/2)^2$. *Here,* $\hat{\mathcal{I}}_k$ *and* $\widehat{\mathcal{A}}_k(\boldsymbol{\theta})$ *are* 8×8 *matrices,* $\hat{\mathcal{I}}_k^{k-1}(\boldsymbol{\theta})$ *is a* 2×8 *matrix,* $\hat{\mathcal{I}}_{k-1}^k(\boldsymbol{\theta})$ *is a* 8×2 *matrix, and* $\widehat{\mathcal{A}}_{k-1}(2\boldsymbol{\theta})$ *is a* 2×2 *matrix.*

If the spaces $E_k^\theta \times E_k^\theta$ are invariant under the smoothing operator $\mathcal{S}_k$, i.e., (the 8×8 matrix) $\hat{\mathcal{S}}_k(\boldsymbol{\theta}) : E_k^\theta \times E_k^\theta \to E_k^\theta \times E_k^\theta$ for all $\boldsymbol{\theta} \in [-\pi/2, \pi/2)^2$, we also have a representation of TG_k^{k-1} on $E_k^\theta \times E_k^\theta$ by a 8×8 matrix given by

$$\widehat{TG}_k^{k-1}(\boldsymbol{\theta}) = \hat{\mathcal{S}}_k(\boldsymbol{\theta})^{m_2}\, \widehat{CG}_k^{k-1}(\boldsymbol{\theta})\, \hat{\mathcal{S}}_k(\boldsymbol{\theta})^{m_1}.$$

We now give the symbols of the operators above in explicit form.

The coarse grid operator $\mathcal{A}_{k-1}$ is

$$\widehat{\mathcal{A}}_{k-1}(2\boldsymbol{\theta}) = \begin{bmatrix} \nu\, \frac{4-2(\cos(2\theta_1)+\cos(2\theta_2))}{h_{k-1}^2} & -1 \\ 1 & \frac{4-2(\cos(2\theta_1)+\cos(2\theta_2))}{h_{k-1}^2} \end{bmatrix}.$$

The fine grid operator is $\mathcal{A}_k$. The symbol $\widehat{\mathcal{A}}_k(\boldsymbol{\theta})$ is given by

$$\begin{bmatrix} \nu\, l(\boldsymbol{\theta}^{(0,0)}) & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & \nu\, l(\boldsymbol{\theta}^{(1,1)}) & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & \nu\, l(\boldsymbol{\theta}^{(1,0)}) & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & \nu\, l(\boldsymbol{\theta}^{(0,1)}) & 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 & l(\boldsymbol{\theta}^{(0,0)}) & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & l(\boldsymbol{\theta}^{(1,1)}) & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & l(\boldsymbol{\theta}^{(1,0)}) & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & l(\boldsymbol{\theta}^{(0,1)}) \end{bmatrix},$$

where

$$l(\boldsymbol{\theta}^{\boldsymbol{\alpha}}) = \frac{4 - 2(\cos(\theta_1^{\alpha_1}) + \cos(\theta_2^{\alpha_2}))}{h_k^2}.$$

The restriction operator is $\mathcal{I}_k^{k-1}$. The symbol $\hat{\mathcal{I}}_k^{k-1}(\boldsymbol{\theta})$ is given by

$$\begin{bmatrix} I(\boldsymbol{\theta}^{(0,0)}) & I(\boldsymbol{\theta}^{(1,1)}) & I(\boldsymbol{\theta}^{(1,0)}) & I(\boldsymbol{\theta}^{(0,1)}) & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I(\boldsymbol{\theta}^{(0,0)}) & I(\boldsymbol{\theta}^{(1,1)}) & I(\boldsymbol{\theta}^{(1,0)}) & I(\boldsymbol{\theta}^{(0,1)}) \end{bmatrix},$$

where

$$I(\boldsymbol{\theta}^{\boldsymbol{\alpha}}) = I_k^{k-1}(\boldsymbol{\theta}^{\boldsymbol{\alpha}}) = \frac{1}{4}(1 + \cos(\theta_1^{\alpha_1}))(1 + \cos(\theta_2^{\alpha_2})).$$

For the prolongation operator we have $\hat{\mathcal{I}}_{k-1}^k(\boldsymbol{\theta}) = \hat{\mathcal{I}}_k^{k-1}(\boldsymbol{\theta})^T$.

For the smoothing iteration $\mathcal{S}_k$ consider the forward CGS scheme as described in section 4.1. On $E_k^\theta \times E_k^\theta$ it is given by $(\overline{(\mathcal{D}_h - \mathcal{A}_h^+)}(\boldsymbol{\theta}))^{-1}\, \overline{(\mathcal{A}_h^-)}(\boldsymbol{\theta})$, where $(\overline{(\mathcal{D}_h - \mathcal{A}_h^+)}(\boldsymbol{\theta}))$ is as follows:

$$\frac{1}{h^2}\begin{bmatrix} \nu\, s_+(\boldsymbol{\theta}^{(0,0)}) & 0 & 0 & 0 & -h^2 & 0 & 0 & 0 \\ 0 & \nu\, s_+(\boldsymbol{\theta}^{(1,1)}) & 0 & 0 & 0 & -h^2 & 0 & 0 \\ 0 & 0 & \nu\, s_+(\boldsymbol{\theta}^{(1,0)}) & 0 & 0 & 0 & -h^2 & 0 \\ 0 & 0 & 0 & \nu\, s_+(\boldsymbol{\theta}^{(0,1)}) & 0 & 0 & 0 & -h^2 \\ h^2 & 0 & 0 & 0 & s_+(\boldsymbol{\theta}^{(0,0)}) & 0 & 0 & 0 \\ 0 & h^2 & 0 & 0 & 0 & s_+(\boldsymbol{\theta}^{(1,1)}) & 0 & 0 \\ 0 & 0 & h^2 & 0 & 0 & 0 & s_+(\boldsymbol{\theta}^{(1,0)}) & 0 \\ 0 & 0 & 0 & h^2 & 0 & 0 & 0 & s_+(\boldsymbol{\theta}^{(0,1)}) \end{bmatrix},$$

and the operator $\overline{(\mathcal{A}_h^-)}(\boldsymbol{\theta})$ is given by

$$\frac{1}{h^2}\begin{bmatrix} \nu\, s_-(\boldsymbol{\theta}^{(0,0)}) & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \nu\, s_-(\boldsymbol{\theta}^{(1,1)}) & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \nu\, s_-(\boldsymbol{\theta}^{(1,0)}) & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \nu\, s_-(\boldsymbol{\theta}^{(0,1)}) & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & s_-(\boldsymbol{\theta}^{(0,0)}) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & s_-(\boldsymbol{\theta}^{(1,1)}) & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & s_-(\boldsymbol{\theta}^{(1,0)}) & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & s_-(\boldsymbol{\theta}^{(0,1)}) \end{bmatrix},$$

TABLE 5.1
Convergence factors and smoothing factors.

	Local Fourier analysis		Experim.
(m_1, m_2)	$\mu_{loc}^{m_1+m_2}$	$\eta(TG_k^{k-1})$	$V(m_1, m_2)$
(1,1)	0.25	0.25	0.30
(2,1)	0.125	0.12	0.12
(2,2)	0.06	0.08	0.08
(3,2)	0.03	0.06	0.06
(3,3)	0.01	0.05	0.05

where

$$s_+(\boldsymbol{\theta}^{\boldsymbol{\alpha}}) = 4 - e^{-i\theta_1^{\alpha_1}} - e^{-i\theta_2^{\alpha_2}} \quad \text{and} \quad s_-(\boldsymbol{\theta}^{\boldsymbol{\alpha}}) = -e^{i\theta_1^{\alpha_1}} - e^{i\theta_2^{\alpha_2}}.$$

Based on the representation on TG_k^{k-1} by a 8×8 matrix $\widehat{TG}_k^{k-1}(\boldsymbol{\theta})$, we can calculate the *convergence factor*:

$$\eta(TG_k^{k-1}) = \sup\{\rho(\widehat{TG}_k^{k-1}(\boldsymbol{\theta})) : \boldsymbol{\theta} \in [-\pi/2, \pi/2)^2\}.$$

Here, $\rho(\widehat{TG}_k^{k-1}(\boldsymbol{\theta}))$ is the spectral radius of $\widehat{TG}_k^{k-1}(\boldsymbol{\theta})$.

Under the invariance property advocated by Theorem 5.1, to measure the smoothing property of the iteration one can assume an *ideal coarse grid correction* which annihilates the low-frequency error components and leaves the high-frequency error components unchanged. That is, one defines the projection operator Q_k^{k-1} on E_k^θ by

$$Q_k^{k-1}\,\phi(\boldsymbol{\theta}, \cdot) = \begin{cases} 0 \text{ if } \boldsymbol{\theta} = \boldsymbol{\theta}^{(0,0)} \in [-\pi/2, \pi/2)^2, \\ \phi(\boldsymbol{\theta}, \cdot) \text{ if } \boldsymbol{\theta} \in \{\boldsymbol{\theta}^{(1,1)}, \boldsymbol{\theta}^{(1,0)}, \boldsymbol{\theta}^{(0,1)}\}. \end{cases}$$

On the space $E_k^\theta \times E_k^\theta$ we then have

$$\widehat{Q}_k^{k-1}(\boldsymbol{\theta}) = \begin{bmatrix} Q_k^{k-1} & 0 \\ 0 & Q_k^{k-1} \end{bmatrix} \qquad \text{for } \boldsymbol{\theta} \in [-\pi/2, \pi/2)^2. \tag{5.2}$$

In this framework the smoothing property of $\mathcal{S}_k$ is defined as follows:

$$\mu_{loc} = \mu(\mathcal{S}_k, m) = \sup\left\{ \sqrt[m]{|\rho(\widehat{Q}_k^{k-1}\,\hat{\mathcal{S}}_k(\boldsymbol{\theta})^m)|} : \boldsymbol{\theta} \in [-\pi/2, \pi/2)^2 \right\}. \tag{5.3}$$

Notice that, assuming an ideal CGC takes place, the convergence factor of the two grid scheme is given by $\mu_{loc}^{m_1+m_2}$.

We complete this section by reporting in Table 5.1 the values of $\eta(TG_k^{k-1})$ and those of $\mu(\mathcal{S}_k, m)$ obtained with the two grid analysis described above. Here the forward Gauss–Seidel smoother is used. For comparison, the observed value of convergence factor defined as the "asymptotic" value of the ratio between the discrete L^2 norms of residuals resulting from two successive multigrid cycles on the finest mesh is reported. Notice that the values reported in Table 5.1 are typical of the standard Poisson model problem. These values have been obtained considering the mesh size value h ranging in the interval $[0.01, 0.25]$ corresponding to the interval of mesh sizes used in the multigrid code. The value of the weight ν has been taken in the interval $[10^{-6}, 1]$.

6. General multigrid convergence theory. In this section, we prove multigrid convergence for the optimal control problem in a more general functional setting. We use the framework in [6, 9, 11] adapted to the nonsymmetric system above. This framework applies directly to elliptic problems with prescribed boundary conditions on bounded polygonal domains.

For the purpose of our analysis, we briefly describe multigrid convergence theory for scalar Poisson equation discretized by the finite difference method on a unit square. Consider

$$\begin{aligned} -\Delta y &= f \text{ in } \Omega, \\ y &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{6.1}$$

The matrix form of this problem is

$$\hat{A}_k y_k = f_k. \tag{6.2}$$

Let $\hat{P}_{k-1} : M_k \to M_{k-1}$ (resp., $I_k^{k-1} : M_k \to M_{k-1}$) be the $\hat{A}_k$ (resp., L_k^2) projections defined by

$$(\hat{A}_{k-1}\hat{P}_{k-1}u, v)_{k-1} = (\hat{A}_k u, I_{k-1}^k v)_k \text{ (resp., } (I_k^{k-1}u, v)_{k-1} = (u, I_{k-1}^k v)_k)$$

for all $u \in M_k$ and $v \in M_{k-1}$. Let $\hat{R}_k : M_k \to M_k$ be an iteration operator. Then the V-cycle multigrid algorithm to solve (6.2) in recursive form is given as follows.

Multigrid Algorithm $V(m_1, m_2)$.

Set $\hat{B}_1 = \hat{A}_1^{-1}$. For $k \geq 2$ define $\hat{B}_k : M_k \to M_k$ in terms of $\hat{B}_{k-1}$ as follows. Let $g \in M_k$.

1. Set $y^0 = 0$.
2. Define y^l for $l = 1, \ldots, m_1$ by
$$y^l = y^{l-1} + \hat{R}_k(g - \hat{A}_k\, y^{l-1}).$$
3. Set $y^{m_1+1} = y^{m_1} + I_{k-1}^k q$, where
$$q = \hat{B}_{k-1} I_k^{k-1}(g - \hat{A}_k\, y^{m_1}).$$
4. Set $\hat{B}_k g = y^{m_1+m_2+1}$, where y^ℓ for $\ell = m_1 + 2, \ldots, m_1 + m_2 + 1$ is given by step 2 ($\hat{R}_k^t$ instead of $\hat{R}_k$).

For the purpose of analysis, we take $m_1 = 1$ and $m_2 = 0$.

From the definition of $\hat{P}_{k-1}$, we see that

$$I_k^{k-1}\hat{A}_k = \hat{A}_{k-1}\hat{P}_{k-1}.$$

Let $\hat{S}_k = I_k - \hat{R}_k\hat{A}_k$ for $k > 1$, where I_k denotes the identity on M_k. Then $\hat{S}_k\, y = y - y^1$. Now for $y \in M_k$, $k = 2, \ldots, L$, we have

$$\begin{aligned} (I_k - \hat{B}_k\hat{A}_k)\, y &= y - y^1 - I_{k-1}^k q \\ &= \hat{S}_k\, y - I_{k-1}^k \hat{B}_{k-1}\hat{A}_{k-1}\hat{P}_{k-1}\hat{S}_k\, y \\ &= [I_k - I_{k-1}^k \hat{B}_{k-1}\hat{A}_{k-1}\hat{P}_{k-1}]\, \hat{S}_k\, y \\ &= [(I_k - I_{k-1}^k\hat{P}_{k-1}) + I_{k-1}^k(I_{k-1} - \hat{B}_{k-1}\hat{A}_{k-1})\, \hat{P}_{k-1}]\hat{S}_k\, y. \end{aligned} \tag{6.3}$$

The convergence results of the multigrid method are expressed in terms of the error operators $\hat{E}_k := I_k - \hat{B}_k\hat{A}_k$ and $\hat{E} := \hat{E}_L$. In the following, let C denote a

generic constant independent of k that can have different values in different places, unless otherwise stated.

In order to prove convergence of the multigrid algorithm, the following two conditions are required. There exists a constant $\bar{C}_{\hat{R}}$ independent of y and k such that

$$\frac{|y|_0^2}{\mu(\hat{A}_k)} \leq \bar{C}_{\hat{R}}(\bar{\hat{R}}y, y) \quad \text{for all } y \in M_k, \tag{6.4}$$

where $\mu(\hat{A}_k)$ denotes the maximum eigenvalue of $\hat{A}_k$, $\bar{\hat{R}} = (I_k - \hat{S}_k^* \hat{S}_k)\,\hat{A}_k^{-1}$, $\hat{S}_k^* = I - \hat{R}_k^t \hat{A}_k$, and $*$ denotes adjoint with respect to the inner product $(\hat{A}_k \cdot, \cdot)$. Next, for $k > 1$ define $\hat{T}_k = \hat{R}_k \hat{A}_k$. We assume that there exists a constant θ, $0 < \theta < 2$, independent of y such that

$$(\hat{A}_k \hat{T}_k\, y, \hat{T}_k\, y)_k \leq \theta(\hat{A}_k \hat{T}_k\, y, y)_k \quad \text{for all } y \in M_k. \tag{6.5}$$

In this paper, we are dealing with multigrid for finite difference method applied to the Poisson equation on rectangular domains. In this case the stiffness matrix is exactly the same as that arising from the finite element case. Hence we have the following result from [9].

THEOREM 6.1. *Let $\hat{R}_k$ satisfy* (6.4) *and* (6.5) *for $k > 1$. Then there exists a positive constant $\hat{\delta} < 1$ such that*

$$(\hat{A}_L \hat{E}_L y, \hat{E}_L y)_L \leq \hat{\delta}^2\,(\hat{A}_L y, y)_L \quad \textit{for all } y \in M_L,$$

where $\hat{\delta} = CL/(CL+1)$.

Remark 3. The dependence of $\hat{\delta}$ on L can be removed by a perturbation analysis given in [16].

Remark 4. For the multigrid algorithm $V(m,0)$ one obtains $\hat{\delta} = CL/(CL+m)$; see [9]. The constant C depends linearly on $\bar{C}_{\hat{R}}$; see [7, 8, 9] to find estimates of these constants. As discussed in [7], the $\hat{\delta}$ estimate in Theorem 6.1 is pessimistic, in the sense that the observed $\hat{\delta}$ is smaller than the theoretical one and their difference becomes larger for larger values of m.

To prove convergence of multigrid for the optimal control optimality system, we first consider the decoupled symmetric system:

$$\begin{aligned} -\nu\Delta y &= \nu g \text{ in } \Omega, \\ y &= 0 \text{ on } \partial\Omega, \\ -\Delta\lambda &= z \text{ in } \Omega, \\ \lambda &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{6.6}$$

This system is exactly two copies of Poisson equation, hence the multigrid convergence theory for this system inherits the properties of the scalar case. In fact, if we define

$$\hat{\mathcal{A}}_k = \begin{pmatrix} \nu\,\hat{A}_k & 0 \\ 0 & \hat{A}_k \end{pmatrix}, \tag{6.7}$$

and analogously $\hat{\mathcal{B}}_k$, $\hat{\mathcal{E}}_k$, etc., as the system counterparts of $\hat{B}_k$, $\hat{E}_k$, etc., then the multigrid algorithm has exactly the same form as (6.3) with $\hat{\mathcal{B}}_k$, $\hat{\mathcal{A}}_k$, etc., replacing $\hat{B}_k$, $\hat{A}_k$, etc. As a consequence we have the following theorem.

THEOREM 6.2. *Under the assumption of Theorem* 6.1, *there exists a positive constant* $\hat{\delta} < 1$ *such that*

$$(\hat{\mathcal{A}}_L\hat{\mathcal{E}}_L(y,\lambda),\hat{\mathcal{E}}_L(y,\lambda))_L \le \hat{\delta}^2(\hat{\mathcal{A}}_L(y,\lambda),(y,\lambda))_L \quad \textit{for all } (y,\lambda)\in\mathcal{M}_L, \tag{6.8}$$

where $\hat{\delta}$ *has the same form as in Theorem* 6.1.

To analyze the optimality system we let

$$\mathcal{A}_k = \hat{\mathcal{A}}_k + d_k,$$

where

$$d_k = \begin{pmatrix} 0 & -I_k \\ I_k & 0 \end{pmatrix}.$$

We note that

$$|(d_k(u,v),(y,\lambda))| \le C\,|(u,v)|_0\,|(y,\lambda)|_0, \tag{6.9}$$

for some constant C. Now, the multigrid algorithm corresponding to this nonsymmetric problem has exactly the same recursive form as (6.3) with $\mathcal{B}_k$, $\mathcal{A}_k$, etc., replacing $\hat{B}_k$, $\hat{A}_k$, etc., and thus,

$$\mathcal{E}_k = \mathcal{I}_k - \mathcal{B}_k\mathcal{A}_k = [\mathcal{I}_k - \mathcal{I}_{k-1}^k\mathcal{P}_{k-1} + \mathcal{I}_{k-1}^k(\mathcal{I}_{k-1} - \mathcal{B}_{k-1}\mathcal{A}_{k-1})\,\mathcal{P}_{k-1}]\,\mathcal{S}_k, \tag{6.10}$$

where $\mathcal{I}_k$ is the identity operator on $\mathcal{M}_k$. We need a subspace decomposition of $\mathcal{M}_k$. Let

$$\mathcal{M}_k = \sum_{i=1}^{\ell}\mathcal{M}_k^i, \tag{6.11}$$

where ℓ is the number of grid points of the discrete domain and $\mathcal{M}_k^i$ is a two dimensional subspace of $\mathcal{M}_k$ consisting of nodal functions with zero nodal values except at the grid point i. Denote the decomposition of $\mathcal{A}_k$ (resp., $\hat{\mathcal{A}}_k$) with respect to the subspace $\mathcal{M}_k^i$ by $\mathcal{A}_k^i : \mathcal{M}_k^i \to \mathcal{M}_k^i$ (resp., $\hat{\mathcal{A}}_k^i$), satisfying

$$(\mathcal{A}_k^i\mathbf{w},\boldsymbol{\chi})_k = (\mathcal{A}_k\mathbf{w},\boldsymbol{\chi})_k \quad \text{for all } \boldsymbol{\chi}\in\mathcal{M}_k^i,\ \mathbf{w}\in\mathcal{M}_k^i.$$

Define $\mathcal{P}_k^i : \mathcal{M}_k \to \mathcal{M}_k^i$ (resp., $\hat{\mathcal{P}}_k^i$) by

$$(\mathcal{A}_k\mathcal{P}_k^i\mathbf{w},\boldsymbol{\chi})_k = (\mathcal{A}_k\mathbf{w},\boldsymbol{\chi})_k \quad \text{for all } \boldsymbol{\chi}\in\mathcal{M}_k^i,\ \mathbf{w}\in\mathcal{M}_k. \tag{6.12}$$

We use the notation $(\mathbf{w},\boldsymbol{\chi})_{0,i} = (\mathbf{w},\boldsymbol{\chi})_0$ and $(\mathbf{w},\boldsymbol{\chi})_{1,i} = (\mathbf{w},\boldsymbol{\chi})_1$ for $\boldsymbol{\chi}\in\mathcal{M}_k^i$. In the case of a CGS smoother, we obtain a product representation (see [8]) of $\mathcal{S}_k$: For this purpose we set $\mathbf{w}^0 = 0$, for $i = 1,\dots,\ell$,

$$\mathbf{w}^i = \mathbf{w}^{i-1} + (\mathcal{A}_k^i)^{-1}\mathcal{Q}_k^i\,(\mathbf{f}_k - \mathcal{A}_k\mathbf{w}^{i-1}), \tag{6.13}$$

and $\mathcal{R}_k\mathbf{f}_k = \mathbf{w}^\ell$. From the identity $\mathcal{A}_k^i\mathcal{P}_k^i = \mathcal{Q}_k^i\mathcal{A}_k$ on $\mathcal{M}_k^i$ it follows that $\mathcal{S}_k = \mathcal{I}_k - \mathcal{R}_k\mathcal{A}_k = \prod_{i=1}^{\ell}(\mathcal{I}_k - \mathcal{P}_k^i)$. Here, the operator $\mathcal{Q}_k^i : \mathcal{M}_k \to \mathcal{M}_k^i$ represents the orthogonal projection onto $\mathcal{M}_k^i$ with respect to $(\cdot,\cdot)_k$. Theorem 3.2 of [8] applies here to prove that (6.13) satisfies (6.4) and (6.5).

LEMMA 6.3. *For* $\mathbf{v}, \mathbf{w} \in \mathcal{M}_k$, *we have*

$$|(\hat{\mathcal{A}}_k \hat{\mathcal{P}}_k^i \mathbf{w}, \mathbf{v})_k| \leq C\, |\mathbf{w}|_1\, |\mathbf{v}|_1 \tag{6.14}$$

and

$$|(\hat{A}_k(\hat{\mathcal{P}}_k^i - \mathcal{P}_k^i)\mathbf{w}, \mathbf{v})| \leq C\, h_k\, |\mathbf{w}|_1\, |\mathbf{v}|_1. \tag{6.15}$$

Proof. By coercivity, Lemma 3.2, and the Poincaré inequality, it follows that there exists a positive constant α such that

$$\begin{aligned} \alpha\, |\hat{\mathcal{P}}_k^i\, \mathbf{w}|_1^2 &\leq (\hat{\mathcal{A}}_k \hat{\mathcal{P}}_k^i \mathbf{w}, \hat{\mathcal{P}}_k^i \mathbf{w})_k = (\hat{\mathcal{A}}_k \mathbf{w}, \hat{\mathcal{P}}_k^i \mathbf{w})_k \\ &\leq C\, |\mathbf{w}|_1\, |\hat{\mathcal{P}}_k^i \mathbf{w}|_1. \end{aligned}$$

Hence $\hat{\mathcal{P}}_k^i$ is bounded in the discrete energy norm and (6.14) is obtained by the Cauchy–Schwarz inequality. For (6.15), we have

$$\begin{aligned} (\hat{\mathcal{A}}_k(\hat{\mathcal{P}}_k^i - \mathcal{P}_k^i)\mathbf{w}, \mathbf{v})_k &= (\hat{\mathcal{A}}_k(\hat{\mathcal{P}}_k^i - \mathcal{P}_k^i)\mathbf{w}, \hat{\mathcal{P}}_k^i \mathbf{v})_k \\ &= (\hat{\mathcal{A}}_k \mathbf{w}, \hat{\mathcal{P}}_k^i \mathbf{v})_k - (\hat{\mathcal{A}}_k \mathcal{P}_k^i \mathbf{w}, \hat{\mathcal{P}}_k^i \mathbf{v})_k \\ &= (\hat{\mathcal{A}}_k \mathbf{w}, \hat{\mathcal{P}}_k^i \mathbf{v})_k - (\mathcal{A}_k \mathbf{w}, \hat{\mathcal{P}}_k^i \mathbf{v})_k + (d_k\, \mathcal{P}_k^i \mathbf{w}, \hat{\mathcal{P}}_k^i \mathbf{v})_k \\ &= -(d_k\, \mathbf{w}, \hat{\mathcal{P}}_k^i \mathbf{v})_k + (d_k\, \mathcal{P}_k^i \mathbf{w}, \hat{\mathcal{P}}_k^i \mathbf{v})_k \\ &= -(d_k\, (\mathcal{I}_k - \mathcal{P}_k^i)\mathbf{w}, \hat{\mathcal{P}}_k^i \mathbf{v})_k. \end{aligned}$$

Taking the absolute value we get by the Poincaré inequality,

$$\begin{aligned} |(\hat{\mathcal{A}}_k(\hat{\mathcal{P}}_k^i - \mathcal{P}_k^i)\mathbf{w}, \mathbf{v})_k| &\leq C\, |(\mathcal{I}_k - \mathcal{P}_k^i)\mathbf{w}|_{0,i}\, |\hat{\mathcal{P}}_k^i \mathbf{v}|_{1,i} \\ &\leq C\, h_k\, |\mathbf{w}|_1\, |\mathbf{v}|_1, \end{aligned}$$

where the boundedness of $\hat{\mathcal{P}}_k^i$ is used for the second inequality. ∎

The proof of the following lemma is based on subspace decomposition and proved with the aid of Lemma 6.3 exactly in the same way as that of Theorem 3.1 in [11]. We skip the details.

LEMMA 6.4. *There exists some constant* $C_{\mathcal{S}}$ *independent of* k *such that*

$$|(\hat{\mathcal{A}}_k(\mathcal{S}_k - \hat{\mathcal{S}}_k)\mathbf{w}, \mathbf{v})_k| \leq C_{\mathcal{S}}\, h_k\, |\mathbf{w}|_1\, |\mathbf{v}|_1 \tag{6.16}$$

for all $\mathbf{w}, \mathbf{v} \in \mathcal{M}_k$.

LEMMA 6.5. *The following inequalities hold:*

$$|(\hat{\mathcal{A}}_{k-1}(\hat{\mathcal{P}}_{k-1} - \mathcal{P}_{k-1})\mathbf{w}, \mathbf{v})_{k-1}| \leq C_{\mathcal{P}}\, h_{k-1}\, |\mathbf{w}|_1 |\mathbf{v}|_1 \text{ for } \mathbf{w} \in \mathcal{M}_k,\ \mathbf{v} \in \mathcal{M}_{k-1} \tag{6.17}$$

and

$$|(\hat{\mathcal{A}}_k(\mathcal{I}_k - \mathcal{I}_{k-1}^k \mathcal{P}_{k-1})\mathbf{w}, \mathbf{v})_k| \leq C_I\, h_k\, |\mathbf{w}|_1\, |\mathbf{v}|_1 \text{ for } \mathbf{w} \in \mathcal{M}_k,\ \mathbf{v} \in \mathcal{M}_k, \tag{6.18}$$

where $C_{\mathcal{P}}$ *and* C_I *are some constants independent of* k.

Proof. Let us first prove (6.17) with $\hat{\mathcal{A}}_{k-1}$ replaced by $\mathcal{A}_{k-1}$. We have for $\mathbf{w} \in \mathcal{M}_k$ and $\mathbf{v} \in \mathcal{M}_{k-1}$,

$$\begin{aligned}
&|(\mathcal{A}_{k-1}\hat{\mathcal{P}}_{k-1}\mathbf{w}, \mathbf{v})_{k-1} - (\mathcal{A}_{k-1}\mathcal{P}_{k-1}\mathbf{w}, \mathbf{v})_{k-1}| \\
&= |(\hat{\mathcal{A}}_{k-1}\hat{\mathcal{P}}_{k-1}\mathbf{w}, \mathbf{v})_{k-1} + (d_{k-1}\,\hat{\mathcal{P}}_{k-1}\mathbf{w}, \mathbf{v})_{k-1} - (\mathcal{A}_k\mathbf{w}, \mathcal{I}_{k-1}^k\mathbf{v})_k| \\
&= |(\hat{\mathcal{A}}_k\mathbf{w}, \mathcal{I}_{k-1}^k\mathbf{v})_k + (d_{k-1}\,\hat{\mathcal{P}}_{k-1}\mathbf{w}, \mathbf{v})_{k-1} - (\mathcal{A}_k\mathbf{w}, \mathcal{I}_{k-1}^k\mathbf{v})_k| \\
&= |(d_{k-1}\,\hat{\mathcal{P}}_{k-1}\mathbf{w}, \mathbf{v})_{k-1} - (d_k\,\mathbf{w}, \mathcal{I}_{k-1}^k\mathbf{v})_k| \\
&= |(d_{k-1}\,\hat{\mathcal{P}}_{k-1}\mathbf{w}, \mathbf{v})_{k-1} - (d_{k-1}\,\mathcal{I}_k^{k-1}\mathbf{w}, \mathbf{v})_{k-1}| \\
&= |(d_{k-1}\,(\hat{\mathcal{P}}_{k-1} - \mathcal{I}_k^{k-1})\mathbf{w}, \mathbf{v})_{k-1}| \\
&\le C\,|(\hat{\mathcal{P}}_{k-1} - \mathcal{I}_k^{k-1})\mathbf{w}|_0\,|\mathbf{v}|_1,
\end{aligned}$$

where the last inequality is obtained as follows: Let us denote by $\bar{\hat{\mathcal{P}}}_{k-1}$ the elliptic projection of the linear finite element method and denote by $\bar{\mathcal{I}}_{k-1} : \mathcal{M}_k \to \mathcal{M}_{k-1}$ the fine-to-coarse injection. We have

$$|(\hat{\mathcal{P}}_{k-1} - \mathcal{I}_k^{k-1})\mathbf{w}|_0 \le |(\hat{\mathcal{P}}_{k-1} - \bar{\hat{\mathcal{P}}}_{k-1})\mathbf{w}|_0 + |(\bar{\hat{\mathcal{P}}}_{k-1} - \bar{\mathcal{I}}_{k-1})\mathbf{w}|_0 + |(\bar{\mathcal{I}}_{k-1} - \mathcal{I}_k^{k-1})\mathbf{w}|_0.$$

Here, inequality (6.7), (6.8), (6.9), and (6.11) of [9], and the approximation property of $\mathcal{I}_k^{k-1}$, are used to obtain the estimate $|(\hat{\mathcal{P}}_{k-1} - \mathcal{I}_k^{k-1})\mathbf{w}|_0 \le C\,h_{k-1}\,|\mathbf{w}|_1$.

It follows that $\|\hat{\mathcal{P}}_{k-1} - \mathcal{P}_{k-1}\|_{\mathcal{A}_{k-1}} \le C\,h_{k-1}$, where $\|\cdot\|_{\mathcal{A}_{k-1}}$ denotes the usual operator norm induced by $\mathcal{A}_{k-1}$. Because $(\mathcal{A}_{k-1}\mathbf{v}, \mathbf{v})_{k-1} = (\hat{\mathcal{A}}_{k-1}\mathbf{v}, \mathbf{v})_{k-1}$ from the definition of d_{k-1}, we also have $\|\hat{\mathcal{P}}_{k-1} - \mathcal{P}_{k-1}\|_{\hat{\mathcal{A}}_{k-1}} \le C\,h_{k-1}$, which is the desired result.

The second assertion (6.18) follows directly from (6.17). ∎

With these preparations we can show the following theorem.

THEOREM 6.6. *There exist positive constants h_0 and $\delta < 1$ such that for all $h_1 < h_0$ we have*

$$(\hat{\mathcal{A}}_L\mathcal{E}_L\mathbf{w}, \mathcal{E}_L\mathbf{w})_L \le \delta^2(\hat{\mathcal{A}}_L\mathbf{w}, \mathbf{w})_L \quad \textit{for all } \mathbf{w} \in \mathcal{M}_L,$$

where $\delta = \hat{\delta} + Ch_1$ and $\hat{\delta}$ is as in Theorem 6.2.

Proof. Denoting the operator norm $\|\cdot\|_{\hat{\mathcal{A}}_k}$ by $\|\cdot\|$, we show that $\|\mathcal{E}_k - \hat{\mathcal{E}}_k\| \le c_k h_1$, where c_k is uniformly bounded. The error operator $\mathcal{E}_k$ can be written as

$$\mathcal{E}_k = (\mathcal{I}_k - \mathcal{I}_{k-1}^k\mathcal{B}_{k-1}\mathcal{A}_{k-1}\mathcal{P}_{k-1})\,\mathcal{S}_k,$$

and $\hat{\mathcal{E}}_k$ has similar representation. We compare the error operators and write their difference as

$$\begin{aligned}
\mathcal{E}_k - \hat{\mathcal{E}}_k = &(\mathcal{I}_k - \mathcal{I}_{k-1}^k\mathcal{B}_{k-1}\mathcal{A}_{k-1}\mathcal{P}_{k-1})\,(\mathcal{S}_k - \hat{\mathcal{S}}_k) \\
&- \mathcal{I}_{k-1}^k\mathcal{B}_{k-1}\mathcal{A}_{k-1}(\mathcal{P}_{k-1} - \hat{\mathcal{P}}_{k-1})\,\hat{\mathcal{S}}_k + \mathcal{I}_{k-1}^k(\mathcal{E}_{k-1} - \hat{\mathcal{E}}_{k-1})\,\hat{\mathcal{P}}_{k-1}\,\hat{\mathcal{S}}_k.
\end{aligned}$$

Thus in terms of the operator norm, we have by (4.16)

$$\begin{aligned}
\|\mathcal{E}_k - \hat{\mathcal{E}}_k\| \le &\|\mathcal{I}_k - \mathcal{I}_{k-1}^k\mathcal{B}_{k-1}\mathcal{A}_{k-1}\mathcal{P}_{k-1}\|\,\|\mathcal{S}_k - \hat{\mathcal{S}}_k\| \\
&+ \|\mathcal{B}_{k-1}\mathcal{A}_{k-1}\|\,\|\mathcal{P}_{k-1} - \hat{\mathcal{P}}_{k-1}\|\,\|\hat{\mathcal{S}}_k\| \\
&+ \|\mathcal{E}_{k-1} - \hat{\mathcal{E}}_{k-1}\|\,\|\hat{\mathcal{P}}_{k-1}\hat{\mathcal{S}}_k\|.
\end{aligned} \tag{6.19}$$

Let us make the induction hypothesis: $\|\mathcal{E}_{k-1} - \hat{\mathcal{E}}_{k-1}\| \le c_{k-1}h_1$, where c_{k-1} is a constant to be defined below. By the triangle inequality and Theorem 6.2,

$$\|\mathcal{E}_{k-1}\| \le \hat{\delta} + c_{k-1}h_1 \tag{6.20}$$

and

$$\|\mathcal{B}_{k-1}\mathcal{A}_{k-1}\| \le 1 + \hat{\delta} + c_{k-1}h_1. \tag{6.21}$$

Using the induction hypothesis, (4.14), Lemma 6.4, and Lemma 6.5, we have

$$\begin{aligned}
&\|\mathcal{I}_k - \mathcal{I}_{k-1}^k \mathcal{B}_{k-1}\mathcal{A}_{k-1}\mathcal{P}_{k-1}\| \\
&\qquad \le \|\mathcal{I}_k - \mathcal{I}_{k-1}^k \mathcal{P}_{k-1}\| + \|\mathcal{I}_{k-1} - \mathcal{B}_{k-1}\mathcal{A}_{k-1}\|\,\|\mathcal{P}_{k-1}\| && (6.22)\\
&\qquad \le C_I h_{k-1} + \|\mathcal{E}_{k-1}\|\,(1 + C_{\mathcal{P}} h_{k-1}) && (6.23)\\
&\qquad \le C_I\,(h_{k-1} + \hat{\delta} + c_{k-1}h_1), && (6.24)
\end{aligned}$$

where we assumed C_I sufficiently large so that $1 + C_{\mathcal{P}} h_{k-1} \le C_I$. To prove the second inequality (6.23) we used the fact that $\|\hat{\mathcal{P}}_{k-1}\| \le 1$ and the chain of inequalities $\|\mathcal{P}_{k-1}\| \le \|\hat{\mathcal{P}}_{k-1}\| + \|\mathcal{P}_{k-1} - \hat{\mathcal{P}}_{k-1}\| \le 1 + C_{\mathcal{P}} h_{k-1}$. The stability of $\hat{\mathcal{P}}_{k-1}$ results from Lemma 4.1 and the identity $(\hat{\mathcal{A}}_{k-1}\hat{\mathcal{P}}_{k-1}\mathbf{w}, \mathbf{v})_{k-1} = (\hat{\mathcal{A}}_k \mathbf{w}, \mathcal{I}_{k-1}^k \mathbf{v})_k$.

Collecting (6.19) through (6.21), and using (4.14), Lemma 6.4, Lemma 6.5, and (6.22)–(6.24), we see that

$$\begin{aligned}
\|\mathcal{E}_k - \hat{\mathcal{E}}_k\| &\le C_I\, C_{\mathcal{S}}\,(h_{k-1} + \hat{\delta} + c_{k-1}h_1)\, h_k \\
&\quad + C_{\mathcal{P}}\,(1 + \hat{\delta} + c_{k-1}h_1)\, h_{k-1} + c_{k-1}h_1 \\
&\le \left(\frac{C_I C_{\mathcal{S}}}{2} + C_{\mathcal{P}}\right) h_{k-1}\,(1 + \hat{\delta} + c_{k-1}h_1) + c_{k-1}h_1
\end{aligned}$$

for all k.

Now let $\hat{C} := \frac{C_I C_{\mathcal{S}}}{2} + C_{\mathcal{P}}$ and define

$$c_k := c_{k-1} + \hat{C}h_1^{-1}h_{k-1}(1 + \hat{\delta} + c_{k-1}h_1). \tag{6.25}$$

To see that the sequence c_k is uniformly bounded in k, one notes that $c_j \le c_k$ for $j \le k$ and hence

$$\begin{aligned}
c_k &= c_{k-1} + \hat{C}h_1^{-1}(1 + \hat{\delta} + c_{k-1}h_1)h_{k-1} \\
&= c_1 + \hat{C}h_1^{-1}\sum_{j=2}^{k}(1 + \hat{\delta} + c_{j-1}h_1)h_{j-1} \\
&\le c_1 + \hat{C}h_1^{-1}\sum_{j=2}^{k}(1 + \hat{\delta} + c_k h_1)h_{j-1} \\
&\le c_1 + 2\hat{C}(1 + \hat{\delta}) + 2\hat{C}h_1 c_k.
\end{aligned}$$

Now move the c_k term to the left to get

$$c_k \le (c_1 + 2\hat{C}(1 + \hat{\delta}))/(1 - 2\hat{C}h_1),$$

provided that h_1 is small enough. Therefore, if the coarsest grid is sufficiently fine, we have $\delta = \hat{\delta} + Ch_1 < 1$. ∎

We conclude this section with remarks on some of the constants appearing in the proofs. In case of collective CGS iteration, the constant $\bar{C}_{\hat{R}}$ appears to be almost independent from the value of the weight of the cost of the control. Its value is close to that of the Gauss–Seidel scheme applied to the scalar Poisson problem.

The constants in (6.16), (6.17), and (6.18) depend on the features of the optimality system as, for example, nonsymmetry. They account for the induction hypothesis where the coarsest mesh size, h_1, enters in the analysis and results in the estimate $\delta = \hat{\delta} + C\,h_1$. The requirement for a sufficiently small h_1 has no correspondence to our numerical experience (using CGS). However, the estimate of Theorem 6.6 states that, for sufficiently small h_1, we have $\delta \approx \hat{\delta}$, that is, the convergence factor of the multigrid method applied to the optimality system is close to the convergence factor of the multigrid scheme applied to the scalar Poisson problem. This fact agrees with our numerical experience and the results reported in Table 5.1.

7. Conclusions. We have presented a systematic study of a finite difference multigrid method for a class of optimality systems arising from optimal control of elliptic partial differential equations. In this emerging field of scientific computing there is an increasing interest in the development of accurate and efficient solution methods for optimal control problems. In the first part we have proved optimal-order error estimates in the discrete L^2 norm and in the discrete H^1 norm under minimum regularity requirements on the data. In the second part, two complementary analytical tools for multigrid convergence theory have been discussed. In the framework of local Fourier analysis it is possible to obtain sharp convergence estimates which are very important in the first phase of development of the multigrid components. The other analytical tool presented here is important from the theoretical point of view. It makes it possible to prove optimal convergence of the multigrid process under weak regularity assumptions. The general multigrid convergence theory discussed in this paper is developed in two steps. First, the multigrid method applied to the uncoupled differential system is considered. Then, the nondifferential coupling part characterizing the optimality system is introduced. By analyzing the difference between the operators obtained with and without coupling, we are able to estimate the convergence factor of multigrid for optimality systems based on the estimates available for the uncoupled problem.

REFERENCES

[1] E. Arian and S. Ta'asan, *Smoothers for optimization problems*, in Seventh Copper Mountain Conference on Multigrid Methods, Vol. CP3339, N. Duane Melson, T.A. Manteuffel, S.F. McCormick, and C.C. Douglas, eds., NASA Conference Publication, NASA, Hampton, VA, 1995, pp. 15–30.

[2] A. Borzì and K. Kunisch, *The numerical solution of the steady state solid fuel ignition model and its optimal control*, SIAM J. Sci. Comput., 22 (2000), pp. 263–284.

[3] A. Borzì and K. Kunisch, *A multigrid method for optimal control of time-dependent reaction diffusion processes*, in Fast Solution of Discretized Optimization Problems, Internat. Ser. Numer. Math. 138, K.H. Hoffmann, R. Hoppe, and V. Schulz, eds., Birkhäuser, Basel, 2001.

[4] A. Borzì, K. Kunisch, and M. Vanmaele, *A multi-grid approach to the optimal control of solid fuel ignition problems*, in Multigrid Methods, VI (Ghent, 1999), E. Dick, K. Riemslagh, and J. Vierendeels, eds., Lect. Notes Comput. Sci. Eng. 14, Springer–Verlag, Berlin, 2000, pp. 59–65.

[5] J.H. Bramble, *Multigrid Methods*, Pitman Res. Notes Math. Ser. 294, John Wiley, New York, 1993.

[6] J.H. Bramble, D. Y. Kwak, and J.E. Pasciak, *Uniform convergence of multigrid V-cycle iterations for indefinite and nonsymmetric problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1746–1763.
[7] J.H. Bramble and J.E. Pasciak, *New convergence estimates for multigrid algorithms*, Math. Comp., 49 (1987), pp. 311–329.
[8] J.H. Bramble and J.E. Pasciak, *The analysis of smoothers for multigrid algorithms*, Math. Comp., 58 (1992), pp. 467–488.
[9] J.H. Bramble, J.E. Pasciak, and J. Xu, *The analysis of multigrid algorithms with nonnested spaces or noninherited quadratic forms*, Math. Comp., 56 (1991), pp. 1–34.
[10] A. Brandt, *Rigorous quantitative analysis of multigrid,* I: *Constant coefficients two-level cycle with L_2-norm*, SIAM J. Numer. Anal., 31 (1994), pp. 1695–1730.
[11] S.H. Chou and D.Y. Kwak, *V-cycle multigrid for a vertex-centered covolume method for elliptic problems*, Numer. Math., 90 (2002), pp. 441–458.
[12] W. Hackbusch, *Fast solution of elliptic control problems*, J. Optim. Theory Appl., 31 (1980), pp. 565–581.
[13] W. Hackbusch, *Multi-Grid Methods and Applications*, Springer–Verlag, New York, 1985.
[14] W. Hackbusch, *Elliptic Differential Equations*, Springer–Verlag, New York, 1992.
[15] D.Y. Kwak, *V-cycle multigrid for cell-centered finite differences*, SIAM J. Sci. Comput., 21 (1999), pp. 552–564.
[16] D.Y. Kwak, *A general multigrid framework for a class of perturbed problems*, in Fluid Flow and Transport in Porous Media: Mathematical and Numerical Treatment, AMS, Providence, RI, 2002, pp. 317–325.
[17] J.L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer–Verlag, Berlin, 1971.
[18] P. Neittaanmäki and D. Tiba, *Optimal Control of Nonlinear Parabolic Systems*, Marcel Dekker, New York, 1994.
[19] V. Schulz and G. Wittum, *Multigrid optimization methods for stationary parameter identification problems in groundwater flow*, in Multigrid Methods V, Lect. Notes Comput. Sci. Eng. 3, W. Hackbusch and G. Wittum, eds., Springer–Verlag, Berlin, pp. 276–288.
[20] E. Süli, *Convergence of finite volume schemes for Poisson's equation on nonuniform meshes*, SIAM J. Numer. Anal., 28 (1991), pp. 1419–1430.
[21] U. Trottenberg, C. Oosterlee, and A. Schüller, *Multigrid*, Academic Press, London, 2001.
[22] S. Vanka, *Block-implicit multigrid calculation of two-dimensional recirculating flows*, Comp. Methods Appl. Mech. Engrg., 59 (1986), pp. 29–48.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 5, pp. 1498–1531

FEEDBACK CLASSIFICATION OF NONLINEAR SINGLE-INPUT CONTROL SYSTEMS WITH CONTROLLABLE LINEARIZATION: NORMAL FORMS, CANONICAL FORMS, AND INVARIANTS*

ISSA AMADOU TALL† AND WITOLD RESPONDEK†

Abstract. We study the feedback group action on single-input nonlinear control systems. We follow an approach of Kang and Krener based on analyzing, step by step, the action of homogeneous transformations on the homogeneous part of the same degree of the system. We construct a dual normal form and dual invariants with respect to those obtained by Kang. We also propose a canonical form and a dual canonical form and show that two systems are equivalent via a formal feedback if and only if their canonical forms (resp., their dual canonical forms) coincide. We give an explicit construction of transformations bringing the system to its normal, dual normal, canonical, and dual canonical forms. We illustrate our results by simple examples on $\mathbb{R}^3$ and $\mathbb{R}^4$.

Key words. feedback equivalence, normal forms, canonical forms, invariants

AMS subject classifications. 93B11, 93B17, 93B27

PII. S0363012900381753

1. Introduction.

The problem of transforming the nonlinear control single-input system

$$\Sigma \ : \ \dot{\xi} = f(\xi) + g(\xi)u$$

by a feedback transformation of the form

$$\Gamma \ : \quad \begin{array}{l} x = \phi(\xi), \\ u = \alpha(\xi) + \beta(\xi)v \end{array}$$

to a simpler form has been extensively studied during the last twenty years. The transformation Γ brings Σ to the system

$$\tilde{\Sigma} \ : \ \dot{x} = \tilde{f}(x) + \tilde{g}(x)v,$$

whose dynamics are given by

$$\begin{aligned} \tilde{f} &= \phi_*(f + g\alpha), \\ \tilde{g} &= \phi_*(g\beta), \end{aligned}$$

where for any vector field f and any diffeomorphism ϕ we denote

$$(\phi_* f)(x) = d\phi(\phi^{-1}(x)) \cdot f(\phi^{-1}(x)).$$

A natural question to ask is whether we can find a transformation Γ such that the

*Received by the editors November 28, 2000; accepted for publication (in revised form) April 21, 2002; published electronically January 14, 2003.
http://www.siam.org/journals/sicon/41-5/38175.html
†Institut National des Sciences Appliquées de Rouen, Laboratoire de Mathématiques de l'INSA, Pl. Emile Blondel, 76 131 Mont Saint Aignan Cedex, France (tall@insa-rouen.fr, wresp@insa-rouen.fr). The first author worked on this paper while on leave from the Department of Mathematics, Université Cheikh Anta Diop, Dakar, Senegal.

transformed system $\tilde{\Sigma}$ is linear, that is, whether we can linearize the system Σ via feedback. Necessary and sufficient geometric conditions for this to be the case have been given in [13] and [18]. Those conditions, except for the planar case, turn out to be restrictive, and a natural problem that arises is to find normal forms for nonlinearizable systems. Although natural, this problem is very involved and has been extensively studied during the last twenty years. Four basic methods have been proposed for studying feedback equivalence problems. The first method is based on the theory of singularities of vector fields and distributions, and their invariants, and using this method a large variety of feedback classification problems have been solved; see, e.g., [4], [7], [14], [15], [18], [19], [27], [29], [32], [38]. The second approach, proposed by Gardner [9], uses Cartan's method of equivalence [6] and describes the geometry of feedback equivalence [10], [11], [12], [28]. The third method, inspired by the Hamiltonian formalism for optimal control problems, was developed by Bonnard [3], [4] and Jakubczyk [16], [17] and has led to a very nice description of feedback invariants in terms of singular extremals. Finally, a very fruitful approach was proposed by Kang and Krener [26] and then followed by Kang [21], [22]. Their idea, which is closely related with Poincaré's classical technique for linearization of dynamical systems (see, e.g., [1]), is to analyze the system Σ and the feedback transformation Γ step by step and, as a consequence, to produce a simpler equivalent system $\tilde{\Sigma}$ also step by step.

Our paper is deeply inspired by those of Kang and Krener [26], [21] and can be considered as a completion of their results. In [21], Kang constructed a normal form for single-input nonlinear control systems with controllable linearization using successively homogeneous feedback transformations, and he proved that the homogeneous terms of a given degree of his normal form are unique under homogeneous feedback transformations of the same degree. He also showed that a nonlinear system can admit different normal forms under feedback resulting from the action of lower order terms of the feedback transformation on higher order terms of the system. The main goal of our paper is to propose a canonical form for the class of single-input systems with controllable linearization and to prove that two systems are equivalent, via a formal feedback, if and only if their canonical forms coincide.

In [26] Kang and Krener constructed two normal forms for the quadratic part of a single-input system. In the first normal form, all components of the linear part of the control vector field are annihilated and all nonremovable quadratic nonlinearities are grouped in the drift; in the second normal form, all quadratic terms of the drift are annihilated and all nonremovable nonlinearities are present in the control vector field. Kang normal form is a generalization, for higher order terms, of the first normal form. In this paper, we generalize the second one and produce a dual normal form for higher order terms. We also construct dual invariants of homogeneous feedback transformations. They contain the same information, as Kang invariants, encoded in a different way. We also give a dual canonical form and prove that two systems are equivalent, via formal feedback, if and only if their dual canonical forms coincide.

The third aim of the paper is to construct explicit homogeneous feedback transformations which bring the homogeneous part of the system of the same degree into its normal, or dual normal, form. For any fixed degree, our transformations are easily computable via differentiation and integration of polynomials. A successive application of those transformations gives formal feedbacks that bring any system to its normal form, dual normal form, canonical form, and dual canonical form.

The theory of normal forms initialized and developed by Kang and Krener [26] and Kang [21], [22] and continued in the present paper (and in [33], [34]) has proved to be very useful in analyzing structural properties of nonlinear control systems. It

has been used to study bifurcations of nonlinear systems [23], [24], [25], has led to a complete description of symmetries around equilibrium [30], [31], and has allowed us to characterize systems equivalent to feedforward forms [35], [36], [37].

The paper is organized as follows. In section 2 we will introduce, following [21] and [26], homogeneous feedback transformations. We give a normal form obtained by Kang and discuss invariants of homogeneous transformations, also obtained by him. We provide an explicit construction of transformations bringing the system to Kang normal form. In section 3 we construct a canonical form and give one of our main results stating that two control systems are feedback equivalent if and only if their canonical forms coincide. Proofs of results presented in sections 2 and 3 are given in section 4.

Section 5 dualizes the main results of section 2: we give a dual normal form, explicitly construct transformations bringing the system to that form, and define dual invariants of homogeneous transformations. Similarly to normal forms, a given system can admit different dual normal forms. In section 6 we thus dualize the results of section 3 by constructing a dual canonical form and proving that two control systems are feedback equivalent if and only if their dual canonical forms coincide. Section 7 contains proofs of results presented in sections 5 and 6. Throughout the paper, we illustrate our results by simple examples on $\mathbb{R}^3$ and $\mathbb{R}^4$.

2. Normal form and m-invariants. All objects, that is, functions, maps, vector fields, control systems, etc., are considered in a neighborhood of $0 \in \mathbb{R}^n$ and assumed to be C^∞-smooth. Let h be a smooth $\mathbb{R}$-valued function. By

$$h(x) = h^{[0]}(x) + h^{[1]}(x) + h^{[2]}(x) + \cdots = \sum_{m=0}^{\infty} h^{[m]}(x)$$

we denote its Taylor series expansion at $0 \in \mathbb{R}^n$, where $h^{[m]}(x)$ stands for a homogeneous polynomial of degree m.

Similarly, for a map ϕ of an open subset of $\mathbb{R}^n$ to $\mathbb{R}^n$ (resp., for a vector field f on an open subset of $\mathbb{R}^n$) we will denote by $\phi^{[m]}$ (resp., by $f^{[m]}$) the homogeneous term of degree m of its Taylor series expansion at $0 \in \mathbb{R}^n$, that is, each component $\phi_j^{[m]}$ of $\phi^{[m]}$ (resp., $f_j^{[m]}$ of $f^{[m]}$) is a homogeneous polynomial of degree m in x.

We will denote by $H^{[m]}(x)$ the space of homogeneous polynomials of degree m of the variables $x_1, \ldots, x_n$ and by $H^{\geq m}(x)$ the space of formal power series of the variables $x_1, \ldots, x_n$ starting from terms of degree m.

Analogously, we will denote by $R^{[m]}(x)$ the space of homogeneous vector fields whose components are in $H^{[m]}(x)$ and by $R^{\geq m}(x)$ the space of vector fields formal power series whose components are in $H^{\geq m}(x)$.

Consider the Taylor series expansion of the system Σ given by

$$\Sigma^\infty \ : \ \dot{\xi} = F\xi + Gu + \sum_{m=2}^{\infty} \left(f^{[m]}(\xi) + g^{[m-1]}(\xi)u \right), \tag{2.1}$$

where $F = \frac{\partial f}{\partial \xi}(0)$ and $G = g(0)$. We will assume throughout the paper that $f(0) = 0$ and $g(0) \neq 0$.

Consider also the Taylor series expansion Γ^∞ of the feedback transformation Γ

given by

$$\Gamma^\infty : \begin{aligned} x &= T\xi + \sum_{m=2}^{\infty} \phi^{[m]}(\xi), \\ u &= K\xi + Lv + \sum_{m=2}^{\infty} \left(\alpha^{[m]}(\xi) + \beta^{[m-1]}(\xi)v\right), \end{aligned} \tag{2.2}$$

where T is an invertible matrix and $L \neq 0$. Let us analyze the action of Γ^∞ on the system Σ^∞ step by step.

To start with, consider the linear system

$$\dot{\xi} = F\xi + Gu.$$

Throughout the paper we will assume that it is controllable. It can be thus transformed by a linear feedback transformation of the form

$$\Gamma^1 : \begin{aligned} x &= T\xi, \\ u &= K\xi + Lv \end{aligned}$$

to the Brunovský canonical form (A, B); see, e.g., [20]. Assuming that the linear part (F, G), of the system Σ^∞ given by (2.1), has been transformed to the Brunovský canonical form (A, B), we follow an idea of Kang and Krener [26], [21] and apply successively a series of transformations

$$\Gamma^m : \begin{aligned} x &= \xi + \phi^{[m]}(\xi), \\ u &= v + \alpha^{[m]}(\xi) + \beta^{[m-1]}(\xi)v \end{aligned} \tag{2.3}$$

for $m = 2, 3, \ldots$. A feedback transformation defined as a series of successive compositions of Γ^m, $m = 1, 2, \ldots$, will also be denoted by Γ^∞ because, as a formal power series, it is of the form (2.2). We will not address the problem of convergence and will call such a series of successive compositions a *formal feedback transformation*.

Observe that each transformation Γ^m for $m \geq 2$ leaves invariant all homogeneous terms of degree smaller than m of the system Σ^∞, and we will call Γ^m a *homogeneous feedback transformation of degree m*. We will study the action of Γ^m on the following *homogeneous system*:

$$\Sigma^{[m]} \ : \ \dot{\xi} = A\xi + Bu + f^{[m]}(\xi) + g^{[m-1]}(\xi)u. \tag{2.4}$$

Consider another homogeneous system, $\tilde{\Sigma}^{[m]}$, given by

$$\tilde{\Sigma}^{[m]} \ : \ \dot{x} = Ax + Bv + \tilde{f}^{[m]}(x) + \tilde{g}^{[m-1]}(x)v. \tag{2.5}$$

We will say that the homogeneous system $\Sigma^{[m]}$ is feedback equivalent to the homogeneous system $\tilde{\Sigma}^{[m]}$ if there exists a homogeneous feedback transformation of the form (2.3), which brings $\Sigma^{[m]}$ into $\tilde{\Sigma}^{[m]}$ modulo terms in $R^{\geq m+1}(x, v)$.

Notation. Because of various normal forms and various transformations that are used throughout the paper, we will keep the following notation. We will denote, respectively, by $\Sigma^{[m]}$ and Σ^{∞} the following systems:

$$\begin{aligned}\Sigma^{[m]} &: \dot{\xi} = A\xi + Bu + f^{[m]}(\xi) + g^{[m-1]}(\xi)u,\\ \Sigma^{\infty} &: \dot{\xi} = A\xi + Bu + \sum_{k=2}^{\infty}\left(f^{[k]}(\xi) + g^{[k-1]}(\xi)u\right).\end{aligned}$$

The systems $\Sigma^{[m]}$ and Σ^{∞} will stand for the systems under consideration. Their state vector will be denoted by ξ and their control by u. The system $\Sigma^{[m]}$ (resp., the system Σ^{∞}) transformed via feedback will be denoted by $\tilde{\Sigma}^{[m]}$ (resp., by $\tilde{\Sigma}^{\infty}$). Its state vector will be denoted by x, its control by v, and the vector fields, defining its dynamics, by $\tilde{f}^{[k]}$ and $\tilde{g}^{[k-1]}$. Feedback equivalence of homogeneous systems $\Sigma^{[m]}$ and $\tilde{\Sigma}^{[m]}$ will be established via a smooth feedback, that is, precisely, via a homogeneous feedback Γ^m. On the other hand, feedback equivalence of systems Σ^{∞} and $\tilde{\Sigma}^{\infty}$ will be established via a formal feedback Γ^{∞}.

We will introduce two kinds of normal forms, Kang normal forms and dual normal forms, as well as canonical forms and dual canonical forms. The "bar" symbol will correspond to the vector field $\bar{f}^{[m]}$ defining the Kang normal forms $\Sigma_{NF}^{[m]}$ and Σ_{NF}^{∞} and the canonical form Σ_{CF}^{∞} as well as to the vector field $\bar{g}^{[m-1]}$ defining the dual normal forms $\Sigma_{DNF}^{[m]}$ and Σ_{DNF}^{∞} and the dual canonical form Σ_{DCF}^{∞}. Analogously, the m-invariants (resp., dual m-invariants) of the system $\Sigma^{[m]}$ will be denoted by $a^{[m]j,i+2}$ (resp., by $b_j^{[m-1]}$) and the m-invariants (resp., dual m-invariants) of the normal form $\Sigma_{NF}^{[m]}$ (resp., dual normal form $\Sigma_{DNF}^{[m]}$) by $\bar{a}^{[m]j,i+2}$ (resp., by $\bar{b}_j^{[m-1]}$).

The starting point is the following result, proved by Kang [21].

PROPOSITION 1. *The homogeneous feedback transformation Γ^m, defined by* (2.3)*, brings the system $\Sigma^{[m]}$, given by* (2.4)*, into $\tilde{\Sigma}^{[m]}$, given by* (2.5)*, if and only if the relations*

$$\begin{cases} L_{A\xi}\phi_j^{[m]}(\xi) - \phi_{j+1}^{[m]}(\xi) = \tilde{f}_j^{[m]}(\xi) - f_j^{[m]}(\xi),\\ L_B\phi_j^{[m]}(\xi) = \tilde{g}_j^{[m-1]}(\xi) - g_j^{[m-1]}(\xi),\\ L_{A\xi}\phi_n^{[m]}(\xi) + \alpha^{[m]}(\xi) = \tilde{f}_n^{[m]}(\xi) - f_n^{[m]}(\xi),\\ L_B\phi_n^{[m]}(\xi) + \beta^{[m-1]}(\xi) = \tilde{g}_n^{[m-1]}(\xi) - g_n^{[m-1]}(\xi)\end{cases} \tag{2.6}$$

hold for any $1 \le j \le n-1$, where $\phi_j^{[m]}$ are the components of $\phi^{[m]}$.

This proposition represents the essence of the method developed by Kang and Krener and used in our paper. The problem of studying the feedback equivalence of two systems Σ and $\tilde{\Sigma}$ requires, in general, solving a system of first order partial differential equations. On the other hand, if we perform the analysis step by step, then the problem of establishing the feedback equivalence of two systems $\Sigma^{[m]}$ and $\tilde{\Sigma}^{[m]}$ reduces to solving the algebraic system (2.6).

Using the above proposition, Kang [21] proved the following result.

THEOREM 1. *The homogeneous system $\Sigma^{[m]}$ can be transformed, via a homogeneous feedback transformation Γ^m, into the normal form*

$$\Sigma_{NF}^{[m]} \; : \; \begin{cases} \dot{x}_1 = x_2 + \sum_{i=3}^{n} x_i^2 P_{1,i}^{[m-2]}(x_1, \dots, x_i), \\ \qquad \vdots \\ \dot{x}_j = x_{j+1} + \sum_{i=j+2}^{n} x_i^2 P_{j,i}^{[m-2]}(x_1, \dots, x_i), \\ \qquad \vdots \\ \dot{x}_{n-2} = x_{n-1} + x_n^2 P_{n-2,n}^{[m-2]}(x_1, \dots, x_n), \\ \dot{x}_{n-1} = x_n, \\ \dot{x}_n = v, \end{cases} \tag{2.7}$$

where $P_{j,i}^{[m-2]}(x_1, \dots, x_i)$ *are homogeneous polynomials of degree* $m-2$ *depending on the indicated variables.*

In order to construct invariants of homogeneous feedback transformations, let us define

$$X_i^{m-1}(\xi) = (-1)^i ad^i_{A\xi + f^{[m]}(\xi)}(B + g^{[m-1]}(\xi))$$

and let $X_i^{[m-1]}$ be its homogeneous part of degree $m-1$. By π_i we will denote the projection on the subspace

$$W_i = \{\xi \in R^n : \quad \xi_{i+1} = \cdots = \xi_n = 0\}, \tag{2.8}$$

that is,

$$\pi_i(\xi) = (\xi_1, \dots, \xi_i, 0, \dots, 0).$$

Following Kang [21], we denote by $a^{[m]j,i+2}(\xi)$ the homogeneous part of degree $m-2$ of the polynomials

$$CA^{j-1}\left[X_i^{m-1}, X_{i+1}^{m-1}\right](\pi_{n-i}(\xi)),$$

where $C = (1, 0, \dots, 0)^T \in \mathbb{R}^n$ and $(j,i) \in \Delta \subset \mathbb{N} \times \mathbb{N}$, defined by

$$\Delta = \{(j,i) \in \mathbb{N} \times \mathbb{N} : \; 1 \le j \le n-2, \; 0 \le i \le n-j-2\}.$$

The homogeneous polynomials $a^{[m]j,i+2}$ for $(j,i) \in \Delta$ will be called *m-invariants* of $\Sigma^{[m]}$.

The following result of Kang [21] asserts that m-invariants $a^{[m]j,i+2}$ for $(j,i) \in \Delta$ are complete invariants of homogeneous feedback and, moreover, illustrates their meaning for the homogeneous normal form $\Sigma_{NF}^{[m]}$.

Consider two homogeneous systems $\Sigma^{[m]}$ and $\tilde{\Sigma}^{[m]}$ and let

$$\{\, a^{[m]j,i+2} : \; (j,i) \in \Delta \,\}$$

and

$$\{\, \tilde{a}^{[m]j,i+2} : \; (j,i) \in \Delta \,\}$$

denote, respectively, their m-invariants. The following theorem was proved by Kang [21].

THEOREM 2. *The m-invariants have the following properties:*

(i) *Two homogeneous systems* $\Sigma^{[m]}$ *and* $\tilde{\Sigma}^{[m]}$ *are equivalent via a homogeneous feedback transformation* Γ^m *if and only if*

$$a^{[m]j,i+2} = \tilde{a}^{[m]j,i+2}$$

for any $(j,i) \in \Delta$.

(ii) *The m-invariants* $\bar{a}^{[m]j,i+2}$ *of the homogeneous normal form* $\Sigma^{[m]}_{NF}$, *defined by* (2.7), *are given by*

$$\bar{a}^{[m]j,i+2}(x) = \frac{\partial^2}{\partial x_{n-i}^2} x_{n-i}^2 P^{[m-2]}_{j,n-i}(x_1, \ldots, x_{n-i}) \tag{2.9}$$

for any $(j,i) \in \Delta$.

Our first aim is to find explicitly feedback transformations bringing the homogeneous system $\Sigma^{[m]}$ to its normal form $\Sigma^{[m]}_{NF}$. Define the homogeneous polynomials $\psi^{[m-1]}_{j,i}(\xi)$ by setting $\psi^{[m-1]}_{j,0}(\xi) = \psi^{[m-1]}_{1,1}(\xi) = 0$,

$$\psi^{[m-1]}_{j,i}(\xi) = -CA^{j-1}\left(ad^{n-i}_{A\xi} g^{[m-1]} + \sum_{t=1}^{n-i} (-1)^t ad^{t-1}_{A\xi} ad_{A^{n-i-t}B} f^{[m]} \right) \tag{2.10}$$

if $1 \le j < i \le n$ and

$$\begin{aligned} \psi^{[m-1]}_{j,i}(\xi) = {} & L_{A^{n-i}B} f^{[m]}_{j-1}(\pi_i(\xi)) + L_{A\xi} \psi^{[m-1]}_{j-1,i}(\pi_i(\xi)) \\ & + \psi^{[m-1]}_{j-1,i-1}(\pi_{i-1}(\xi)) + \int_0^{\xi_i} L_{A^{n-i+1}B} \psi^{[m-1]}_{j-1,i}(\pi_i(\xi)) d\xi_i \end{aligned} \tag{2.11}$$

if $1 \le i \le j$, where $\psi^{[m-1]}_{j,i}(\pi_i(\xi))$ is the restriction of $\psi^{[m-1]}_{j,i}(\xi)$ to the submanifold W_i. Define the components $\phi^{[m]}_j$ of $\phi^{[m]}$ for $1 \le j \le n$ and the feedback $(\alpha^{[m]}, \beta^{[m-1]})$ by

$$\begin{aligned} \phi^{[m]}_j(\xi) &= \sum_{i=1}^{n} \int_0^{\xi_i} \psi^{[m-1]}_{j,i}(\pi_i(\xi)) d\xi_i, \qquad 1 \le j \le n-1, \\ \phi^{[m]}_n(\xi) &= f^{[m]}_{n-1}(\xi) + L_{A\xi} \phi^{[m]}_{n-1}(\xi), \\ \alpha^{[m]}(\xi) &= -\left(f^{[m]}_n(\xi) + L_{A\xi} \phi^{[m]}_n(\xi) \right), \\ \beta^{[m-1]}(\xi) &= -\left(g^{[m-1]}_n(\xi) + L_B \phi^{[m]}_n(\xi) \right). \end{aligned} \tag{2.12}$$

We have the following result.

THEOREM 3. *The homogeneous feedback transformation*

$$\Gamma^m \ : \ \begin{aligned} x &= \xi + \phi^{[m]}(\xi), \\ u &= v + \alpha^{[m]}(\xi) + \beta^{[m-1]}(\xi) v, \end{aligned}$$

where $\alpha^{[m]}$, $\beta^{[m-1]}$, *and the components* $\phi^{[m]}_j$ *of* $\phi^{[m]}$ *are defined by* (2.12), *brings the homogeneous system* $\Sigma^{[m]}$ *into its normal form* $\Sigma^{[m]}_{NF}$ *given by* (2.7).

Proof of Theorem 3. Denote by

$$\tilde{\Sigma}^{[m]} \ : \ \dot{x} = Ax + Bv + \tilde{f}^{[m]}(x) + \tilde{g}^{[m-1]}(x) v$$

the system $\Sigma^{[m]}$ transformed via the feedback transformation Γ^m defined by (2.12).

From the expressions of $\alpha^{[m]}(\xi)$ and $\beta^{[m-1]}(\xi)$ given by (2.12) and the last two equations of (2.6), we get

$$\tilde{f}_n^{[m]}(x) = 0 \qquad \text{and} \qquad \tilde{g}_n^{[m-1]}(x) = 0.$$

Plugging $\phi_j^{[m]}$, defined by (2.12), into the second equation of (2.6) gives

$$\psi_{j,n}^{[m-1]}(x) = \tilde{g}_j^{[m-1]}(x) - g_j^{[m-1]}(x),$$

which, by (2.10), implies $\tilde{g}_j^{[m-1]}(x) = 0$ for $1 \le j \le n-1$. Now we consider the first equation of (2.6). From the expression of $\phi_n^{[m]}$ we get $\tilde{f}_{n-1}^{[m]}(x) = 0$, and for any $1 \le i \le n$, we obtain by differentiating

$$\frac{\partial \tilde{f}_j^{[m]}}{\partial x_i} = \frac{\partial f_j^{[m]}}{\partial x_i} + L_{Ax}\frac{\partial \phi_j^{[m]}}{\partial x_i} + \frac{\partial \phi_j^{[m]}}{\partial x_{i-1}} - \frac{\partial \phi_{j+1}^{[m]}}{\partial x_i}. \tag{2.13}$$

In the above formula, the term $\frac{\partial \phi_j^{[m]}}{\partial x_{i-1}}$ is not present in the case $i = 1$.

If $i \ge j+1$, we get

$$\begin{aligned}\frac{\partial \tilde{f}_j^{[m]}}{\partial x_i}(\pi_{i-1}(x)) &= \left(\frac{\partial f_j^{[m]}}{\partial x_i} + L_{Ax}\frac{\partial \phi_j^{[m]}}{\partial x_i} + \frac{\partial \phi_j^{[m]}}{\partial x_{i-1}} - \frac{\partial \phi_{j+1}^{[m]}}{\partial x_i}\right)(\pi_{i-1}(x)) \\ &= \frac{\partial f_j^{[m]}}{\partial x_i}(\pi_{i-1}(x)) + L_{Ax}\psi_{j,i}^{[m-1]}(\pi_{i-1}(x)) \\ &\quad + \psi_{j,i-1}^{[m-1]}(\pi_{i-1}(x)) - \psi_{j+1,i}^{[m-1]}(\pi_{i-1}(x)).\end{aligned}$$

Hence, by an induction argument, we obtain

$$\frac{\partial f_j^{[m]}}{\partial x_i}(\pi_{i-1}(x)) + L_{Ax}\psi_{j,i}^{[m-1]}(\pi_{i-1}(x)) + \psi_{j,i-1}^{[m-1]}(\pi_{i-1}(x)) - \psi_{j+1,i}^{[m-1]}(\pi_{i-1}(x)) = 0$$

and, finally, we get

$$\frac{\partial \tilde{f}_j^{[m]}}{\partial x_i}(\pi_{i-1}(x)) = 0. \tag{2.14}$$

If $1 \le i \le j$, then, using (2.12) and (2.13), we obtain

$$\begin{aligned}\frac{\partial \tilde{f}_j^{[m]}}{\partial x_i}(\pi_i(x)) &= \left(\frac{\partial f_j^{[m]}}{\partial x_i} + L_{Ax}\frac{\partial \phi_j^{[m]}}{\partial x_i} + \frac{\partial \phi_j^{[m]}}{\partial x_{i-1}} - \frac{\partial \phi_{j+1}^{[m]}}{\partial x_i}\right)(\pi_i(x)) \\ &= \frac{\partial f_j^{[m]}}{\partial x_i}(\pi_i(x)) + L_{Ax}\psi_{j,i}^{[m-1]}(\pi_i(x)) + \psi_{j,i-1}^{[m-1]}(\pi_{i-1}(x)) \\ &\quad + \int_0^{x_i} \frac{\partial \psi_{j,i}^{[m-1]}(\pi_i(x))}{\partial x_{i-1}} dx_i - \psi_{j+1,i}^{[m-1]}(\pi_i(x)).\end{aligned}$$

Using the expression (2.11), it follows that

$$\frac{\partial \tilde{f}_j^{[m]}}{\partial x_i}(\pi_i(x)) = 0. \tag{2.15}$$

From the relations (2.14) and (2.15), we conclude that

$$\tilde{f}_j^{[m]}(x) = \sum_{i=j+2}^{n} x_i^2 P_{j,i}^{[m-2]}(x_1, \dots, x_i),$$

which proves that $\tilde{\Sigma}^{[m]}$ is a normal form satisfying (2.7). Thus the system $\Sigma^{[m]}$ given by (2.4) is feedback equivalent to the normal form $\Sigma_{NF}^{[m]}$ given by (2.7). □

Example 1. To illustrate results of this section, we consider the system $\Sigma^{[m]}$, given by (2.4) on $\mathbb{R}^3$. Theorem 1 implies that the system $\Sigma^{[m]}$ is equivalent, via a homogeneous feedback transformation Γ^m defined by (2.12), to its normal form $\Sigma_{NF}^{[m]}$ (see (2.7))

$$\begin{aligned} \dot{x}_1 &= x_2 + x_3^2 P^{[m-2]}(x_1, x_2, x_3), \\ \dot{x}_2 &= x_3, \\ \dot{x}_3 &= v, \end{aligned}$$

where $P^{[m-2]}(x_1, x_2, x_3)$ is a homogeneous polynomial of degree $m-2$ of the variables x_1, x_2, x_3. □

We would like now to discuss the interest of Theorem 3. As we have already mentioned, Poincaré's method allows us to replace a partial differential equation by solving successively linear algebraic equations defined by the homological equation (2.6); see [26] and [21], and Proposition 1. The solvability of this equation was proved in [26] and [21], while Theorem 3 provides an explicit solution (in the form of the transformations (2.12), which are easily computable via differentiation and integration of homogeneous polynomials) to the homological equation. As a consequence, for any given control system, Theorem 3 gives transformations bringing the homogeneous part of the system to its normal form. For example, if the system is feedback linearizable, up to order $m_0 - 1$ (see [27]), then a diffeomorphism and a feedback compensating all nonlinearities of degree lower than m_0 can be calculated explicitly without solving partial differential equations. More generally, by a successive application of transformations given by (2.12) we can bring the system, without solving partial differential equations, to its normal form given in Theorem 4 below.

Consider the system Σ^∞ of the form (2.1) and recall that we assume the linear part (F, G) to be controllable. Apply successively to Σ^∞ a series of transformations Γ^m, $m = 1, 2, \dots$, such that each Γ^m brings $\Sigma^{[m]}$ to its normal form $\Sigma_{NF}^{[m]}$; for instance we can take a series of transformations defined by (2.12). Successive repeating of Theorem 1 gives the following result of Kang [21].

THEOREM 4. *There exists a formal feedback transformation* Γ^∞ *which brings the*

system Σ^∞ *to a normal form* Σ^∞_{NF} *given by*

$$\Sigma^\infty_{NF}\ :\ \begin{cases} \dot{x}_1 = x_2 + \displaystyle\sum_{i=3}^{n} x_i^2 P_{1,i}(x_1,\dots,x_i), \\ \qquad\vdots \\ \dot{x}_j = x_{j+1} + \displaystyle\sum_{i=j+2}^{n} x_i^2 P_{j,i}(x_1,\dots,x_i), \\ \qquad\vdots \\ \dot{x}_{n-2} = x_{n-1} + x_n^2 P_{n-2,n}(x_1,\dots,x_n), \\ \dot{x}_{n-1} = x_n, \\ \dot{x}_n = v, \end{cases} \tag{2.16}$$

where $P_{j,i}(x_1,\dots,x_i)$ *are formal power series depending on the indicated variables.*

Example 2. Consider a system Σ defined on $\mathbb{R}^3$ whose linear part is controllable. Theorem 4 implies that the system Σ is equivalent, via a formal feedback transformation Γ^∞, to its normal form Σ^∞_{NF}

$$\begin{aligned} \dot{x}_1 &= x_2 + x_3^2 P(x_1,x_2,x_3), \\ \dot{x}_2 &= x_3, \\ \dot{x}_3 &= v, \end{aligned}$$

where $P(x_1,x_2,x_3)$ is a formal power series of the variables x_1,x_2,x_3. □

3. Canonical form. As proved by Kang and recalled in Theorem 2, the normal form $\Sigma^{[m]}_{NF}$ is unique under homogeneous feedback transformation Γ^m. The normal form Σ^∞_{NF} is constructed by a successive application of homogeneous transformations Γ^m for $m \geq 1$ which bring the corresponding homogeneous systems $\Sigma^{[m]}$ into their normal forms $\Sigma^{[m]}_{NF}$. Therefore a natural and fundamental question which arises is whether the system Σ^∞ can admit two different normal forms, that is, whether the normal forms given by Theorem 4 are in fact canonical forms under a general formal feedback transformations of the form Γ^∞. It turns out that a given system can admit different normal forms, as the following example of Kang [21] shows. The main reason for the nonuniqueness of the normal form Σ^∞_{NF} is that, although the normal form $\Sigma^{[m]}_{NF}$ is unique, homogeneous feedback transformation Γ^m bringing $\Sigma^{[m]}$ into $\Sigma^{[m]}_{NF}$ is not. It is this small group of homogeneous feedback transformations of order m that preserve $\Sigma^{[m]}_{NF}$ (described by Proposition 2 below), which causes the nonuniqueness of Σ^∞_{NF}.

The aim of this section is thus to construct a canonical form for Σ^∞ under feedback transformation Γ^∞.

Example 3. Consider the system

$$\begin{aligned} \dot{\xi}_1 &= \xi_2 + \xi_3^2 - 2\xi_1\xi_3^2, \\ \dot{\xi}_2 &= \xi_3, \\ \dot{\xi}_3 &= u \end{aligned} \tag{3.1}$$

on $\mathbb{R}^3$. Clearly, this system is in Kang normal form (compare with Theorem 4). The

feedback transformation

$$\Gamma^{\leq 3}\ :\ \begin{aligned} x_1 &= \xi_1 - \xi_1^2 - \tfrac{4}{3}\xi_2^3,\\ x_2 &= \xi_2 - 2\xi_1\xi_2,\\ x_3 &= \xi_3 - 2(\xi_2^2 + \xi_1\xi_3) - 2\xi_2\xi_3^2,\\ u &= v + 6\xi_2\xi_3 + 12\xi_1\xi_2\xi_3 - 4\xi_3^3 + 2(\xi_1 + 2\xi_1^2 + 2\xi_2\xi_3)v \end{aligned}$$

brings the system (3.1) into the form

$$\begin{aligned} \dot{x}_1 &= x_2 + x_3^2,\\ \dot{x}_2 &= x_3,\\ \dot{x}_3 &= v \end{aligned}$$

modulo terms in $R^{\geq 4}(x, v)$. Applying successively homogeneous feedback transformations Γ^m given, for any $m \geq 4$, by (2.12), we transform the above system into the normal form

$$\begin{aligned} \dot{x}_1 &= x_2 + x_3^2 + x_3^2 P(x),\\ \dot{x}_2 &= x_3,\\ \dot{x}_3 &= v, \end{aligned} \tag{3.2}$$

where P is a formal power series whose 1-jet at $0 \in \mathbb{R}^3$ vanishes. The systems (3.1) and (3.2) are in their normal forms and, moreover, feedback equivalent, but the latter system does not contain any term of degree 3. As a consequence, the normal form Σ^{∞}_{NF} is not unique under formal feedback transformations. □

Consider the system Σ^{∞} of the form (2.1). Since its linear part (F, G) is assumed to be controllable, we bring it, via a linear transformation and linear feedback, to the Brunovský canonical form (A, B). Let the first homogeneous term of Σ^{∞} which cannot be annihilated by a feedback transformation be of degree m_0. As proved by Krener [27], the degree m_0 is given by the largest integer p such that all distributions $\mathcal{D}^k = \operatorname{span}\{g, \dots, ad_f^{k-1} g\}$ for $1 \leq k \leq n-1$ are involutive modulo terms of order $p-2$. We can thus, due to Theorems 1 and 2, assume that, after applying a suitable feedback, Σ^{∞} takes the form

$$\dot{\xi} = A\xi + Bu + \bar{f}^{[m_0]}(\xi) + \sum_{m=m_0+1}^{\infty} \left(f^{[m]}(\xi) + g^{[m-1]}(\xi)u \right),$$

where (A, B) is in Brunovský canonical form and the first nonvanishing homogeneous vector field $\bar{f}^{[m_0]}$ is of the form

$$\bar{f}_j^{[m_0]}(\xi) = \begin{cases} \displaystyle\sum_{i=j+2}^{n} \xi_i^2 P_{j,i}^{[m_0-2]}(\xi_1, \dots, \xi_i), & 1 \leq j \leq n-2,\\ 0, & n-1 \leq j \leq n. \end{cases}$$

Let $(i_1, \dots, i_{n-s})$, where $i_1 + \cdots + i_{n-s} = m_0$ and $i_{n-s} \geq 2$, be the largest, in the lexicographic ordering, $(n-s)$-tuple of nonnegative integers such that for some $1 \leq j \leq n-2$, we have

$$\frac{\partial^{m_0} \bar{f}_j^{[m_0]}}{\partial \xi_1^{i_1} \cdots \partial \xi_{n-s}^{i_{n-s}}} \neq 0.$$

Define

$$j^* = \sup\left\{ j = 1, \dots, n-2 \ : \ \frac{\partial^{m_0} \bar{f}_j^{[m_0]}}{\partial \xi_1^{i_1} \cdots \partial \xi_{n-s}^{i_{n-s}}} \neq 0 \right\}.$$

We have the following result.

THEOREM 5. *The system* Σ^∞ *given by* (2.1) *is equivalent by a formal feedback* Γ^∞ *to a system of the form*

$$\Sigma_{CF}^\infty \ : \ \dot{x} = Ax + Bv + \sum_{m=m_0}^{\infty} \bar{f}^{[m]}(x), \tag{3.3}$$

where, for any $m \geq m_0$,

$$\bar{f}_j^{[m]}(x) = \begin{cases} \displaystyle\sum_{i=j+2}^{n} x_i^2 P_{j,i}^{[m-2]}(x_1, \dots, x_i), & 1 \leq j \leq n-2, \\ 0, & n-1 \leq j \leq n; \end{cases} \tag{3.4}$$

additionally, we have

$$\frac{\partial^{m_0} \bar{f}_{j^*}^{[m_0]}}{\partial x_1^{i_1} \cdots \partial x_{n-s}^{i_{n-s}}} = \pm 1 \tag{3.5}$$

and, moreover, for any $m \geq m_0 + 1$,

$$\frac{\partial^{m_0} \bar{f}_{j^*}^{[m]}}{\partial x_1^{i_1} \cdots \partial x_{n-s}^{i_{n-s}}}(x_1, 0, \dots, 0) = 0. \tag{3.6}$$

The form Σ_{CF}^∞ satisfying (3.4), (3.5), and (3.6) will be called the *canonical form* of Σ^∞. The name is justified by the following.

THEOREM 6. *Two systems* Σ_1^∞ *and* Σ_2^∞ *are formally feedback equivalent if and only if their canonical forms* $\Sigma_{1,CF}^\infty$ *and* $\Sigma_{2,CF}^\infty$ *coincide.*

Proofs of Theorems 5 and 6 are given in section 4.

Kang [21], generalizing [26], proved that any system Σ^∞ can be brought by a formal feedback into the normal form (3.3) for which (3.4) is satisfied. He also observed that his normal forms are not unique; see Example 3. Our results, Theorems 5 and 6, complete his study. We show that for each degree m of homogeneity we can use a one-dimensional subgroup of feedback transformations which preserves the "triangular" structure of (3.4) and at the same time allows us to normalize one higher order term. The form of (3.5) and (3.6) is a result of this normalization. These one-dimensional subgroups of feedback transformations are given by the following proposition.

PROPOSITION 2. *The transformation* Γ^m *given by* (2.3) *leaves invariant the system* $\Sigma^{[m]}$ *defined by* (2.4) *if and only if*

$$\begin{aligned} \phi_j^{[m]} &= a_m L_{A\xi}^{j-1} \xi_1^m, \qquad 1 \leq j \leq n, \\ \alpha^{[m]} &= -a_m L_{A\xi}^n \xi_1^m, \\ \beta^{[m-1]} &= -a_m L_B L_{A\xi}^{n-1} \xi_1^m, \end{aligned} \tag{3.7}$$

where a_m is an arbitrary real parameter.

Proof of Proposition 2. Observe that, following Proposition 1, the transformation Γ^m leaves invariant the system $\Sigma^{[m]}$ if and only if it satisfies the following system of equations:

$$\begin{cases} L_{A\xi}\phi_j^{[m]} - \phi_{j+1}^{[m]}(\xi) = 0, & 1 \le j \le n-1, \\ L_B\phi_j^{[m]} = 0, & 1 \le j \le n-1, \\ L_{A\xi}\phi_n^{[m]} + \alpha^{[m]}(\xi) = 0, \\ L_B\phi_n^{[m]} + \beta^{[m-1]}(\xi) = 0. \end{cases}$$

In order to solve the above system, let us remark, using the second equation of the system, that for any j such that $1 \le j \le n-1$, the component $\phi_j^{[m]}$ does not depend to the variable ξ_n. Putting $j = n-2$ into the first equation, we get

$$\frac{\partial \phi_{n-2}^{[m]}}{\partial \xi_1}\xi_2 + \cdots + \frac{\partial \phi_{n-2}^{[m]}}{\partial \xi_{n-1}}\xi_n = \phi_{n-1}^{[m]}.$$

Since $\phi_{n-1}^{[m]}$ and $\phi_{n-2}^{[m]}$ do not depend on the variable ξ_n, we conclude that $\phi_{n-2}^{[m]}$ does not depend on the variable ξ_{n-1}. An inductive argument shows that $\phi_1^{[m]}$ depends only on the variable ξ_1, that is, $\phi_1^{[m]}(\xi) = a_m\xi_1^m$. Now, all equations of (3.7) follow easily. □

Theorem 5 establishes an equivalence of the system Σ^∞ with its canonical form Σ_{CF}^∞ via a formal feedback. Its direct corollary yields the following result for equivalence under a smooth feedback of the form

$$\Gamma \ : \quad \begin{array}{l} x = \phi(\xi), \\ u = \alpha(\xi) + \beta(\xi)v, \end{array}$$

up to an arbitrary order.

Corollary 1. *Consider a smooth control system*

$$\Sigma \ : \ \dot{\xi} = f(\xi) + g(\xi)u.$$

For any positive integer k we have the following:

(i) *There exists a smooth feedback Γ transforming Σ, locally around $0 \in \mathbb{R}^n$, into its canonical form $\Sigma_{CF}^{\le k}$ given by*

$$\Sigma_{CF}^{\le k} \ : \ \dot{x} = Ax + Bv + \sum_{m=m_0}^{k} \bar{f}^{[m]}(x),$$

modulo $O(x,v)^{k+1}$, where $\bar{f}^{[m]}(x)$, for any $m_0 \le m \le k$, satisfies (3.4), (3.5), (3.6).

(ii) *Feedback equivalence of Σ and $\Sigma_{CF}^{\le k}$, modulo $O(x,v)^{k+1}$, can be established via a polynomial feedback transformation $\Gamma^{\le k}$ of degree k.*

(iii) *Two smooth systems Σ_1 and Σ_2 are feedback equivalent modulo terms of order $O(x,v)^{k+1}$ if and only if their canonical forms $\Sigma_{1,CF}^{\le k}$ and $\Sigma_{2,CF}^{\le k}$ coincide.*

This corollary follows directly from Theorem 5 and its proof, given in section 4, which provides explicit polynomial transformations (4.4)–(4.5) bringing, step by step, the system into its canonical form.

We will illustrate results of this section by two examples.

Example 4. Let us reconsider the system Σ given by Example 2. It is feedback equivalent to the normal form

$$\begin{aligned}\dot{x}_1 &= x_2 + x_3^2 P(x_1, x_2, x_3),\\ \dot{x}_2 &= x_3,\\ \dot{x}_3 &= v,\end{aligned}$$

where $P(x_1, x_2, x_3)$ is a formal power series. Assume, for simplicity, that $m_0 = 2$, which is equivalent to the following generic condition: g, $ad_f g$, and $[g, ad_f g]$ are linearly independent at $0 \in \mathbb{R}^3$. This implies that we can express $P = P(x_1, x_2, x_3)$ as

$$P = c + P_1(x_1) + x_2 P_2(x_1, x_2) + x_3 P_3(x_1, x_2, x_3),$$

where $c \neq 0$ and $P_1(0) = 0$. Observe that any $P(x_1, x_2, x_3)$, of the above form, gives a normal form Σ_{NF}^{∞}. In order to get the canonical form Σ_{CF}^{∞}, we use Theorem 5, which ensures the existence of a feedback transformation Γ^{∞} of the form

$$\begin{aligned}\tilde{x} &= \phi(x),\\ v &= \alpha(x) + \beta(x)\tilde{v},\end{aligned}$$

which normalizes the constant c and annihilates the formal power series $P_1(x_1)$. More precisely, Γ^{∞} transforms Σ into its canonical form Σ_{CF}^{∞},

$$\begin{aligned}\dot{\tilde{x}}_1 &= \tilde{x}_2 + \tilde{x}_3^2 \tilde{P}(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3),\\ \dot{\tilde{x}}_2 &= \tilde{x}_3,\\ \dot{\tilde{x}}_3 &= \tilde{v},\end{aligned}$$

where the formal power series $\tilde{P}(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3)$ is of the form

$$\tilde{P}(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) = 1 + \tilde{x}_2 \tilde{P}_2(\tilde{x}_1, \tilde{x}_2) + \tilde{x}_3 \tilde{P}_3(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3).$$

Now, we give an example of constructing the canonical form for a physical model of variable length pendulum. □

Example 5. Consider the variable length pendulum of Bressan and Rampazzo [5] (see also [2] and [8]). We denote by ξ_1 the length of the pendulum, by ξ_2 its velocity, by ξ_3 the angle with respect to the horizontal, and by ξ_4 the angular velocity. The control $u = \dot{\xi}_4 = \ddot{\xi}_3$ is the angular acceleration. The mass is normalized to 1. The equations are (compare [5] and [8])

$$\begin{aligned}\dot{\xi}_1 &= \xi_2,\\ \dot{\xi}_2 &= -g \sin \xi_3 + \xi_1 \xi_4^2,\\ \dot{\xi}_3 &= \xi_4,\\ \dot{\xi}_4 &= u,\end{aligned}$$

where g denotes the gravity. Notice that if we suppose to control the angular velocity $\xi_4 = \dot{\xi}_3$, which is the case of [5] and [8], then the system is three-dimensional but the control enters nonlinearly.

At any equilibrium point $\xi_0 = (\xi_{10}, \xi_{20}, \xi_{30}, \xi_{40})^T = (\xi_{10}, 0, 0, 0)^T$, the linear part of the system is controllable. Our goal is to produce, for the variable length pendulum,

a normal form and the canonical form as well as to answer the question whether the systems corresponding to various values of the gravity constant g are feedback equivalent. In order to get a normal form, put

$$\begin{aligned} x_1 &= \xi_1, \\ x_2 &= \xi_2, \\ x_3 &= -g \sin \xi_3, \\ x_4 &= -g\xi_4 \cos \xi_3, \\ v &= g\xi_4^2 \sin \xi_3 - ug \cos \xi_3. \end{aligned}$$

The system becomes

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= x_3 + x_4^2 \frac{x_1}{g^2 - x_3^2}, \\ \dot{x}_3 &= x_4, \\ \dot{x}_4 &= v, \end{aligned}$$

which gives a normal form. Indeed, we rediscover Σ_{NF}^{∞}, given by (2.16), with $P_{1,3} = 0$, $P_{1,4} = 0$, and

$$P_{2,4} = \frac{x_1}{g^2 - x_3^2}.$$

In order to bring the system to its canonical form Σ_{CF}^{∞}, first observe that $m_0 = 3$. Indeed, the function $x_4^2 \frac{x_1}{g^2 - x_3^2}$ starts with third order terms, which corresponds to the fact that the invariants $a^{[2]j,i+2}$ vanish for any $1 \leq j \leq 2$ and any $0 \leq i \leq 2 - j$. The only nonzero component of $f^{[3]}$ is $f_2^{[3]} = x_4^2 P_{2,4}^{[1]}$. Hence $j^* = 2$ and the only, and thus largest, quadruplet (i_1, i_2, i_3, i_4) of nonnegative integers, satisfying $i_1 + i_2 + i_3 + i_4 = 3$ and such that

$$\frac{\partial^3 f_2^{[3]}}{\partial x_1^{i_1} \cdots \partial x_4^{i_4}} \neq 0,$$

is $(i_1, i_2, i_3, i_4) = (1, 0, 0, 2)$. In order to normalize $f_2^{[3]}$, put

$$\begin{aligned} \tilde{x}_i &= a_1 x_i, \quad 1 \leq i \leq 4, \\ \tilde{v} &= a_1 v, \end{aligned}$$

where $a_1 = 1/g$. We get the following canonical form for the variable length pendulum:

$$\begin{aligned} \dot{\tilde{x}}_1 &= \tilde{x}_2, \\ \dot{\tilde{x}}_2 &= \tilde{x}_3 + \tilde{x}_4^2 \frac{\tilde{x}_1}{1 - \tilde{x}_3^2}, \\ \dot{\tilde{x}}_3 &= \tilde{x}_4, \\ \dot{\tilde{x}}_4 &= \tilde{v}. \end{aligned}$$

Independently of the value of the gravity constant g, all systems are feedback equivalent to each other. □

4. Proofs of Theorems 5 and 6.

Proof of Theorem 5. The proof of this theorem will be done in two steps. In the first step we will deal with terms of degree m_0. Then we will prove the general step by an induction argument.

First step. Let us consider the system Σ^∞ given by (2.1) and let m_0 be the degree of the first nonlinearizable homogeneous part. We can assume that (see Theorems 1 and 2), after applying a suitable feedback transformation, the system Σ^∞ given by (2.1) takes the form

$$\dot{\xi} = A\xi + Bu + \bar{f}^{[m_0]}(\xi) + \sum_{m=m_0+1}^{\infty} \left(f^{[m]}(\xi) + g^{[m-1]}(\xi)u \right), \tag{4.1}$$

where (A, B) is in Brunovský canonical form and the first nonvanishing vector field $\bar{f}^{[m_0]}$ is of the form

$$\bar{f}_j^{[m_0]}(\xi) = \begin{cases} \displaystyle\sum_{i=j+2}^{n} \xi_i^2 P_{j,i}^{[m_0-2]}(\xi_1, \dots, \xi_i), & 1 \le j \le n-2, \\ 0, & n-1 \le j \le n. \end{cases}$$

Notice that the linear feedback transformation

$$\Gamma^1 \ : \ \begin{array}{l} x = a_1 \xi, \\ u = \frac{1}{a_1} v, \end{array}$$

where $a_1 \in \mathbb{R}$ and $a_1 \neq 0$, brings the system (4.1) into the following one:

$$\dot{x} = Ax + Bv + \frac{1}{a_1^{m_0-1}} \bar{f}^{[m_0]}(x) + \sum_{m=m_0+1}^{\infty} \left(\tilde{f}^{[m]}(x) + \tilde{g}^{[m-1]}(x)v \right).$$

By the definitions of $(i_1, \dots, i_{n-s})$ and j^*, we have

$$\frac{\partial^{m_0} \bar{f}_{j^*}^{[m_0]}}{\partial x_1^{i_1} \cdots \partial x_{n-s}^{i_{n-s}}} \neq 0,$$

and thus we can suitably choose the parameter a_1 such that

$$\frac{\partial^{m_0} \bar{f}_{j^*}^{[m_0]}}{\partial x_1^{i_1} \cdots \partial x_{n-s}^{i_{n-s}}} = \pm 1.$$

General step. Now, we assume that, for some $l \ge 1$, the system Σ^∞ given by (2.1), takes the form

$$\Sigma^\infty \ : \ \dot{\xi} = A\xi + Bu + \sum_{m=m_0}^{m_0+l-1} \bar{f}^{[m]}(\xi) + f^{[m_0+l]}(\xi) + g^{[m_0+l-1]}(\xi)u + r(\xi, u), \tag{4.2}$$

where $r(\xi, u) \in R^{\ge m_0+l+1}(\xi, u)$ and the vector fields $\bar{f}^{[m]}(\xi)$ for any m such that $m_0 \le m \le m_0 + l - 1$ satisfy the conditions (3.4), (3.5), and (3.6). We will construct a transformation Γ^∞ which preserves all terms of degree smaller than $m_0 + l$ while taking those of degree $m_0 + l$ into the canonical form defined by (3.4) and (3.6).

Consider the following feedback transformation

$$\Gamma^{\infty} \; : \; \begin{aligned} x &= \xi + \sum_{m=l+1}^{\infty} \phi^{[m]}(\xi), \\ u &= v + \sum_{m=l+1}^{\infty} \left(\alpha^{[m]}(\xi) + \beta^{[m-1]}(\xi) v \right), \end{aligned} \tag{4.3}$$

where, for any m such that $m_0 \leq m \leq m_0 + l - 1$, the triplet $(\phi^{[m]}, \alpha^{[m]}, \beta^{[m-1]})$ is given by (3.7) and $\phi^{[m]} = 0$, $\alpha^{[m]} = 0$, and $\beta^{[m-1]} = 0$ for $m \geq m_0 + l + 1$.

The transformation Γ^{∞} is actually a polynomial transformation $\Gamma^{\leq m_0+l}$ and can be viewed as a composition of a transformation $\Gamma^{\leq m_0+l-1}$ and a homogeneous transformation Γ^{m_0+l} defined, respectively, by

$$\Gamma^{\leq m_0+l-1} \; : \; \begin{aligned} y &= \xi + \sum_{m=l+1}^{m_0+l-1} \phi^{[m]}(\xi), \\ u &= w + \sum_{m=l+1}^{m_0+l-1} \left(\alpha^{[m]}(\xi) + \beta^{[m-1]}(\xi) w \right) \end{aligned} \tag{4.4}$$

and

$$\Gamma^{m_0+l} \; : \; \begin{aligned} x &= y + \phi^{[m_0+l]}(y), \\ w &= v + \alpha^{[m_0+l]}(y) + \beta^{[m_0+l-1]}(y) v. \end{aligned} \tag{4.5}$$

Let us denote by $\tilde{\Sigma}^{\infty}$ the system Σ^{∞}, given by (4.2), transformed via $\Gamma^{\leq m_0+l-1}$. Since

$$\bar{f}^{[m_0]}(\xi) = \bar{f}^{[m_0]}(y - \phi^{[l+1]}(y) - \cdots) = \bar{f}^{[m_0]}(y) - \frac{\partial \bar{f}^{[m_0]}}{\partial y} \phi^{[l+1]}(y) + r_1(y),$$

where $r_1(y) \in R^{\geq m_0+l+1}(y)$ and for any $m \geq m_0 + 1$,

$$\bar{f}^{[m]}(\xi) = \bar{f}^{[m]}(y - \phi^{[l+1]}(y) - \cdots) = \bar{f}^{[m]}(y) + r_2(y),$$

where $r_2(y) \in R^{\geq m_0+l+1}(y)$, we get

$$\tilde{\Sigma}^{\infty} \; : \; \dot{y} = Ay + Bw + \sum_{m=m_0}^{m_0+l-1} \bar{f}^{[m]}(y) + \tilde{f}^{[m_0+l]}(y) + \tilde{g}^{[m_0+l-1]}(y) w + r_3(y, w), \tag{4.6}$$

where $r_3(y, w) \in R^{\geq m_0+l+1}(y, w)$ and

$$\begin{aligned} \tilde{f}^{[m_0+l]} &= f^{[m_0+l]} + [\bar{f}^{[m_0]}, \phi^{[l+1]}], \\ \tilde{g}^{[m_0+l-1]} &= g^{[m_0+l-1]}. \end{aligned}$$

Let

$$\left\{ a^{[m_0+l]j,i+2} \; : \; (j, i) \in \Delta \right\}$$

and

$$\left\{\tilde{a}^{[m_0+l]j,i+2} \ : \ (j,i)\in\Delta\right\}$$

denote, respectively, the sets of (m_0+l)-invariants associated with the homogeneous parts of degree m_0+l of the systems (4.2) and (4.6). We have

$$\tilde{a}^{[m_0+l]j,i+2} = a^{[m_0+l]j,i+2} + \hat{a}^{[m_0+l]j,i+2}, \tag{4.7}$$

where

$$\begin{aligned}\hat{a}^{[m_0+l]j,i+2} = CA^{j-1}\Bigg[&\sum_{k=0}^{i}(-1)^{i+k}\left(ad_{A^iB}ad_{A\xi}^{i-k}ad_{A^kB}\left[\bar{f}^{[m_0]},\phi^{[l+1]}\right]\right)(\pi_{n-i}(\xi))\\ &+\sum_{k=0}^{i-1}(-1)^{i+k}\left(ad_{A^{i+1}B}ad_{A\xi}^{i-k-1}ad_{A^kB}\left[\bar{f}^{[m_0]},\phi^{[l+1]}\right]\right)(\pi_{n-i}(\xi))\Bigg].\end{aligned}$$

Since the identity

$$ad_{A^{n-1}B}^{k}ad_{A\xi}^{i}h = ad_{A\xi}^{i}ad_{A^{n-1}B}^{k}h$$

holds for any vector field h and any $k,i\geq 0$, we get by differentiating

$$\begin{aligned}&L_{A^{n-1}B}^{i_1+l}\hat{a}^{[m_0+l]j,i+2}\\ &= CA^{j-1}\Bigg[\sum_{k=0}^{i}(-1)^{i+k}\left(ad_{A^iB}ad_{A\xi}^{i-k}ad_{A^kB}ad_{A^{n-1}B}^{i_1+l}\left[\bar{f}^{[m_0]},\phi^{[l+1]}\right]\right)(\pi_{n-i}(\xi))\\ &\quad+\sum_{k=0}^{i-1}(-1)^{i+k}\left(ad_{A^{i+1}B}ad_{A\xi}^{i-k-1}ad_{A^kB}ad_{A^{n-1}B}^{i_1+l}\left[\bar{f}^{[m_0]},\phi^{[l+1]}\right]\right)(\pi_{n-i}(\xi))\Bigg].\end{aligned} \tag{4.8}$$

Due to the definition of the $(n-s)$-tuple $(i_1,\dots,i_{n-s})$, we obtain

$$\begin{aligned}ad_{A^{n-1}B}^{i_1+l}\left[\bar{f}^{[m_0]},\phi^{[l+1]}\right] &= c_1\left[ad_{A^{n-1}B}^{i_1}\bar{f}^{[m_0]},ad_{A^{n-1}B}^{l}\phi^{[l+1]}\right]\\ &\quad+c_2\left[ad_{A^{n-1}B}^{i_1-1}\bar{f}^{[m_0]},ad_{A^{n-1}B}^{l+1}\phi^{[l+1]}\right],\end{aligned} \tag{4.9}$$

where c_1 and c_2 are strictly positive integers. From the relations

$$\begin{aligned}ad_{A^{n-1}B}^{l}\phi^{[l+1]} &= a_{l+1}(l+1)!(\xi_1,\xi_2,\dots,\xi_n)^T,\\ ad_{A^{n-1}B}^{l+1}\phi^{[l+1]} &= a_{l+1}(l+1)!(1,0,\dots,0)^T,\end{aligned}$$

one can easily deduce that identity (4.10) can be rewritten as

$$ad_{A^{n-1}B}^{i_1+l}\left[\bar{f}^{[m_0]},\phi^{[l+1]}\right] = \gamma_l ad_{A^{n-1}B}^{i_1}\bar{f}^{[m_0]},$$

where we set $\gamma_l = -a_{l+1}(l+1)!\,(c_1(m_0-i_1+1)+c_2)$. Plugging the above identity into the formula (4.8), we obtain

$$\begin{aligned}&L_{A^{n-1}B}^{i_1+l}\hat{a}^{[m_0+l]j,i+2}\\ &= \gamma_l CA^{j-1}\Bigg[\sum_{k=0}^{i}(-1)^{i+k}\left(ad_{A^{n-1}B}^{i_1}ad_{A^iB}ad_{A\xi}^{i-k}ad_{A^kB}\bar{f}^{[m_0]}\right)(\pi_{n-i}(\xi))\\ &\quad+\sum_{k=0}^{i-1}(-1)^{i+k}\left(ad_{A^{n-1}B}^{i_1}ad_{A^{i+1}B}ad_{A\xi}^{i-k-1}ad_{A^kB}\bar{f}^{[m_0]}\right)(\pi_{n-i}(\xi))\Bigg].\end{aligned}$$

Since $\bar{f}^{[m_0]}$ is of the form (3.4), we get for any k such that $0 \leq k \leq i-1$,

$$ad_{A^kB}\bar{f}^{[m_0]}(\pi_{n-i}(\xi)) = 0$$

and for any $t \geq 0$,

$$\left(ad^t_{A\xi}ad_{A^kB}\bar{f}^{[m_0]}\right)(\pi_{n-i}(\xi)) = 0.$$

Thus, we can deduce the relation

$$\frac{\partial^{i_1+l}\hat{a}^{[m_0+l]j,i+2}}{\partial\xi_1^{i_1+l}} = \gamma_l CA^{j-1}\frac{\partial^{i_1+2}\bar{f}^{[m_0]}}{\partial\xi_1^{i_1}\partial\xi_{n-i}^2}(\pi_{n-i}(\xi)),$$

which leads, after differentiating and setting $j = j^*$ and $i = s$, to the following one:

$$\frac{\partial^{m_0+l-2}\hat{a}^{[m_0+l]j^*,s+2}}{\partial\xi_1^{i_1+l}\partial\xi_2^{i_2}\cdots\partial\xi_{n-s}^{i_{n-s}-2}} = \gamma_l\frac{\partial^{m_0}\bar{f}^{[m_0]}_{j^*}}{\partial\xi_1^{i_1}\partial\xi_2^{i_2}\cdots\partial\xi_{n-s}^{i_{n-s}}}.$$

Differentiating (4.7) and using the above identity, we get

$$\frac{\partial^{m_0+l-2}\tilde{a}^{[m_0+l]j^*,s+2}}{\partial\xi_1^{i_1+l}\partial\xi_2^{i_2}\cdots\partial\xi_{n-s}^{i_{n-s}-2}} = \frac{\partial^{m_0+l-2}a^{[m_0+l]j^*,s+2}}{\partial\xi_1^{i_1+l}\partial\xi_2^{i_2}\cdots\partial\xi_{n-s}^{i_{n-s}-2}} + \gamma_l\frac{\partial^{m_0}\bar{f}^{[m_0]}_{j^*}}{\partial\xi_1^{i_1}\partial\xi_2^{i_2}\cdots\partial\xi_{n-s}^{i_{n-s}}}.$$

We can choose suitably the parameter a_{l+1} (recall the definition of γ_l) such that we obtain

$$\frac{\partial^{m_0+l-2}\tilde{a}^{[m_0+l]j^*,s+2}}{\partial\xi_1^{i_1+l}\partial\xi_2^{i_2}\cdots\partial\xi_{n-i}^{i_{n-i}-2}} = 0.$$

Now, transforming the homogeneous part of degree m_0+l of the system (4.6) to its normal form via a homogeneous transformation Γ^{m_0+l} and taking into account Theorem 2, we bring the system (4.6) into the form

$$\Sigma^\infty \ : \ \dot{x} = Ax + Bv + \sum_{m=m_0}^{m_0+l}\bar{f}^{[m]}(x) + r(x,v),$$

where $r(x,v) \in R^{\geq m_0+l+1}(x,v)$ and the vector fields $\bar{f}^{[m]}$, for any m such that $m_0 \leq m \leq m_0+l$, satisfy the conditions (3.4), (3.5), and (3.6). This completes the proof of Theorem 5. □

In our proof of Theorem 6, we will use the following result.

LEMMA 1. *A transformation Γ^∞ leaves invariant all terms of degree smaller than m_0+l of the system* (4.2) *if and only if Γ^∞ is of the form*

$$\Gamma^\infty \ : \ \begin{aligned} x &= \xi + \sum_{m=l+1}^{\infty}\phi^{[m]}(\xi), \\ u &= v + \sum_{m=l+1}^{\infty}\left(\alpha^{[m]}(\xi) + \beta^{[m-1]}(\xi)v\right), \end{aligned} \tag{4.10}$$

where, for any m such that $m_0 \leq m \leq m_0+l-1$, the triplet $(\phi^{[m]}, \alpha^{[m]}, \beta^{[m-1]})$ is given by (3.7).

Proof of Lemma 1. We have shown, when proving Theorem 5, that the transformation Γ^∞, defined by (4.10) and (3.7), leaves invariant all terms of degree smaller than $m_0 + l$ of the system (4.2).

Conversely, assume that the transformation Γ^∞ leaves invariant all terms of degree smaller than $m_0 + l$ of the system (4.2). Without loss of generality, we can take

$$\Gamma^\infty \ : \begin{aligned} x &= \xi + \sum_{m=k+1}^{\infty} \phi^{[m]}(\xi), \\ u &= v + \sum_{m=k+1}^{\infty} (\alpha^{[m]}(\xi) + \beta^{[m-1]}(\xi)v), \end{aligned}$$

where $k+1$ denotes the smallest degree among degrees of all nonvanishing components $\phi_j^{[m]}$ of the transformation Γ^∞. There is nothing to prove if $k+1 \geq m_0 + l$. We thus focus our attention on the case $k+2 \leq m_0 + l$. Since Γ^∞ leaves invariant all terms of degree smaller than $m_0 + l$ of the system (4.2), in particular it leaves invariant terms of degree $k+1$, which implies that $(\phi^{[k+1]}, \alpha^{[k+1]}, \beta^{[k]})$ satisfies the condition (3.7). By induction, we show that $(\phi^{[m]}, \alpha^{[m]}, \beta^{[m-1]})$ also satisfies the condition (3.7) for any m such that $k+1 \leq m \leq m_0 + k - 1$. Thus it remains only to prove that $k \geq l$. Assume this is false; that is, suppose $k \leq l - 1$. We can see that the transformation Γ^∞ brings the system (4.2) into the following one:

$$\dot{x} = Ax + Bv + \sum_{m=m_0}^{m_0+k-1} \bar{f}^{[m]}(x) + \tilde{f}^{[m_0+k]}(x) + r(x, v), \tag{4.11}$$

where $r(x, v) \in R^{\geq m_0+k+1}(x, v)$ and the vector field $\bar{f}^{[m]}(x)$, for any m such that $m_0 \leq m \leq m_0 + k - 1$, is of the form (3.4) and (3.6) and

$$\tilde{f}^{[m_0+k]} = \bar{f}^{[m_0+k]} + [\bar{f}^{[m_0]}, \phi^{[k+1]}].$$

Since the transformation Γ^∞ leaves invariant all terms of degree smaller than $m_0 + l$ of the system (4.2), in particular it leaves invariant all terms of degree $m_0 + k$, which is equivalent to

$$[\bar{f}^{[m_0]}, \phi^{[k+1]}] = 0.$$

Repeating the calculations done in the proof of Theorem 5 we deduce, by differentiating, the identity

$$\frac{\partial^{m_0+k} C A^{j^*-1} [\bar{f}^{[m_0]}, \phi^{[k+1]}]}{\partial x_1^{i_1+k} \partial x_2^{i_2} \cdots \partial x_{n-s}^{i_{n-s}}} = \gamma_k \frac{\partial^{m_0} \bar{f}_{j^*}^{[m_0]}}{\partial x_1^{i_1} \partial x_2^{i_2} \cdots \partial x_{n-s}^{i_{n-s}}} = 0.$$

Thus, due to the fact that

$$\frac{\partial^{m_0} \bar{f}_{j^*}^{[m_0]}}{\partial x_1^{i_1} \partial x_2^{i_2} \cdots \partial x_{n-s}^{i_{n-s}}} \neq 0,$$

we obtain $\gamma_k = 0$ and hence $(\phi^{[k+1]}, \alpha^{[k+1]}, \beta^{[k]}) = 0$, which contradicts the definition of $k+1$. As a conclusion, it follows that the transformation Γ^∞ is of the form (4.10) and (3.7). □

Proof of Theorem 6. Let us consider two systems Σ_1^∞ and Σ_2^∞ and let

$$\Sigma_{1,CF}^\infty \ : \ \dot{x} = Ax + Bv + \sum_{m=m_0}^{\infty} \bar{f}^{[m]}(x)$$

and

$$\Sigma_{2,CF}^\infty \ : \ \dot{z} = Az + Bw + \sum_{m=m_1}^{\infty} \tilde{f}^{[m]}(z)$$

denote, respectively, their canonical forms, where m_0 and m_1 denote the degrees of the first nonlinearizable homogeneous parts. It is obvious that Σ_1^∞ and Σ_2^∞ are feedback equivalent if their canonical forms $\Sigma_{1,CF}^\infty$ and $\Sigma_{2,CF}^\infty$ coincide. To prove the converse, we assume that the systems Σ_1^∞ and Σ_2^∞ are feedback equivalent while their canonical forms fail to coincide. Since Σ_1^∞ and Σ_2^∞ are feedback equivalent, so are their canonical forms $\Sigma_{1,CF}^\infty$ and $\Sigma_{2,CF}^\infty$. It means that there exists a transformation Γ^∞ which brings $\Sigma_{1,CF}^\infty$ into $\Sigma_{2,CF}^\infty$. First remark that, from the definition of the integers m_0 and m_1, we necessarily have $m_0 = m_1$. Then, Theorem 2 and the fact that the components $\bar{f}_{j^*}^{[m_0]}$ and $\tilde{f}_{j^*}^{[m_0]}$ are normalized (see (3.5)) ensure that $\bar{f}^{[m_0]} = \tilde{f}^{[m_1]}$. Let l be the largest integer such that for any $i \leq l$, we have $\bar{f}^{[m_0+i-1]} = \tilde{f}^{[m_0+i-1]}$. This means that the transformation Γ^∞ leaves invariant all terms of degree smaller than m_0+l of the system $\Sigma_{1,CF}^\infty$. Then Lemma 1 shows that the transformation Γ^∞ is of the form (4.10). Since the transformation Γ^∞ brings $\Sigma_{1,CF}^\infty$ into $\Sigma_{2,CF}^\infty$, we deduce that

$$\tilde{f}^{[m_0+l]} = \bar{f}^{[m_0+l]} + [\bar{f}^{[m_0]}, \phi^{[l+1]}]. \tag{4.12}$$

Following arguments in the proof of Theorem 5, we obtain

$$\frac{\partial^{m_0+l-2}\tilde{a}^{[m_0+l]j^*,s+2}}{\partial x_1^{i_1+l}\partial x_2^{i_2}\cdots\partial x_{n-s}^{i_{n-s}-2}} = \frac{\partial^{m_0+l-2}\bar{a}^{[m_0+l]j^*,s+2}}{\partial x_1^{i_1+l}\partial x_2^{i_2}\cdots\partial x_{n-s}^{i_{n-s}-2}} + \gamma_l \frac{\partial^{m_0}\bar{f}_{j^*}^{[m_0]}}{\partial x_1^{i_1}\partial x_2^{i_2}\cdots\partial x_{n-s}^{i_{n-s}}},$$

where

$$\left\{\bar{a}^{[m_0+l]j,i+2} \ : \ (j,i) \in \Delta\right\}$$

and

$$\left\{\tilde{a}^{[m_0+l]j,i+2} \ : \ (j,i) \in \Delta\right\}$$

denote, respectively, the set of (m_0+l)-invariants associated with the homogeneous parts of degree m_0+l of the systems $\Sigma_{1,CF}^\infty$ and $\Sigma_{2,CF}^\infty$. Using Theorem 2, the last identity can be rewritten as

$$\frac{\partial^{m_0+l}\tilde{f}_{j^*}^{[m_0+l]}}{\partial x_1^{i_1+l}\partial x_2^{i_2}\cdots\partial x_{n-s}^{i_{n-s}}} = \frac{\partial^{m_0+l}\bar{f}_{j^*}^{[m_0+l]}}{\partial x_1^{i_1+l}\partial x_2^{i_2}\cdots\partial x_{n-s}^{i_{n-s}}} + \gamma_l \frac{\partial^{m_0}\bar{f}_{j^*}^{[m_0]}}{\partial x_1^{i_1}\partial x_2^{i_2}\cdots\partial x_{n-s}^{i_{n-s}}}. \tag{4.13}$$

Since

$$\frac{\partial^{m_0}\tilde{f}_{j^*}^{[m_0+l]}}{\partial x_1^{i_1}\cdots\partial x_{n-s}^{i_{n-s}}}(x_1,0,\ldots,0) = \frac{\partial^{m_0}\bar{f}_{j^*}^{[m_0+l]}}{\partial x_1^{i_1}\cdots\partial x_{n-s}^{i_{n-s}}}(x_1,0,\ldots,0) = 0,$$

the identity (4.13) gives

$$\gamma_l \frac{\partial^{m_0} \bar{f}_{j^*}^{[m_0]}}{\partial x_1^{i_1} \partial x_2^{i_2} \cdots \partial x_{n-s}^{i_{n-s}}} = 0,$$

which implies $\gamma_l = 0$, that is (recall the definition of γ_l), we have $a_{l+1} = 0$, and consequently $(\phi^{[l+1]}, \alpha^{[l+1]}, \beta^{[l]}) = 0$. Then the identity (4.12) reduces to

$$\tilde{f}^{[m_0+l]} = \bar{f}^{[m_0+l]},$$

which contradicts the definition of l. We conclude that the canonical forms $\Sigma^{\infty}_{1,CF}$ and $\Sigma^{\infty}_{2,CF}$ coincide. □

5. Dual normal form and dual m-invariants. In the normal form $\Sigma^{[m]}_{NF}$ given by (2.7), all the components of the control vector field $g^{[m-1]}$ are annihilated and all nonremovable nonlinearities are grouped in $f^{[m]}$. Kang and Krener in their pioneering paper [26] showed that it is possible to transform, via a transformation Γ^2 of degree 2, the homogeneous system

$$\Sigma^{[2]}: \ \dot{\xi} = A\xi + Bu + f^{[2]}(\xi) + g^{[1]}(\xi)u$$

into a dual normal form. In that form the components of the drift $f^{[2]}$ are annihilated, while this time all nonremovable nonlinearities are present in $g^{[1]}$. The aim of this section is to propose, for an arbitrary m, a dual normal form for the system $\Sigma^{[m]}$ and a dual normal form for the system Σ^{∞}. Our dual normal form on the one hand generalizes, for higher order terms, that given in [26] for second order terms, and on the other hand dualizes the normal form $\Sigma^{[m]}_{NF}$. The structure of this section will follow that of section 2: we will give the dual normal form, then define and study dual m-invariants; finally, we give an explicit construction of transformations bringing the system into its dual normal form.

Our first result asserts that we can always bring $\Sigma^{[m]}$ to a dual normal form.

THEOREM 7. *The homogeneous system $\Sigma^{[m]}$ is equivalent, via a homogeneous feedback transformation Γ^m, to the dual normal form $\Sigma^{[m]}_{DNF}$ given by*

$$\Sigma^{[m]}_{DNF}: \begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = x_3 + v x_n Q^{[m-2]}_{2,n}(x_1, \dots, x_n), \\ \qquad \vdots \\ \dot{x}_j = x_{j+1} + v \sum_{i=n-j+2}^{n} x_i Q^{[m-2]}_{j,i}(x_1, \dots, x_i), \\ \qquad \vdots \\ \dot{x}_{n-1} = x_n + v \sum_{i=3}^{n} x_i Q^{[m-2]}_{j,i}(x_1, \dots, x_i), \\ \dot{x}_n = v, \end{cases} \tag{5.1}$$

where $Q^{[m-2]}_{j,i}(x_1, \dots, x_i)$ are homogeneous polynomials of degree $m-2$ depending on the indicated variables.

Theorem 7 follows from Theorem 9, which gives explicit transformation bringing $\Sigma^{[m]}$ to its dual normal form $\Sigma^{[m]}_{DNF}$, and thus we omit its proof.

Now we will define dual m-invariants. To start with, recall that the homogeneous vector field $X_i^{[m-1]}$ is defined by taking the homogeneous part of degree $m-1$ of the vector field

$$X_i^{m-1} = (-1)^i ad^i_{A\xi+f^{[m]}}(B + g^{[m-1]}).$$

By $X_i^{[m-1]}(\pi_i(\xi))$ we will denote the vector field $X_i^{[m-1]}$ evaluated at the point $\pi_i(\xi) = (\xi_1, \dots, \xi_i, 0, \dots, 0)$ of the submanifold

$$W_i = \{\xi \in \mathbb{R}^n \ : \ \xi_{i+1} = \cdots = \xi_n = 0 \ \}.$$

Consider the system $\Sigma^{[m]}$ and, for any j such that $2 \leq j \leq n-1$, define the polynomial $b_j^{[m-1]}$ by setting

$$b_j^{[m-1]} = g_j^{[m-1]} + \sum_{k=1}^{j-1} L_B L_{A\xi}^{j-k-1} f_k^{[m]} - \sum_{i=1}^{n} L_B L_{A\xi}^{j-1} \int_0^{\xi_i} CX_{n-i}^{[m-1]}(\pi_i(\xi)) d\xi_i.$$

The homogeneous polynomials $b_j^{[m-1]}$ for $2 \leq j \leq n-1$ will be called the *dual m-invariants* of the homogeneous system $\Sigma^{[m]}$.

Consider two systems $\Sigma^{[m]}$ and $\tilde{\Sigma}^{[m]}$ of the form (2.4) and (2.5). Let

$$\{\, b_j^{[m-1]} \ : \ 2 \leq j \leq n-1 \,\}$$

and

$$\{\, \tilde{b}_j^{[m-1]} \ : \ 2 \leq j \leq n-1 \,\}$$

denote, respectively, their dual m-invariants. The following result gives a dualization of Theorem 2.

THEOREM 8. *The dual m-invariants have the following properties:*

(i) *Two systems $\Sigma^{[m]}$ and $\tilde{\Sigma}^{[m]}$ are equivalent via a homogeneous feedback transformation Γ^m if and only if*

$$b_j^{[m-1]} = \tilde{b}_j^{[m-1]}$$

for any $2 \leq j \leq n-1$.

(ii) *The dual m-invariants $\bar{b}_j^{[m-1]}$ of the dual normal form $\Sigma_{DNF}^{[m]}$, defined by* (5.1)*, are given by*

$$\bar{b}_j^{[m-1]}(x) = \sum_{i=n-j+2}^{n} x_i Q_{j,i}^{[m-2]}(x_1, \dots, x_i)$$

for any $2 \leq j \leq n-1$.

The above result asserts that the dual m-invariants, as do the m-invariants, form a set of complete invariants of the homogeneous feedback transformation. Notice, however, that the same information is encoded in both sets of invariants in different ways. We will give a proof of Theorem 8 in section 7.

Now, we define the following homogeneous polynomials:

$$
\begin{aligned}
\phi_1^{[m]} &= -\sum_{i=1}^{n} \int_0^{\xi_i} CX_{n-i}^{[m-1]}(\pi_i(\xi))d\xi_i, \\
\phi_{j+1}^{[m]} &= f_j^{[m]} + L_{A\xi}\phi_j^{[m]}, \qquad 1 \le j \le n-1, \\
\alpha^{[m]} &= -\left(f_n^{[m]} + L_{A\xi}\phi_n^{[m]}\right), \\
\beta^{[m-1]} &= -\left(g_n^{[m-1]} + L_B\phi_n^{[m]}\right).
\end{aligned} \tag{5.2}
$$

The next result gives an explicit construction of feedback transformations bringing the system $\Sigma^{[m]}$ to its dual normal form $\Sigma_{DNF}^{[m]}$.

THEOREM 9. *The feedback transformation*

$$
\Gamma^m \ : \ \begin{aligned} x &= \xi + \phi^{[m]}(\xi), \\ u &= v + \alpha^{[m]}(\xi) + \beta^{[m-1]}(\xi)v, \end{aligned}
$$

where $\alpha^{[m]}$, $\beta^{[m-1]}$, and the components $\phi_j^{[m]}$ of $\phi^{[m]}$ are defined by (5.2)*, brings the system $\Sigma^{[m]}$ into its dual normal form $\Sigma_{DNF}^{[m]}$ given by* (5.1).

6. Dual canonical form. Consider the system Σ^∞ of the form (2.1) and assume that its linear part (F, G) is controllable. Apply successively to it a series of transformations Γ^m, $m = 1, 2, \ldots$, such that each Γ^m brings $\Sigma^{[m]}$ to its dual normal form $\Sigma_{DNF}^{[m]}$; for instance we can take a series of transformations defined by (5.2). Successive repeating of Theorem 9 gives the following dual normal form.

THEOREM 10. *The system Σ^∞ can be transformed via a formal feedback transformation Γ^∞ into the dual normal form Σ_{DNF}^∞ given by*

$$
\Sigma_{DNF}^\infty \ : \ \begin{cases}
\dot{x}_1 = x_2, \\
\dot{x}_2 = x_3 + vx_nQ_{2,n}(x_1, \ldots, x_n), \\
\qquad \vdots \\
\dot{x}_j = x_{j+1} + v \displaystyle\sum_{i=n-j+2}^{n} x_iQ_{j,i}(x_1, \ldots, x_i), \\
\qquad \vdots \\
\dot{x}_{n-1} = x_n + v\displaystyle\sum_{i=3}^{n} x_iQ_{j,i}(x_1, \ldots, x_i), \\
\dot{x}_n = v,
\end{cases} \tag{6.1}
$$

where $Q_{j,i}(x_1, \ldots, x_i)$ are formal power series depending on the indicated variables.

Naturally, as with normal forms, a given system can admit different dual normal forms. We are thus interested in constructing a dual canonical form. Assuming that the linear part (F, G) of the system Σ^∞, of the form (2.1), is controllable, we denote by m_0 the degree of the first homogeneous term of the system Σ^∞ which cannot be annihilated by a feedback transformation. Thus, using Theorems 8 and 9, we can assume, after applying a suitable feedback, that Σ^∞ takes the form

$$
\Sigma^\infty \ : \ \dot{\xi} = A\xi + Bu + \bar{g}^{[m_0-1]}(\xi)u + \sum_{m=m_0+1}^{\infty} \left(f^{[m]}(\xi) + g^{[m-1]}(\xi)u\right),
$$

where (A, B) is in Brunovský canonical form and the first nonvanishing homogeneous vector field $\bar{g}^{[m_0-1]}$ is of the form

$$\bar{g}_j^{[m_0-1]}(\xi) = \begin{cases} \displaystyle\sum_{i=n-j+2}^{n} \xi_i Q_{j,i}^{[m_0-2]}(\xi_1, \dots, \xi_i), & 2 \le j \le n-1, \\ 0, & j = 1 \text{ and } j = n. \end{cases}$$

Define

$$j_* = \inf \left\{ j = 2, \dots, n-1 \ : \ \bar{g}_j^{[m_0-1]}(\xi) \neq 0 \right\}$$

and let $(i_1, \dots, i_n)$ such that $i_1 + \cdots + i_n = m_0 - 1$ be the largest, in the lexicographic ordering, n-tuple of nonnegative integers such that

$$\frac{\partial^{m_0-1} \bar{g}_{j_*}^{[m_0-1]}}{\partial \xi_1^{i_1} \cdots \partial \xi_n^{i_n}} \neq 0.$$

We get the following result.

THEOREM 11. *There exists a formal feedback transformation Γ^∞ which brings the system Σ^∞ into the following one:*

$$\Sigma_{DCF}^\infty : \dot{x} = Ax + Bv + \sum_{m=m_0}^{\infty} \bar{g}^{[m-1]}(x) v,$$

where for any $m \ge m_0$,

$$\bar{g}_j^{[m-1]} = \begin{cases} \displaystyle\sum_{i=n-j+2}^{n} x_i Q_{j,i}^{[m-2]}(x_1, \dots, x_i), & 2 \le j \le n-1, \\ 0, & j = 1 \text{ and } j = n. \end{cases} \tag{6.2}$$

Moreover,

$$\frac{\partial^{m_0-1} \bar{g}_{j_*}^{[m_0-1]}}{\partial x_1^{i_1} \cdots \partial x_n^{i_n}} = \pm 1, \tag{6.3}$$

and for any $m \ge m_0 + 1$

$$\frac{\partial^{m_0-1} \bar{g}_{j_*}^{[m-1]}}{\partial x_1^{i_1} \cdots \partial x_n^{i_n}}(x_1, 0, \dots, 0) = 0. \tag{6.4}$$

The form Σ_{DCF}^∞, which satisfies (6.2), (6.3), and (6.4), will be called the *dual canonical form* of Σ^∞. The name is justified by the following.

THEOREM 12. *The two systems Σ_1^∞ and Σ_2^∞ are formally feedback equivalent if and only if their dual canonical forms $\Sigma_{1,DCF}^\infty$ and $\Sigma_{2,DCF}^\infty$ coincide.*

Example 6. Let us consider the system

$$\Sigma \ : \ \dot{\xi} = f(\xi) + g(\xi) u, \quad \xi(\cdot) \in \mathbb{R}^3, \ u(\cdot) \in \mathbb{R},$$

whose linear part is assumed to be controllable. Theorem 10 ensures that the system Σ is formally feedback equivalent to the dual normal form Σ^{∞}_{DNF} given by

$$\begin{aligned}\dot{x}_1 &= x_2,\\ \dot{x}_2 &= x_3 + v x_3 Q(x_1, x_2, x_3),\\ \dot{x}_3 &= v,\end{aligned}$$

where $Q(x_1, x_2, x_3)$ is a formal power series of variables x_1, x_2, x_3.

Assume for simplicity that $m_0 = 2$, which is equivalent to the condition that g, $ad_f g$, and $[g, ad_f g]$ are linearly independent at $0 \in \mathbb{R}^3$. This implies that we can represent $Q = Q(x_1, x_2, x_3)$ by

$$Q = c + x_1 Q_1(x_1) + x_2 Q_2(x_1, x_2) + x_3 Q_3(x_1, x_2, x_3),$$

where $c \in \mathbb{R}$, $c \neq 0$.

Observe that any Q of the above form gives a dual normal form Σ^{∞}_{DNF}. In order to get the dual canonical form we use Theorem 11, which ensures that the system Σ is formally feedback equivalent to its dual canonical form Σ^{∞}_{DCF} defined by

$$\begin{aligned}\dot{\tilde{x}}_1 &= \tilde{x}_2,\\ \dot{\tilde{x}}_2 &= \tilde{x}_3 + \tilde{v}\tilde{x}_3 \tilde{Q}(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3),\\ \dot{\tilde{x}}_3 &= \tilde{v},\end{aligned}$$

where $\tilde{Q}(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3)$ is a formal power series such that

$$\tilde{Q}(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) = 1 + \tilde{x}_2 \tilde{Q}_2(\tilde{x}_1, \tilde{x}_2) + \tilde{x}_3 \tilde{Q}_3(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3). \qquad \square$$

7. Proofs of dual results. In this section, we prove our dual results. The proof of Theorem 7 will be omitted because in the proof of Theorem 9 we give an explicit homogeneous feedback transformation bringing a given homogeneous system into its dual normal form. Theorem 10 follows from a successive application of Theorem 7. We will thus prove Theorems 8, 9, 11, and 12.

Proof of Theorem 8. (i) We will prove that if the system $\Sigma^{[m]}$ is equivalent to $\tilde{\Sigma}^{[m]}$ via a transformation Γ^m, then their dual m-invariants $b_j^{[m-1]}$ and $\tilde{b}_j^{[m-1]}$ coincide. The action of Γ^m can be decomposed into that of a pure feedback of the form

$$u = v + \alpha^{[m]}(\xi) + \beta^{[m-1]}(\xi) v$$

followed by that of a diffeomorphism

$$x = \xi + \phi^{[m]}(\xi)$$

of the state space. Since the first $n-1$ components of the vector fields $f^{[m]}$ and $g^{[m-1]}$, as well as those of $X_{n-i}^{[m-1]}$, are invariant under pure feedback, we can conclude that the functions $b_j^{[m-1]}$ for $2 \leq j \leq n-1$ are invariant under pure feedback. It remains to prove that they are also invariant under any diffeomorphism $x = \Phi(\xi)$ of the form $\Phi(\xi) = \xi + \phi^{[m]}(\xi)$.

The diffeomorphism Φ brings the system $\Sigma^{[m]}$ into the form

$$\tilde{\Sigma}^{[m]} \ : \ \dot{x} = Ax + Bu + \tilde{f}^{[m]}(x) + \tilde{g}^{[m-1]}(x) u,$$

where

$$\tilde{f}^{[m]} = f^{[m]} + [Ax, \phi^{[m]}],$$
$$\tilde{g}^{[m-1]} = g^{[m-1]} + L_B\phi^{[m]}.$$

Denoting by $b_j^{[m-1]}$ and $\tilde{b}_j^{[m-1]}$ for $2 \leq j \leq n-1$ the dual m-invariants associated, respectively, with the homogeneous systems $\Sigma^{[m]}$ and $\tilde{\Sigma}^{[m]}$, we get

$$\tilde{b}_j^{[m-1]} = b_j^{[m-1]} + \hat{b}_j^{[m-1]},$$

where

$$\hat{b}_j^{[m-1]}(x) = L_B\phi_j^{[m]}(x) + \sum_{k=0}^{j-2} L_B L_{Ax}^{j-k-2} C A^k ad_{Ax}\phi^{[m]} - \sum_{i=1}^{n} L_B L_{Ax}^{j-1} \int_0^{x_i} C\hat{X}_{n-i}^{[m-1]}(\pi_i(x))dx_i$$

and

$$\hat{X}_{n-i}^{[m-1]}(x) = (-1)^{n-i} ad_{Ax+[Ax,\phi^{[m]}]}^{n-i}(B + L_B\phi^{[m]}) = (-1)^{n-i}(\Phi_* ad_{A\xi}^{n-i}(B))(x)$$
$$= A^{n-i}B + L_{A^{n-i}B}\phi^{[m]}(x).$$

We can deduce that

$$\hat{b}_j^{[m-1]}(x) = L_B\phi_j^{[m]}(x) + \sum_{k=1}^{j-1} L_B L_{Ax}^{j-k}\phi_k^{[m]} - \sum_{k=2}^{j} L_B L_{Ax}^{j-k}\phi_k^{[m]} - \sum_{i=1}^{n} L_B L_{Ax}^{j-1} \int_0^{x_i} L_{A^{n-i}B}\phi_1^{[m]}(\pi_i(x))dx_i$$
$$= L_B\phi_j^{[m]}(x) + \sum_{k=1}^{j-1} L_B L_{Ax}^{j-k}\phi_k^{[m]} - \sum_{k=2}^{j} L_B L_{Ax}^{j-k}\phi_k^{[m]} - L_B L_{Ax}^{j-1}\phi_1^{[m]}(x) = 0,$$

which gives

$$\tilde{b}_j^{[m-1]} = b_j^{[m-1]}.$$

Thus the functions $b_j^{[m-1]}$ are invariant under any diffeomorphism of the form $x = \Phi(\xi) = \xi + \phi^{[m]}(\xi)$. Therefore they remain invariant under the transformation Γ^m.

The fact that two homogeneous systems, whose dual m-invariants coincide, are feedback equivalent follows clearly from item (ii) of the theorem, which will be proved below. Indeed, by item (ii), both systems coincide when transformed to their canonical forms.

(ii) Denote by $\bar{b}_j^{[m-1]}$ for $2 \leq j \leq n-1$ the dual m-invariants associated with the dual normal form $\Sigma_{DNF}^{[m]}$. They are given by

$$\bar{b}_j^{[m-1]} = \bar{g}_j^{[m-1]} - \sum_{i=1}^{n} L_B L_{Ax}^{j-1} \int_0^{x_i} C\bar{X}_{n-i}^{[m-1]}(\pi_i(x))dx_i,$$

where the components $\bar{g}_j^{[m-1]}$ are given by (5.1) and

$$C\bar{X}_{n-i}^{[m-1]} = (-1)^{n-i} C ad_{Ax}^{n-i} \bar{g}^{[m-1]}.$$

It suffices to observe (see Lemma 2 below) that, on the one hand, $C\bar{X}_{n-i}^{[m-1]}$ is a linear combination of functions $L_{Ax}^s \bar{g}_j^{[m-1]}$ for $0 \le s \le n-i$ and $1 \le j \le n-i+1$ and, on the other hand, $\bar{g}_j^{[m-1]}(\pi_i(x)) = 0$ for all j such that $1 \le j \le n-i+1$. We thus conclude that $C\bar{X}_{n-i}^{[m-1]}(\pi_i(x)) = 0$, which implies

$$\bar{b}_j^{[m-1]} = \bar{g}_j^{[m-1]}$$

for any j such that $2 \le j \le n-1$. □

Proof of Theorem 9. Denote by

$$\tilde{\Sigma}^{[m]} \;:\; \dot{x} = Ax + Bv + \tilde{f}^{[m]}(x) + \tilde{g}^{[m-1]}(x)v$$

the system $\Sigma^{[m]}$ transformed via a homogeneous feedback transformation Γ^m defined by (5.2). From Proposition 1, it follows that for $\tilde{\Sigma}^{[m]}$ we have

$$\begin{aligned} &\tilde{f}_j^{[m]} = 0 \qquad \text{for } \; 1 \le j \le n, \\ &\tilde{g}_j^{[m-1]} = 0 \qquad \text{for } \; j = 1 \text{ and } j = n, \\ &\tilde{g}_j^{[m-1]} = g_j^{[m-1]} + L_B \phi_j^{[m]} \qquad \text{for } \; 2 \le j \le n-1. \end{aligned} \tag{7.1}$$

It thus suffices to show that the components $\tilde{g}_j^{[m-1]}$ for $2 \le j \le n-1$ are in the dual normal form (5.1). We prove easily by an induction argument that

$$\phi_{j+1}^{[m]} = \sum_{k=1}^{j} L_{A\xi}^{j-k} f_k^{[m]} + L_{A\xi}^j \phi_1^{[m]},$$

$$L_B L_{A\xi}^j \phi_1^{[m]} = \sum_{k=0}^{j} \binom{j}{k} L_{A\xi}^{j-k} L_{A^k B} \phi_1^{[m]},$$

which allows us to show that

$$\tilde{g}_{j+1}^{[m-1]} = g_{j+1}^{[m-1]} + \sum_{k=1}^{j} L_B L_{A\xi}^{k-1} f_{j-k+1}^{[m]} + \sum_{k=0}^{j} \binom{j}{k} L_{A\xi}^{j-k} L_{A^k B} \phi_1^{[m]}.$$

Now, from the identity

$$L_{A^k B} \phi_1^{[m]} = -CX_k^{[m-1]}(\pi_{n-k}(\xi)) - \sum_{i=n-k+1}^{n} \int_0^{\xi_i} \frac{\partial CX_{n-i}^{[m-1]}(\pi_i(\xi))}{\partial \xi_{n-k}} d\xi_i,$$

we can deduce that

$$\begin{aligned} \tilde{g}_{j+1}^{[m-1]} = g_{j+1}^{[m-1]} &+ \sum_{k=1}^{j} L_B L_{A\xi}^{k-1} f_{j-k+1}^{[m]} - \sum_{k=0}^{j} \binom{j}{k} L_{A\xi}^{j-k} CX_k^{[m]}(\pi_{n-k}(\xi)) \\ &- \sum_{k=1}^{j} \sum_{i=n-k+1}^{n} \binom{j}{k} L_{A\xi}^{j-k} \left(\int_0^{\xi_i} \frac{\partial CX_{n-i}^{[m-1]}(\pi_i(\xi))}{\partial \xi_{n-k}} d\xi_i \right). \end{aligned}$$

Taking into account that for any k such that $0 \leq k \leq j$ we have

$$L_{A\xi}^{j-k} CX_k^{[m]}(\pi_{n-k} \circ \pi_{n-j}(\xi)) = L_{A\xi}^{j-k} CX_k^{[m]}(\pi_{n-j}(\xi))$$

and that for any $i \geq n-j+1$ we have

$$\left(\int_0^{\xi_i} \frac{\partial CX_{n-i}^{[m-1]}(\pi_i(\xi))}{\partial \xi_{n-k}} d\xi_i \right) (\pi_{n-j}(\xi)) = 0,$$

we can conclude that

$$\tilde{g}_{j+1}^{[m-1]}(\pi_{n-j}(\xi)) = \left(g_{j+1}^{[m-1]} + \sum_{k=1}^{j} L_B L_{A\xi}^{j-k} f_k^{[m]} - \sum_{k=0}^{j} \binom{j}{k} L_{A\xi}^{j-k} CX_k^{[m]} \right) (\pi_{n-j}(\xi)).$$

Using Lemma 2 given below, we thus obtain

$$\tilde{g}_{j+1}^{[m-1]}(\pi_{n-j}(\xi)) = 0,$$

which proves that $\tilde{g}_j^{[m-1]}$ is in the dual normal form (5.1). □

LEMMA 2. *Let $X_i^{[m-1]}$ be the homogeneous part of degree $m-1$ of*

$$X_i^{m-1} = (-1)^i ad_{A\xi + f^{[m]}}^i (B + g^{[m-1]}).$$

Then the following identities hold:

(i) *For any $j \geq 1$, we have*

$$CA^j X_1^{[m-1]} = \sum_{k=1}^{j} L_{AB} L_{A\xi}^{j-k} f_k^{[m]} - \sum_{k=0}^{j} \binom{j}{k} L_{A\xi}^{j-k} CX_{k+1}^{[m-1]}.$$

(ii) *For any j such that $0 \leq j \leq n-1$, we have*

$$\sum_{k=0}^{j} \binom{j}{k} L_{A\xi}^{j-k} CX_k^{[m-1]} = g_{j+1}^{[m-1]} + \sum_{k=1}^{j} L_B L_{A\xi}^{j-k} f_k^{[m]}.$$

Both identities can be proved by a direct calculation.

Proof of Theorem 11. In the first step we will normalize terms of degree at most m_0 while in the general step we will normalize terms of order $m_0 + l$.

First step. Consider the system Σ^∞ and recall that m_0 is the degree of the first nonlinearizable homogeneous part. We can assume (see Theorems 7 and 8) that after applying a suitable feedback transformation, the system Σ^∞ takes the form

$$\dot{\xi} = A\xi + Bu + \bar{g}^{[m_0-1]}(\xi)u + \sum_{m=m_0+1}^{\infty} \left(f^{[m]}(\xi) + g^{[m-1]}(\xi)u \right), \tag{7.2}$$

where the vector field $\bar{g}^{[m_0-1]}$ defined by

$$\bar{g}_j^{[m_0-1]}(\xi) = \begin{cases} \displaystyle\sum_{i=n-j+2}^{n} \xi_i Q_{j,i}^{[m_0-2]}(\xi_1, \ldots, \xi_i), & 2 \leq j \leq n-1, \\ 0, & j = 1 \text{ or } j = n, \end{cases}$$

is the first nonlinearizable homogeneous part. We can notice that the linear feedback transformation

$$\Gamma^1 \ : \quad \begin{aligned} x &= a_1\xi, \\ u &= \tfrac{1}{a_1}v, \end{aligned}$$

where $a_1 \in \mathbb{R}$ and $a_1 \neq 0$, brings the system (7.2) into the following one:

$$\dot{x} = Ax + Bv + \frac{1}{a_1^{m_0-1}}\bar{g}^{[m_0-1]}(x)v + \sum_{m=m_0+1}^{\infty}\left(\tilde{f}^{[m]}(x) + \tilde{g}^{[m-1]}(x)v\right).$$

Due to the definitions of $(i_1, \dots, i_n)$ and j_*, we can suitably choose the parameter a_1 such that

$$\frac{\partial^{m_0-1}\bar{g}_{j_*}^{[m_0-1]}}{\partial x_1^{i_1}\cdots\partial x_n^{i_n}} = \pm 1.$$

General step. Now we assume that, for some $l \geq 1$, the system Σ^∞ takes the form

$$\dot{\xi} = A\xi + Bu + \sum_{m=m_0}^{m_0+l-1}\bar{g}^{[m-1]}(\xi)u + f^{[m_0+l]}(\xi) + g^{[m_0+l-1]}(\xi)u + r(\xi,u), \tag{7.3}$$

where $r(\xi,u) \in R^{\geq m_0+l+1}(\xi,u)$ and, for any m such that $m_0 \leq m \leq m_0+l-1$ and any $1 \leq j \leq n$, the components $\bar{g}_j^{[m-1]}$ satisfy the conditions (6.2), (6.3), and (6.4). Consider the transformation Γ^∞ given by (4.3), satisfying (3.7), and its decomposition $\Gamma^\infty = \Gamma^{\leq m_0+l} = \Gamma^{m_0+l} \circ \Gamma^{\leq m_0+l-1}$ given by (4.4)–(4.5). We can easily see that the transformation $\Gamma^{\leq m_0+l-1}$ brings the system (7.3) into the system

$$\dot{y} = Ay + Bw + \sum_{m=m_0}^{m_0+l-1}\bar{g}^{[m-1]}(y)w + \tilde{f}^{[m_0+l]}(y) + \tilde{g}^{[m_0+l-1]}(y)w + r(y,w), \tag{7.4}$$

where $r(y,w) \in R^{\geq m_0+l+1}(y,w)$ and

$$\begin{aligned} \tilde{f}^{[m_0+l]} &= f^{[m_0+l]} + \bar{g}^{[m_0-1]}\alpha^{[l+1]}, \\ \tilde{g}^{[m_0+l-1]} &= g^{[m_0+l-1]} + \left[\bar{g}^{[m_0-1]}, \phi^{[l+1]}\right] + \bar{g}^{[m_0-1]}\beta^{[l]}. \end{aligned}$$

Let $b_j^{[m_0+l-1]}$ and $\tilde{b}_j^{[m_0+l-1]}$ be the dual (m_0+l)-invariants associated, respectively, to the homogeneous parts of degree m_0+l of the systems (7.3) and (7.4). We thus deduce that

$$\tilde{b}_j^{[m_0+l-1]} = b_j^{[m_0+l-1]} + \hat{b}_j^{[m_0+l-1]}, \tag{7.5}$$

where

$$\hat{b}_j^{[m_0+l-1]} = CA^{j-1}\left[\bar{g}^{[m_0-1]}, \phi^{[l+1]}\right] + \beta^{[l]}\bar{g}_j^{[m_0-1]} - \sum_{i=1}^{n} L_B L_{Ay}^{j-1}\int_0^{y_i} C\hat{X}_{n-i}^{[m-1]}(\pi_i(y))dy_i$$

and

$$\begin{aligned} \hat{X}_{n-i}^{[m-1]} = (-1)^{n-i}ad_{Ay}^{n-i}&\left(\left[\bar{g}^{[m_0-1]}, \phi^{[l+1]}\right] + \bar{g}^{[m_0-1]}\beta^{[l]}\right) \\ &+ \sum_{k=0}^{n-i-1}(-1)^k ad_{Ay}^k ad_{A^{n-i-k-1}B}\left(\bar{g}^{[m_0-1]}\alpha^{[l+1]}\right). \end{aligned}$$

First notice (see Lemma 2) that $C\hat{X}_{n-i}^{[m-1]}$ is a linear combination, over the ring of polynomials, of the components $CA^{j-1}\bar{g}^{[m_0-1]}$ and $CA^{j-1}\left[\bar{g}^{[m_0-1]},\phi^{[l+1]}\right]$, $1\leq j\leq n-i+1$, and their derivatives. Since

$$CA^{j-1}\left[\bar{g}^{[m_0-1]},\phi^{[l+1]}\right]=\sum_{k=1}^{j}\frac{\partial\phi_j^{[l+1]}}{\partial y_k}\bar{g}_k^{[m_0-1]}-\frac{\partial\bar{g}_j^{[m_0-1]}}{\partial y}\phi^{[l+1]},$$

it follows that $C\hat{X}_{n-i}^{[m-1]}$ is a linear combination, over the ring of polynomials, of the components $\bar{g}_j^{[m_0-1]}$, $1\leq j\leq n-i+1$, and their derivatives. Taking into account the fact that $\bar{g}^{[m_0-1]}$ satisfies (6.2), we obtain, for any $1\leq j\leq n-i+1$,

$$\bar{g}_j^{[m_0-1]}(\pi_i(y))=0.$$

Thus, we deduce that

$$\hat{X}_{n-i}^{[m-1]}(\pi_i(y))=0,$$

which leads to the identity

$$\hat{b}_j^{[m_0+l-1]}=CA^{j-1}\left[\bar{g}^{[m_0-1]},\phi^{[l+1]}\right]+\beta^{[l]}\bar{g}_j^{[m_0-1]}.$$

Putting $j=j_*$ and due to the fact that $\bar{g}_1^{[m_0-1]}=\cdots=\bar{g}_{j_*-1}^{[m_0-1]}=0$, we get

$$\hat{b}_{j_*}^{[m_0+l-1]}=-\frac{\partial\bar{g}_{j_*}^{[m_0-1]}}{\partial y}\phi^{[l+1]}+L_{A^{n-j_*}B}(\phi_{j_*}^{[l+1]})\bar{g}_{j_*}^{[m_0-1]}+\beta^{[l]}\bar{g}_{j_*}^{[m_0-1]}.$$

Since the triplet $(\phi^{[l+1]},\alpha^{[l+1]},\beta^{[l]})$ satisfies the condition (3.7), it is easy to see that for any $1\leq j\leq n-1$, we have

$$L_{A^{n-j}B}\phi_j^{[l+1]}+\beta^{[l]}=0$$

and then conclude that

$$\hat{b}_{j_*}^{[m_0+l-1]}=-\frac{\partial\bar{g}_{j_*}^{[m_0-1]}}{\partial y}\phi^{[l+1]}.$$

Now, let us differentiate this last expression, taking into account that $(i_1,\dots,i_n)$ is the largest n-tuple of nonnegative integers such that

$$\frac{\partial^{m_0-1}\bar{g}_{j_*}^{[m_0-1]}}{\partial y_1^{i_1}\cdots\partial y_n^{i_n}}\neq 0.$$

We obtain

$$\frac{\partial^{i_1+l}\hat{b}_{j_*}^{[m_0+l-1]}}{\partial y_1^{i_1+l}}=-\left(d_1\frac{\partial^{i_1+1}\bar{g}_{j_*}^{[m_0-1]}}{\partial y\partial y_1^{i_1}}\frac{\partial^l\phi^{[l+1]}}{\partial y_1^l}+d_2\frac{\partial^{i_1}\bar{g}_{j_*}^{[m_0-1]}}{\partial y\partial y_1^{i_1-1}}\frac{\partial^{l+1}\phi^{[l+1]}}{\partial y_1^{l+1}}\right),\tag{7.6}$$

where d_1 and d_2 are strictly positive integers. Since

$$\frac{\partial^l\phi^{[l+1]}}{\partial y_1^l}=a_{l+1}(l+1)!(y_1,y_2,\dots,y_n)^T$$

and

$$\frac{\partial^{l+1}\phi^{[l+1]}}{\partial y_1^{l+1}} = a_{l+1}(l+1)!(1,0,\dots,0)^T,$$

and due to the fact that

$$\sum_{k=2}^{n}\frac{\partial \bar{g}_{j_*}^{[m_0-1]}}{\partial y_k}y_k = (m_0 - i_1 - 1)\bar{g}_{j_*}^{[m_0-1]},$$

the identity (7.6) gives

$$\frac{\partial^{i_1+l}\bar{b}_{j_*}^{[m_0+l-1]}}{\partial y_1^{i_1+l}} = \theta_l \frac{\partial^{i_1}\bar{g}_{j_*}^{[m_0-1]}}{\partial y_1^{i_1}},$$

where $\theta_l = -a_{l+1}(l+1)!\,(d_1(m_0 - i_1 - 1) + d_2)$. Plugging this last expression into (7.5), where we put $j = j_*$, we obtain, after differentiating, the following relation:

$$\frac{\partial^{m_0+l-1}\tilde{b}_{j_*}^{[m_0+l-1]}}{\partial y_1^{i_1+l}\partial y_2^{i_2}\cdots\partial y_n^{i_n}} = \frac{\partial^{m_0+l-1}\bar{b}_{j_*}^{[m_0+l-1]}}{\partial y_1^{i_1+l}\partial y_2^{i_2}\cdots\partial y_n^{i_n}} + \theta_l\frac{\partial^{m_0-1}\bar{g}_{j_*}^{[m_0-1]}}{\partial y_1^{i_1}\partial y_2^{i_2}\cdots\partial y_n^{i_n}}.$$

Because of the definition of θ_l, we can choose suitably the parameter a_{l+1} such that

$$\frac{\partial^{m_0+l-1}\tilde{b}_{j_*}^{[m_0+l-1]}}{\partial y_1^{i_1+l}\partial y_2^{i_2}\cdots\partial y_n^{i_n}} = 0,$$

which is equivalent to

$$\frac{\partial^{m_0-1}\tilde{b}_{j_*}^{[m_0+l-1]}}{\partial y_1^{i_1}\partial y_2^{i_2}\dots\partial y_n^{i_n}}(y_1,0,\dots,0) = 0.$$

Now, transforming the homogeneous part of degree $m_0 + l$ of the system (7.4) to its normal form via a homogeneous transformation Γ^{m_0+l} and taking into account Theorem 8, we bring the system (7.4) into the form

$$\dot{x} = Ax + Bv + \sum_{m=m_0}^{m_0+l}\bar{g}^{[m-1]}(x)v + r(x,v), \tag{7.7}$$

where $r(x,v) \in R^{\geq m_0+l+1}(x,v)$, and for any m such that $m_0 \leq m \leq m_0 + l$, the components $\bar{g}_j^{[m-1]}$ for $2 \leq j \leq n-1$ satisfy the conditions (6.2), (6.3), and (6.4). This ends the proof of Theorem 11. □

Proof of Theorem 12. The proof of this theorem follows the same line as that of Theorem 6. We notice only that the transformation Γ^{∞} leaves invariant all terms of degree smaller than $m_0 + l$ of the system (7.3) if and only if it is of the form (4.10), given by Lemma 1. □

REFERENCES

[1] V.I. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations*, 2nd ed., Springer-Verlag, New York, 1988.

[2] V. Bogaevski and A. Povzner, *Algebraic Methods in Nonlinear Physics*, Springer-Verlag, New York, 1991.
[3] B. Bonnard, *Feedback equivalence for nonlinear systems and the time optimal control problem,* SIAM J. Control Optim., 29 (1991), pp. 1300–1321.
[4] B. Bonnard, *Quadratic control systems,* Math. Control Signals Systems, 4 (1991), pp. 139–160.
[5] A. Bressan and F. Rampazzo, *On differential systems with quadratic impulses and their applications to Lagrangian mechanics,* SIAM J. Control Optim., 31 (1993), pp. 1205–1230.
[6] E. Cartan, *Théorie des groupes finis et continus et la géométrie différentielle traitées par la méthode du repère mobile*, Gauthier-Villars, Paris, 1937.
[7] S. Čelikovsky and H. Nijmeijer, *Equivalence of nonlinear systems to triangular form: The singular case*, Systems Control Lett., 27 (1996), pp. 135–144.
[8] M. Fliess, J. Lévine, P. Martin, and P. Rouchon, *Flatness and defect of nonlinear systems: Introductory theory and examples,* Internat. J. Control, 6 (1995), pp. 1327–1361.
[9] R.B. Gardner, *The Method of Equivalence and Its Applications*, CBMS Reg. Conf. Ser. in Appl. Math. 58, SIAM, Philadelphia, 1989.
[10] R.B. Gardner, W.F. Shadwick, and G.R. Wilkens, *A geometric isomorphism with applications to closed loop controls,* SIAM J. Control Optim., 27 (1989), pp. 1361–1368.
[11] R. Gardner and W. Shadwick, *The GS algorithm for exact linearization to Brunovský normal form,* IEEE Trans. Automat. Control, 37 (1992), pp. 224–230.
[12] R.B. Gardner, W.F. Shadwick, and G.R. Wilkens, *Feedback equivalence and symmetries of Brunovský normal forms,* Contemp. Math., 97 (1989), pp. 115–130.
[13] L.R. Hunt and R. Su, *Linear equivalents of nonlinear time varying systems,* in Proceedings of the 4th International Symposium on Mathematical Theory of Networks and Systems, Santa Monica, CA, 1981, pp. 119–123.
[14] L.R. Hunt, R. Su, and G. Meyer, *Design for multi-input nonlinear systems*, in Differential Geometric Control Theory, R.W. Brockett, R. Millman, and H.J. Sussmann, eds., Birkhäuser, Boston, 1983, pp. 268–298.
[15] B. Jakubczyk, *Equivalence and invariants of nonlinear control systems*, in Nonlinear Controllability and Optimal Control, H.J. Sussmann, ed., Marcel Dekker, New York, Basel, 1990, pp. 177–218.
[16] B. Jakubczyk, *Feedback invariants, critical trajectories, and Hamiltonian formalism*, in Nonlinear Control in the Year 2000, Vol. 1, Lecture Notes in Control and Inform. Sci. 258, A. Isidori, F. Lamnabhi-Lagarrigue, and W. Respondek, eds., Springer-Verlag, London, 2001, pp. 545–568.
[17] B. Jakubczyk, *Critical Hamiltonians and feedback invariants*, in Geometry of Feedback and Optimal Control, B. Jakubczyk and W. Respondek, eds., Marcel Dekker, New York, Basel, 1998, pp. 219–256.
[18] B. Jakubczyk and W. Respondek, *On linearization of control systems,* Bull. Acad. Polon. Sci. Sér. Sci. Math., 28 (1980), pp. 517–522.
[19] B. Jakubczyk and W. Respondek, *Feedback classification of analytic control systems in the plane*, in Analysis of Controlled Dynamical Systems, B. Bonnard et al., eds., Birkhäuser, Boston, 1991, pp. 262–273.
[20] T. Kailath, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
[21] W. Kang, *Extended controller form and invariants of nonlinear control systems with single input,* J. Math. System Estimation Control, 6 (1996), pp. 27–51.
[22] W. Kang, *Quadratic normal forms of nonlinear control systems with uncontrollable linearization,* in Proceedings of the 34th IEEE Conference on Decision and Control, New Orleans, 1995, pp. 608–612.
[23] W. Kang, *Bifurcation and normal form of nonlinear control systems, Part* I, SIAM J. Control Optim., 36 (1998), pp. 193–212.
[24] W. Kang, *Bifurcation and normal form of nonlinear control systems, Part* II, SIAM J. Control Optim., 36 (1998), pp. 213–232.
[25] W. Kang, *Normal form, invariants, and bifurcations of nonlinear control systems in the particle deflection plane*, in Dynamics, Bifurcations and Control, Lecture Notes in Control and Inform. Sci. 273, F. Colonius and L. Grüne, eds., Springer-Verlag, Berlin, Heidelberg, 2002, pp. 67–87.
[26] W. Kang and A.J. Krener, *Extended quadratic controller normal form and dynamic state feedback linearization of nonlinear systems,* SIAM J. Control Optim., 30 (1992), pp. 1319–1337.
[27] A.J. Krener, *Approximate linearization by state feedback and coordinate change,* Systems Control Lett., 5 (1984), pp. 181–185.
[28] I. Kupka, *On feedback equivalence*, in Differential Geometry, Global Analysis, and Topology,

CMS Conf. Proc. 12, AMS, Providence, RI, 1991, pp. 105–117.

[29] W. Respondek, *Feedback classification of nonlinear control systems in* $\mathbb{R}^2$ *and* $\mathbb{R}^3$*,* in Geometry of Feedback and Optimal Control, B. Jakubczyk and W. Respondek, eds., Marcel Dekker, New York, 1998, pp. 347–382.

[30] W. Respondek and I.A. Tall, *How many symmetries does admit a nonlinear single-input control system around an equilibrium?,* in Proceedings of the 40th IEEE Conference on Decision and Control, Orlando, FL, 2001, pp. 1795–1800.

[31] W. Respondek and I.A. Tall, *Nonlinearizable single-input control systems do not admit stationary symmetries,* Systems Control Lett., 46 (2002), pp. 1–16.

[32] W. Respondek and M. Zhitomirskii, *Feedback classification of nonlinear control systems on* 3*-manifolds,* Math. Control Signals Systems, 8 (1995), pp. 299–333.

[33] I.A. Tall, *Classification des systèmes de contrôles non linéaires à une entrée*, Thèse de 3^{eme} cycle, Université de Dakar, Senegal, 1999.

[34] I.A. Tall and W. Respondek, *Normal forms, canonical forms, and invariants of single single-input control systems under feedback,* in Proceedings of the 39th IEEE Conference on Decision and Control, Sydney, 2000, pp. 1625–1630.

[35] I.A. Tall and W. Respondek, *Transforming a single-input nonlinear system to a feedforward form via feedback*, in Nonlinear Control in the Year 2000, Vol. 2, Lecture Notes in Control and Inform. Sci. 259, A. Isidori, F. Lamnabhi-Lagarrigue, and W. Respondek, eds., Springer-Verlag, London, 2000, pp. 527–542.

[36] I.A. Tall and W. Respondek, *Feedback equivalence to feedforward forms for nonlinear single-input control systems,* in Dynamics, Bifurcations and Control, Lecture Notes in Control and Inform. Sci. 273, F. Colonius and L. Grüne, eds., Springer-Verlag, Berlin, Heidelberg, 2002, pp. 269–286.

[37] I.A. Tall and W. Respondek, *Feedback equivalence to feedforward forms for nonlinear single-input systems,* in Dynamics, Bifurcations, and Control, Lecture Notes in Control and Inform. Sci. 273, F. Colonius and L. Grüne, eds., Springer-Verlag, Berlin, Heidelberg, 2002, pp. 269–286.

[38] M. Zhitomirskii and W. Respondek, *Simple germs of corank one affine distributions,* in Singularities Symposium—Łojasiewicz 70, Banach Center Pub. 44, B. Jakubczyk, W. Pawłucki, and J. Stasica, eds., Polish Acad. Sci., Warsaw, 1998, pp. 269–276.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 5, pp. 1532–1553

ON THE GLOBAL CONTROLLABILITY OF NONLINEAR SYSTEMS: FOUNTAINS, RECURRENCE, AND APPLICATIONS TO HAMILTONIAN SYSTEMS*

PETER E. CAINES† AND EKATERINA S. LEMCH‡

Abstract. In this paper, a form of open local accessibility for nonlinear control systems is introduced called the continuous fountain condition. Subject to the conditions that (i) the states of a system are continuous fountains and (ii) one of various recurrence conditions holds, it is established that the system state space is (globally) controllable. It is shown that these controllability results imply the controllability of certain subsets of the state space of Hamiltonian control systems called energy slices. The relations between the notion of a fountain and (i) local accessibility and (ii) the full Lie algebra rank condition for control affine systems are investigated. Finally, the results in this paper are shown to have application to hierarchical hybrid control theory in that they give conditions for a finite analytic partition to satisfy the so-called in-block controllability property; this is illustrated with a linear mass-spring system example.

Key words. nonlinear control systems, controllability, attainable sets, recurrence, hierarchical systems, fountains, accessibility, reachability

AMS subject classifications. 93C10, 93B05, 93B03, 11B37, 93A13

PII. S0363012999361950

1. Introduction.

There is an extensive literature on various forms of the controllability problem for nonlinear systems (see, for example, the texts by Isidori [10], Nijmeijer and van der Schaft [24], and Jurdjevic [11]). In particular, in [11], [12], [13], and [19], it has been shown that control affine systems satisfying a dense recurrence condition and a full Lie algebra rank condition (LARC) are (globally) controllable. In [20], sufficient conditions for the global controllability of conservative systems on a compact manifold are presented. Other results on global controllability based on Poisson stability and local accessibility are presented in [1], [2], and [7]. The notion of weak positive Poisson stability is employed in [18] and [22] for the analysis of the controllability of control affine systems.

In this paper, a form of open local accessibility and coaccessibility for nonlinear control systems is introduced called the continuous fountain condition. Subject to the conditions that (i) the states of a system are continuous fountains and (ii) one of various recurrence conditions holds, it is shown that the system state space is (globally) controllable. As demonstrated in sections 4 and 5, the fountain condition is partially ordered with respect to the local accessibility condition and the LARC condition (i.e., it is not strictly weaker or stronger). In fact, the topological nature of the fountain

*Received by the editors September 23, 1999; accepted for publication (in revised form) January 3, 2002; published electronically January 14, 2003. Research for this paper was partially supported by NSERC grant OGP 0001329, NSERC-FCAR-Nortel grant CRD 180190, and NASA-Ames Research Center grant NAG-2-1040. This paper appeared in a preliminary form in *Proceedings of the* 37*th IEEE Conference on Decision and Control*, Tampa, FL, 1998, pp. 3575–3580.

http://www.siam.org/journals/sicon/41-5/36195.html

†Department of Electrical and Computer Engineering, McGill University, 3480 University Street, Montreal, Quebec, Canada H3A 2A7 (peterc@cim.mcgill.edu). Research for this paper was partly carried out in the Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, 1997–1998. Also affiliated with the Canadian Institute for Advanced Research.

‡Department of Electrical and Computer Engineering, McGill University, 3480 University Street, Montreal, Quebec, Canada H3A 2A7 (lemch@cim.mcgill.ca).

notion allows it to be invoked in controllability analysis even for systems that have nonsmooth dynamics. These characteristics of the fountain notion permit it a different domain of application than the classical results based on a recurrence property and the LARC condition. Furthermore, for various classes of smooth systems, we establish several algebraic conditions that give straightforward algorithmic methods for verifying whether a system state is fountain.

In addition to the general relevance of these results to the study of nonlinear systems, they have value in *hierarchical hybrid control theory* (HHCT) (see section 8). This is because the theory presented in [5], [15], [16], and the references therein invokes the so-called hybrid in-block controllability (HIBC) hypothesis; this requires that each of the blocks of a certain class of decompositions (called finite analytic partitions) of the system state space forms a controllable subsystem. To verify the HIBC condition using the theory of this paper, it is sufficient to establish that the fountain condition and one of the recurrence conditions hold. Furthermore, for the so-called energy slice partitions of Hamiltonian control systems, the dense recurrence condition under a distinguished control is an inherent property which does not need explicit verification whenever each slice is precompact.

The paper is organized as follows. In section 2, we introduce the notion of a *(continuous) fountain*; then we show that, for a large class of nonlinear systems, the fountain condition taken together with an additional condition such as (i) existence of orbits, (ii) control recurrence, or (iii) weak positive Poisson stability, implies the global controllability of the systems. In section 3, we generalize the results obtained in section 2 by employing the weaker notion of a *fountain with respect to a trajectory*. In sections 4 and 5, we investigate the relation of the fountain property to that of local accessibility and, for a control affine system, to the LARC condition. It is shown that, in general, the fountain condition is not strictly weaker or stronger than the local accessibility and the LARC conditions. In section 6, some algebraic criteria for verifying the fountain property are presented. These include controllable linearization, symmetry, and full rank conditions. Section 7 is devoted to the global controllability of Hamiltonian control affine systems which satisfy the fountain condition. In this context, the developed results are applied to certain subsets of the state space of Hamiltonian control systems called energy slices in order to establish their controllability. Furthermore, applications of the results of this paper to HHCT as presented in [5], [15], and [16] are discussed in section 8, and, finally, in section 8.2, the theory is illustrated by use of a linear mass-spring system example.

2. Fountains, recurrence, and controllability. We consider nonlinear differential systems on an open connected state space $E \subset \mathbb{R}^n$ of the form

$$\mathcal{S}: \quad \begin{array}{l} \dot{x} = f(x,u), \quad x(0) = x_0 \in E, \\ x \in E,\ u \in \mathbb{R}^m,\ f \in C^r(E \times \mathbb{R}^m, \mathbb{R}^n), \\ r \in \{1, 2, \cdots, \infty, \omega\}, \end{array} \tag{2.1}$$

where C^ω denotes the class of analytic functions on E, and where the set of *admissible control functions* $\mathcal{U}$ satisfies either

(a) $\mathcal{U} = \mathcal{U}^q(\mathbb{R}; \mathbb{R}^m)$, $q \in \{1, 2, \ldots, \infty, \omega\}$, the set of all $\mathbb{R}^m$ valued bounded piecewise C^q functions of time which are continuous from the right, or

(b) $\mathcal{U} = \mathcal{U}^q(\mathbb{R}^n; \mathbb{R}^m)$, $q \in \{1, 2, \ldots, \infty, \omega\}$, the set of all $\mathbb{R}^m$ valued bounded C^q functions of $x \in \mathbb{R}^n$.

Unless otherwise stated, $\mathcal{U} = \mathcal{U}^q(\mathbb{R};\mathbb{R}^m)$. We note, however, that, if u is specified to lie in the state feedback class $\mathcal{U}^q(\mathbb{R}^n;\mathbb{R}^m)$, then, as a function of time along the system trajectory, the resulting function $u(\cdot)$ lies in $\mathcal{U}^q(\mathbb{R};\mathbb{R}^m)$. Furthermore, *unless otherwise stated*, all definitions, theorems, etc. are stated for some system $\mathcal{S}$ of the form (2.1) with the class of controls $\mathcal{U}$.

Take an arbitrary control $u \in \mathcal{U}$ and any $\varepsilon > 0$. We shall say that $\bar{u}$ is an ε-*approximation of* u if (i) $\bar{u}$ belongs to the class of piecewise constant functions which are continuous from the right (denoted $\mathcal{U}_{pwc}$) and (ii) for any t in the domain of u, $\|u(t)-\bar{u}(t)\| \leq \varepsilon$. From the standard results on the continuity of a solution trajectory of a differential equation $\mathcal{S}$ with respect to small perturbations of the control function in $\mathcal{U}^q(\mathbb{R},\mathbb{R}^m)$, it follows that the solution of $\mathcal{S}$ under u can be approximated (with any required accuracy) by the solution of $\mathcal{S}$ under ε-approximating controls in $\mathcal{U}_{pwc}$. In this context, certain properties (such as local accessibility and controllability) of $\mathcal{S}$ with control functions taken in $\mathcal{U}$ can be studied by considering only piecewise constant controls.

Throughout the paper, we use the term *orbit* to denote any nontrivial, possibly self-intersecting, trajectory ϕ_x joining x to itself; i.e., for some $T > 0$, $\phi_x : [0,T] \to E$ is any nonconstant trajectory satisfying $\phi_x(0) = \phi_x(T) = x$. All sets specified in this paper are taken to be nonempty; the term "nonempty set" is merely used for emphasis or to avoid ambiguity. In addition, the symbol $\subset$ denotes not necessarily strict set inclusion.

DEFINITION 2.1. *The set* $A_T^V(x)$ *of* accessible (at time T) from x states (with respect to $V \subset E$) *is defined by*

$$A_T^V(x) \triangleq \{z \in E;\ \exists u \in \mathcal{U},\ \ x(0) = x,\ \phi(T,x,u) = z,\ \phi(t,x,u) \in V \text{ for all } 0 \leq t \leq T\}.$$

The set of accessible states (at time T) from x *is* $A_T(x) \triangleq A_T^E(x)$*; further,* $A^V(x) \triangleq \bigcup_{0\leq T<\infty} A_T^V(x)$ *with* $A(x) \triangleq A^E(x)$. *The set* $R_T^V(x)$ *of* reachable states from x in time T (with respect to V) *is defined by* $R_T^V(x) \triangleq \bigcup_{0\leq t\leq T} A_t^V(x)$. *A state x is called* locally accessible *if the set* $R_T^V(x)$ *contains an open set for every neighborhood V and every time $T > 0$.*

Dually, we have the following definitions.

DEFINITION 2.2. *The set* $CA_T^V(x)$ *of* coaccessible (at time T) to x states (with respect to $V \subset E$) *is defined by*

$$CA_T^V(x) \triangleq \{z \in E;\ \exists u \in \mathcal{U},\ \ x(0) = z,\ \phi(T,z,u) = x,\ \phi(t,z,u) \in V \text{ for all } 0 \leq t \leq T\}.$$

The set of coaccessible states (at time T) to x *is* $CA_T(x) \triangleq CA_T^E(x)$*; further,* $CA^V(x) \triangleq \bigcup_{0\leq T<\infty} CA_T^V(x)$ *with* $CA(x) \triangleq CA^E(x)$.

DEFINITION 2.3. *We say that E is* controllable (with respect to the class of controls $\mathcal{U}$) *if $A(x) = E$ for all $x \in E$, and that E is* controllable on the interval $[0,T]$ *if $A_T(x) = E$ for all $x \in E$. We say that the system is* locally controllable *at p if, for all sufficiently small $\rho' > 0$, there exists ρ, $0 < \rho < \rho'$, such that any two states x, x' in the open ball $B_\rho(p)$ are mutually accessible with respect to $B_{\rho'}(p)$.*

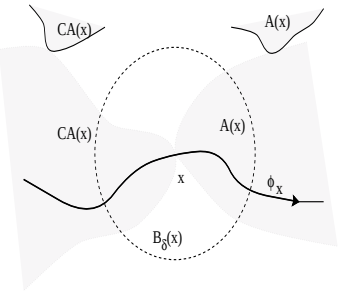


FIG. 2.1. *The fountain condition is satisfied at* x.

DEFINITION 2.4. *A state* $x \in E$ *is called a* fountain *if*

(i) *there exists* $\mu > 0$ *such that, for all* δ, $0 < \delta < \mu$, *the open ball neighborhood* $B_\delta(x)$ *is contained in* E *and* $A^\delta(x) - \{x\}$ *is an open set, where* $A^\delta(x) \underline{\Delta} A^{B_\delta(x)}(x)$ *(in which case* x *is called a* positive fountain*); and*

(ii) *there exists* $\mu' > 0$ *such that, for all* δ', $0 < \delta' < \mu'$, $B_{\delta'}(x) \subset E$ *and* $CA^{\delta'}(x) - \{x\}$ *is an open set, where* $CA^{\delta'}(x) \underline{\Delta} CA^{B_{\delta'}(x)}(x)$ *(in which case* x *is called a* negative fountain*).*

Finally, $\sup\{\mu$; *such that* (i) *holds at* $x\}$ *shall be denoted* $\rho^+(x)$, *and* $\sup\{\mu'$; *such that* (ii) *holds at* $x\}$ *shall be denoted* $\rho^-(x)$; *when* (i) *and* (ii) *hold at* x, *that is, when* x *is a fountain,* $\min(\rho^+(x), \rho^-(x))$ *shall be denoted* $\rho(x)$.

If a set of states $C \subset E$ *(in particular,* E*) is such that each* $x \in C$ *is a fountain, the set* C *is called a fountain.*

Evidently, the fountain definition above requires that, locally (i.e., with respect to all sufficiently small $B_\delta(x)$) the entire accessible set (from a given x, less $\{x\}$) is open; this implies, in particular, that a trajectory ϕ_x passing through x must depart from x within the interior of the accessible set. Clearly, the definition also requires that, locally, the entire coaccessible set (to x, less $\{x\}$) is open, and so such a trajectory ϕ_x must arrive at x within the interior of the coaccessible set. (See Figure 2.1, where the accessible and coaccessible sets are not open, but, at the same time, there exists a sufficiently small neighborhood $B_\delta(x)$ such that the accessibility and coaccessibility sets are open with respect to that neighborhood.)

Example 2.1. Consider the *double integrator* system

$$\begin{aligned} \dot{x} &= u, \quad x(0) = x_0, \\ \dot{y} &= x, \quad y(0) = y_0, \end{aligned} \tag{2.2}$$

where we take $E = \mathbb{R}^2$. Let $p = (x_0, y_0)$ be an arbitrary point in E. In the case when $x_0 > 0$, take $\delta > 0$ sufficiently small so that $B_\delta(p) \subset \{(x, y);\ x > 0\}$. Then, with respect to this δ-ball neighborhood, the accessible set $A^\delta(p) = (\{(x, y);\ y > y_0\} \cup \{p\}) \cap B_\delta(p)$, and the coaccessible set $CA^\delta(p) = (\{(x, y);\ y < y_0\} \cup \{p\}) \cap B_\delta(p)$. This is readily verified by use, for instance, of the parabolic family of controlled trajectories. For any $x_0 < 0$, the situation is reversed with the accessible set lying below and the coaccessible set lying above the y_0 axis. Further, at $p = (0, y_0)$, $A^\delta(p) = B_\delta(p)$ and $CA^\delta(p) = B_\delta(p)$. In all cases, the sets $A^\delta(p) - \{p\}$ and $CA^\delta(p) - \{p\}$ are open. Hence any $p \in E$ is a fountain.

The definition of a positive fountain x is not strictly stronger or weaker than the standard definition of local accessibility from x (see [24]), in that the set of states accessible from a fountain x (relative to the ρ^+ neighborhood in arbitrary time), with

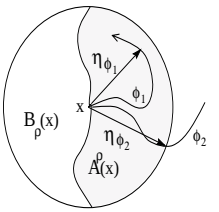


FIG. 2.2. *Throughput function* $\eta_\phi(\cdot)$.

the state x removed, is required to be open; whereas a system is *locally accessible from* x if, for every time $T > 0$ and every neighborhood V of x, the set $R_T^V(x)$ of states reachable from x in time less than or equal to T (with respect to V) contains an open set. The relation of the fountain property to that of local accessibility shall be further discussed in section 4.

Remark. A set of fountains is not necessarily open, which is shown, for instance, by the differential system $\dot{x} = g(x,y)u$, $\dot{y} = 1$, $E = \mathbb{R}^2$, where the function

$$g : E \to \mathbb{R} \text{ is defined to be } g(x,y) = \begin{cases} (x^2 - y^2)^2 & \text{if } x^2 < y^2, \\ 0 & \text{otherwise.} \end{cases}$$

DEFINITION 2.5. *Let $C \subset E$ be an open set of fountains. A state $x \in C$ is called a* continuous positive fountain *if $\rho^+(x)$ is continuous at x, and it is called a* continuous negative fountain *if $\rho^-(x)$ is continuous at x, where a function which is unbounded for all $x \in E$ is taken to be continuous. A fountain is* continuous *if it is both a continuous positive and a continuous negative fountain.*

The set C is called a continuous fountain *if each $x \in C$ is a continuous fountain.*

We observe that, if x is a continuous fountain, then $\rho(x)$ is continuous at x.

For a trajectory ϕ through x, we now introduce a measure of the maximum radius of a ball centered at x which contains an entire segment of ϕ lying in $A^{\rho(x)}(x)$ and $CA^{\rho(x)}(x)$ and is such that the end points of the trajectory segment lie in the boundary of the ball.

DEFINITION 2.6. *For a fountain $x \in E$, the* throughput function on the trajectory $\phi(t,x,u)$ *is defined by $\eta_\phi(x) \underline{\Delta} \sup\{\mu,\ \mu \le \rho(x)$; there exists $T_1 > 0$ such that, for all τ, $0 \le \tau < T_1$, $\phi(\tau,x,u) \in A^\mu(x)$ with $\phi(T_1,x,u) \in \partial B_\mu(x)$; and there exists $T_2 < 0$ such that, for all τ, $T_2 < \tau \le 0$, $\phi(\tau,x,u) \in CA^\mu(x)$ with $\phi(T_2,x,u) \in \partial B_\mu(x)\}$. If $\eta_\phi(x)$ is strictly positive and continuous at x, then x is said to have* continuous strictly positive throughput on the trajectory ϕ.

We observe that throughput functions are strictly positive on any nontrivial trajectory. The definition above is illustrated in Figure 2.2, where η_{ϕ_1}, η_{ϕ_2} represent throughput functions on different trajectories.

LEMMA 2.7. *Let C be an open set of continuous fountains. Then every $x \in C$ has a continuous strictly positive throughput $\eta_\phi(x)$ on any nontrivial trajectory ϕ passing through x; that is, $\eta_\phi(x) > 0$ and*

$$\lim_{n\to\infty} \|\eta_\phi(y_n) - \eta_\phi(x)\| = 0$$

for any sequence $\{y_n;\ y_n \in \phi\}$ converging to x, as $n \to \infty$.

THEOREM 2.8. *Assume that a system $\mathcal{S}$ on the open connected state space E is such that E is a continuous fountain and through each x in E there exists a nontrivial orbit. Then E is controllable.*

Proof. Let Φ_x denote the set of orbits through x. First, we claim that the set $O_x \underline{\Delta} \{z;\ \exists \phi_x \in \Phi_x \text{ s.t. } z \in \phi_x\}$ is open. This is established by showing that any $z \in O_x$ has a neighborhood $N(z)$ contained in O_x.

Consider any $\phi_x \in \Phi_x$. Take an arbitrary point $z \in \phi_x$, and let $z = \phi_x(\tau)$, $0 \leq \tau \leq T$, where the continuous map $\phi_x(\cdot)$ takes $[0,T]$ into the orbit ϕ_x with $\phi_x(0) = \phi_x(T) = x$. Since all states in E are continuous fountains and since ϕ_x is compact (as a continuous image of $[0,T]$ in E), there exists a strictly positive minimum $\eta > 0$ for the function $\eta_{\phi_x}(\cdot)$, which is continuous on ϕ_x by Lemma 2.7.

Then there exist τ_a and τ_b, $\tau_a < \tau < \tau_b$, such that $a = \phi_x(\tau_a) \in B_\eta(z)$ and $b = \phi_x(\tau_b) \in B_\eta(z)$. By the definition of η on ϕ_x, the point z lies in the open set $(A^\eta(a) - \{a\})$ and lies in the open set $(CA^\eta(b) - \{b\})$. Hence there exist open neighborhoods $N_1(z)$ and $N_2(z)$ such that $N_1(z) \subset (A^\eta(a) - \{a\})$ and $N_2(z) \subset (CA^\eta(b) - \{b\})$. Define $N(z) = N_1(z) \cap N_2(z)$. Then, for any $p \in N(z)$, there exists a trajectory from a to p and from p to b; i.e., p lies on a nontrivial orbit through x. Hence $N(z) \subset O_x$. Since this holds for any $\phi_x \in \Phi_x$ and any $z \in \phi_x$, we conclude that O_x is open.

Second, we say that $x, y \in E$ are equivalent, denoted $x \sim y$, if there exists an orbit ϕ such that $x, y \in \phi$. It is clear that $\sim$ is reflexive, symmetric and transitive. Hence the relation $\sim$ defines an equivalence relation on the set E. Thus there exists a partition of the set E into disjoint equivalence classes denoted by $[\cdot]$. For any $x, y \in E$, $y \in [x]$ if and only if $y \in O_x$, i.e., $[x] = O_x$. It follows that the state space E is the disjoint union of open sets. However, E is connected, and hence E consists of just one such orbit class. Now all states $x, y \in E$ lie on orbits passing through each other, and so $A(x) = E$ for all $x \in E$; i.e., E is controllable. □

Example 2.2. Consider the double integrator system described in Example 2.1. As has been shown, every state $p = (x, y) \in \mathbb{R}^2$ is a fountain. Moreover, $\rho(p)$ is unbounded at each p, and hence all states are continuous fountains. Further, applying the state dependent control function $u(x, y) = -y$, we see that there exists a nontrivial orbit trajectory through every state in $\mathbb{R}^2$. Hence, by Theorem 2.8, the system (2.2) is controllable in the space $\mathbb{R}^2$.

2.1. Control recurrence.

DEFINITION 2.9. *A state $x \in E$ is called* control recurrent *if it lies in its nontrivial positive limit set under some control $u \in \mathcal{U}$; i.e., x is not an equilibrium point under u and $x = \lim_{n\to\infty} \phi(t_n, x, u)$ for some sequence $\{t_n;\ n = 1, 2, \cdots\}$ such that $\lim_{n\to\infty} t_n = \infty$.*

Versions of control recurrence have been employed by Lobry [19], [20], Kunita [12], and Jurdjevic [11] to establish the controllability of control affine systems. Using the notion of control recurrence, we obtain the following generalization of Theorem 2.8.

THEOREM 2.10. *Assume that a system $\mathcal{S}$ on the open connected state space E is such that E is a continuous fountain, and, for every x in E there exists a control $u_x \in \mathcal{U}(\mathbb{R}^n; \mathbb{R}^m)$ such that x is control recurrent under u_x. Then E is controllable.*

Proof. Let x be an arbitrary point in E. Consider the associated $u_x(\cdot)$ controlled trajectory ϕ through x. By the fountain condition, there exists a strictly positive radius δ such that the set $N^\delta(x) \underline{\Delta} (CA^\delta(x) - \{x\})$ is open. Take any p such that $p \in \phi$ and $p \in N^\delta(x)$; such a p exists since ϕ is a nontrivial trajectory. Further, let the time interval for the state to pass from p to x be ε. Since x lies in the closure of the forward trajectory, there exists a sequence of instants $\{t_n;\ n = 1, 2, \ldots\}$ such that $\lim_{n\to\infty} t_n = \infty$ and $\lim_{n\to\infty} \phi(t_n, x, u) = x$.

Now consider the sequence $\{\phi(t_n - \varepsilon, x, u_x);\ n = 1, 2, \ldots\}$. By the continuity of solution trajectories with respect to final conditions, this sequence of states converges to p as $n \to \infty$. Hence, for some sufficiently large N, the trajectory ϕ through x over the time interval $[t_n - \varepsilon;\ t_n]$ for all $n > N$ lies in the neighborhood $N^\delta(x)$. Hence, for any such n, the state $\phi(t_n - \varepsilon, x, u_x)$ can be joined to x by a trajectory, using a possibly time-dependent control, where this trajectory segment lies entirely in $N^\delta(x) \cup \{x\}$. This yields a nontrivial orbit through x, and the same construction yields an orbit through any state $y \in E$.

Applying Theorem 2.8, we conclude that E is controllable. □

A vector field F is called *positively Poisson stable* (*PPS*) *on* E (see [24]) if the set of all states $x \in E$ which are control recurrent under F is dense in E. Positive Poisson stability is used by various authors in order to establish the controllability of nonlinear systems. In particular, in [12], it is proven that, if the drift vector field of a control affine system is PPS and the system possesses the LARC condition, then the system is (globally) controllable.

We note that, in the case of the existence of a control u such that the vector field $F \underline{\Delta} f(\cdot, u(\cdot))$ is PPS (in particular, in the case of a control affine system with a PPS drift), the recurrence condition of Theorem 2.10 is satisfied.

DEFINITION 2.11. *A set R is called a* uniform control recurrent set *if there is some state dependent control v such that each $x \in R$ lies in its nontrivial positive limit set under v.*

THEOREM 2.12. *Assume that a system $\mathcal{S}$ on the open connected state space E is such that E is a continuous fountain and there exists a uniform control recurrent set R which is dense in E. Then E is controllable.*

Proof. Take any $x \in E$. By the fountain condition, there exists a strictly positive ρ such that the accessible set $(A^\rho(x) - \{x\})$ and the coaccessible set $(CA^\rho(x) - \{x\})$ are open.

Since R is dense in E, there exists a sequence $\{x_n;\ n = 1, 2, \cdots\} \subset R \cap (CA^\rho(x) - \{x\})$ of control recurrent states converging to x. All x_n are continuous fountains; hence, as has been proved in Theorem 2.10, each x_n lies on a nontrivial orbit ϕ_{x_n}. Moreover, by the recurrence property, the cross-over link creating each orbit can be chosen so that it forms a segment with x_n as the right-hand end point. Hence there exists some interval $[0, \varepsilon]$, where ε is independent of n, such that the trajectory segments commencing at each x_n, subject to the control v, converge uniformly to the trajectory subject to v through x, as $n \to \infty$. So, for some sufficiently large N, the orbit ϕ_{x_N} passes through the set $(A^\rho(x) - \{x\})$, and, since each x_n lies in $(CA^\rho(x) - \{x\})$, the trajectory also passes through this latter set.

Consider any two points p_1 and p_2 such that $p_1, p_2 \in \phi_{x_N}$, $p_1 \in (A^\rho(x) - \{x\})$, $p_2 \in (CA^\rho(x) - \{x\})$. Then there exists an orbit ϕ_x from x to p_1 (since p_1 is accessible from x) to p_2 (since $p_1, p_2 \in \phi_{x_N}$), and then back to x (since p_2 is coaccessible to x).

Hence every state $x \in E$ lies on a nontrivial orbit. Applying Theorem 2.8, we conclude that E is controllable. □

2.2. Weak positive Poisson stability. Let F be a C^r, $r \geq 1$, vector field defined on $E \times \mathbb{R}^1$, and let $\phi^F(t, x)$ denote the flow of F at the time t with the initial condition x; i.e.,

(i) $\phi^F(0, x) = x$,

(ii) $\frac{d}{dt}\phi^F(t, x)|_{t=\tau} = F(\phi^F(\tau, x), \tau)$ for all $\tau \in \mathbb{R}^1$.

Define the set $\phi^F(t, A)$ for $A \subset E$ to be $\{z;\ \exists y \in A,\ \ z = \phi^F(t, y)\}$.

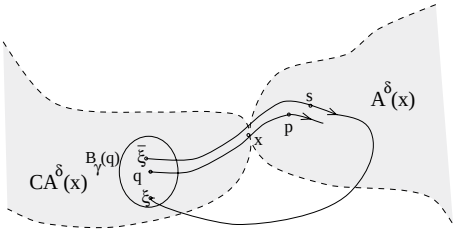


FIG. 2.3. *Theorem* 2.15: *Weak positive Poisson stability.*

DEFINITION 2.13 (see [18]). *A state $x \in E$ is called a* nonwandering state *of F if, for any $T > 0$ and any neighborhood $N(x) \subset E$, there exists a time instant $t > T$, such that $\phi^F(t, N(x)) \cap N(x) \neq \emptyset$. In other words, $\phi^F(t, y) \in N(x)$ for some $y \in N(x)$.*

DEFINITION 2.14 (see [18]). *A smooth (i.e., C^r, $r \geq 1$) vector field F is called* weakly positively Poisson stable (PPS) *if the set of all states which are nonwandering points of F is dense in E.*

The notion of weak positive Poisson stability is employed in [18] and [22] for the analysis of the controllability of control affine systems.

THEOREM 2.15. *For a system $\mathcal{S}$, let there exist a control $\tilde{u} \in \mathcal{U}(\mathbb{R}^n, \mathbb{R}^m)$ such that the vector field $\tilde{X} = f(\cdot, \tilde{u})$ is weakly PPS. Further, let E be a continuous fountain. Then*

(i) *there exists a nontrivial orbit passing through each $x \in E$, and*
(ii) *E is controllable.*

Proof.

(i) Let x be an arbitrary point in E. Choose $\rho > 0$ so that $(A^\rho(x) - \{x\})$ and $(CA^\rho(x) - \{x\})$ are open. Such ρ exists since x is a fountain. Consider the flow $\phi^{\tilde{X}}$ with the initial condition x. Take $p = \phi^{\tilde{X}}(t_1, x)$ and $q = \phi^{\tilde{X}}(t_2, x)$, where $t_1 > 0$ and $t_2 < 0$ are chosen so that $p \in A^\rho(x) - \{x\}$ and $q \in CA^\rho(x) - \{x\}$. Let $t^* = t_1 - t_2$ denote the time needed to reach p from q under the flow $\phi^{\tilde{X}}$. By the continuity of the solution trajectory with respect to initial conditions, there exists an open γ-ball neighborhood $B_\gamma(q) \subset CA^\rho(x) - \{x\}$ such that $\phi^{\tilde{X}}(t^*, B_\gamma(q)) \subset A^\rho(x) - \{x\}$.
Further, since the vector field $\tilde{X}$ is weakly PPS, there exist $\xi, \bar{\xi} \in B_\gamma(q) \subset CA^\rho(x) - \{x\}$ such that $\xi = \phi^{\tilde{X}}(T^*, \bar{\xi})$ for some $T^* > t^*$ (see Figure 2.3). By construction, there exists $s = \phi^{\tilde{X}}(t^*, \bar{\xi}) \in A^\rho(x) - \{x\}$. Hence $\xi = \phi^{\tilde{X}}(T^*, \bar{\xi}) = \phi^{\tilde{X}}(T^* - t^*, s)$, where $T^* - t^* > 0$; i.e., ξ is accessible from s. Therefore, given the facts that ξ is coaccessible to x and s is accessible from x, we conclude that there exists an orbit from x to s to ξ to x.

(ii) The (global) controllability of the state space E follows from part (i) and Theorem 2.8. ∎

Each recurrent point of a vector field F is a nonwandering point of F. Hence we observe that, if a vector field F is PPS (i.e., the set of recurrent points is dense in E), then the nonwandering set is E.

3. Fountains with respect to trajectories. In this section, we assume there exists a compact set $U \subset \mathbb{R}^m$ such that all functions in the set of admissible control functions $\mathcal{U}$ take values in U.

DEFINITION 3.1. *Let x be an arbitrary state in E, and let ϕ_x be an arbitrary trajectory of a system $\mathcal{S}$ passing through x. The state x is called a* fountain with respect to the trajectory ϕ_x *if there exists a time instant $T = T(x, \phi_x) > 0$ such that*

(i) *$\phi_x((0;T))$ lies in the interior of $A(x)$, and*

(ii) *$\phi_x((-T;0))$ lies in the interior of $CA(x)$.*

Further, if $T^* \triangleq \sup\{T;\ \phi_x((0;T)) \subset [A(x)]^\circ$ and $\phi_x((-T;0)) \subset [CA(x)]^\circ\}$ is a continuous function of x, then x is called a continuous fountain with respect to the trajectory ϕ_x.

Note that, by the compactness hypothesis on the set U, T^* is always finite.

Let $x \in E$ be a fountain. Then, as has been noted in section 2, each nontrivial trajectory ϕ_x passing through x has to depart from x within the interior of the accessible set, and it has to arrive at x within the interior of the coaccessible set. Hence x is a fountain with respect to ϕ_x, so Definition 3.1 is a weaker version of Definition 2.4. To illustrate this fact, we consider a simple differential system acting on $\mathbb{R}^2$,

$$\begin{aligned} \dot{x} &= u + 1, \\ \dot{y} &= u^2 x, \end{aligned} \tag{3.1}$$

with the initial condition $p_0 = (1, 0)$. Take a neighborhood $B_\delta(p_0)$ in the open right-half plane $\mathbb{R}^2_+$. Then the accessible set $A^\delta(p_0)$ for all sufficiently small $\delta > 0$ contains a boundary, namely, the set $D \triangleq \{(x, y);\ x > 1; y = 0\} \cap B_\delta(p_0)$. This is because each state p in the topological boundary of D is accessible from p_0 using the control $u = 0$. On the other hand, $A(p_0) = \mathbb{R}^2$, and hence each trajectory lies within the interior of the accessible set. This shows that, for the system (3.1), the state p_0 is not a fountain, but at the same time p_0 is a positive (and, similarly, negative) fountain with respect to any trajectory passing through p_0.

Observe that the fountain property with respect to a specified trajectory is less restrictive than the notion of the *local controllability along a reference trajectory* ϕ (see, for example, [9]), in which the requirement is that all points in some open neighborhood of $\phi(t, x_0)$, $t \geq 0$, can be reached at time t by solutions initiating from x_0. In [23], a sufficient condition for the local controllability along a reference trajectory of an affine control system is formulated in the special case where the reference trajectory is a closed orbit.

THEOREM 3.2. *Assume that a system $\mathcal{S}$ on an open connected state space E is such that, through each state $x \in E$, there exists a nontrivial orbit and every $x \in E$ is a continuous fountain with respect to any orbit $\phi_x \in \Phi(x)$ passing through x. Then E is controllable.*

Proof. Let Φ_x denote the set of orbits through x. First, as in the proof of Theorem 2.8, we claim that the set $O_x \triangleq \{z;\ \exists \phi_x \in \Phi_x$ s.t. $z \in \phi_x\}$ is open. Consider any $\phi_x \in \Phi_x$. Take an arbitrary point $z \in \phi_x$, and let $z = \phi_x(\tau)$, $0 \leq \tau \leq T$, where the continuous map $\phi_x(\cdot)$ takes $[0, T]$ into the orbit ϕ_x with $\phi_x(0) = \phi_x(T) = x$. Since z is a continuous fountain with respect to all orbits passing through z, it is possible to find some q and p on ϕ_x such that some nonempty open neighborhood $N(z)$ is a subset in $[A(q)]^\circ \cap [CA(p)]^\circ$ and hence $N(z) \subset O_x$. We conclude that O_x is open.

As in the proof of Theorem 2.8, the fact that the set O_x is open for any x implies the global controllability of the state space E. ∎

The notion of the fountain with respect to a trajectory can be utilized to establish results similar to the results formulated earlier in this section, namely, in Theorems 2.10, 2.12, and 2.15.

4. Fountain condition for control affine systems. Consider a system $\mathcal{S}$ of the *input-linear* or *control affine* class of nonlinear time-invariant control systems (see [24]),

$$\mathcal{S}: \quad \dot{x} = f(x) + \sum_{i=1}^{m} g_i(x)u_i, \tag{4.1}$$

where we assume that $f, g_1, g_2, \ldots, g_m$ are smooth mappings from $\mathbb{R}^n$ into $\mathbb{R}^n$.

The question of global controllability for control affine systems has been extensively addressed. One of the essential conditions, used for instance in [18] and [20], to establish the controllability of such systems is the (Chow–Hermann) LARC. In [20], it has been proven that any *Poisson stable* system satisfying the LARC is controllable. An extension of this result is presented in [22], where the Poisson stability condition is weakened to the requirement of Poisson stability on the drift vector field $f(x)$ only. One of the most recent of this set of results is to be found in the paper by Lian, Wang, and Fu [18].

THEOREM 4.1 (see [18]). *Consider the system* (4.1) *on the state space* E. *Suppose that* f *is a weakly PPS vector field (see Definition* 2.14*). Then* E *is controllable if the LARC is satisfied.*

Related results can be also found in [3], [1], and [8]. In [21], the LARC condition is used to establish the small-time local controllability for a planar body with unilateral thrusters.

Consequently, it is of interest to investigate the relation between the LARC and the fountain notion. In this section, we show that the fountain condition for control affine systems does not imply the LARC condition and, conversely, the LARC condition does not imply the fountain condition. These facts are illustrated by the following examples.

Example 4.1. Let $h_1(x, y)$, $h_2(x, y)$ be $C^\infty(\mathbb{R} \times \mathbb{R})$ functions such that

(i) $h_1(x, y) \neq 0$, if $(x, y) \in S_1$, and $h_1(x, y) = 0$ if $(x, y) \in \mathbb{R}^2 - S_1$, and

(ii) $h_2(x, y) \neq 0$, if $(x, y) \in S_2$, and $h_2(x, y) = 0$ if $(x, y) \in \mathbb{R}^2 - S_2$,

where $S_1 \subset \mathbb{R}^2$ is the open unit disk with the center at (0,1) and $S_2 \subset \mathbb{R}^2$ is the open unit disk with the center at $(0, -1)$.

Consider the following control affine system:

$$\begin{aligned} \dot{x} &= (h_1(x, y) + h_2(x, y))u, \\ \dot{y} &= 1. \end{aligned} \tag{4.2}$$

Then the accessible set from $p_0 = (0, 0)$ and the coaccessible set to p_0 are such that $A^\delta(p_0) - \{p_0\}$ and $CA^\delta(p_0) - \{p_0\}$ are open for any $\delta > 0$. Hence p_0 is a fountain. On the other hand, the LARC fails at p_0 since the function $g(x, y) \underline{\Delta} h_1(x, y) + h_2(x, y)$ and all of its partial derivatives are equal to zero at p_0 and thus the dimension of the associated Lie algebra of the system (4.2) at p_0 is equal to 1.

Example 4.2. Consider the differential system

$$\begin{aligned} \dot{w} &= z^2 I w + J w, \\ \dot{z} &= u, \end{aligned} \tag{4.3}$$

where $w = [x, y]^T$, $I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, and $J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$.

The system in this example is in the class of control affine systems described by (4.1) with $f = [(z^2 x - y) \quad (z^2 y + x) \quad 0]^T$, $m = 1$, and $g = [0 \quad 0 \quad 1]^T$.

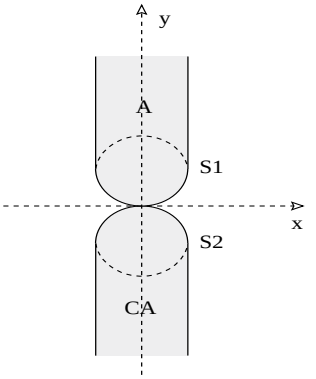


FIG. 4.1. *Example* 4.1: *Fountain condition does not imply LARC.*

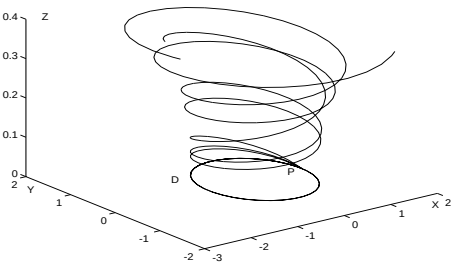


FIG. 4.2. *Example* 4.2: *Fountain condition fails, while the LARC property holds at p.*

The LARC condition holds at the state $p = (1, 0, 0)$. Indeed, $[[f, g], g] = [2x\ 2y\ 0]^T$, and hence, at p, $\dim L(f, g) = \dim span\{[[f, g], g], f, g, \} = 3$.

However, the fountain condition fails at p since the accessible set from p with respect to any neighborhood $B_\delta(p)$ contains accessible boundary states that lie in the set $D \underline{\Delta} \{(x, y, 0);\ x^2 + y^2 = 1\}$ (see Figure 4.2). To verify this, we note that $u \equiv 0$ implies $z \equiv 0$, in which case the motion is a rotation around the unit circle, and so there certainly exist states $p' \epsilon D \cap B_\delta(p)$ accessible under the zero control from p. Next we consider the function $R \underline{\Delta} \frac{1}{2}(x^2 + y^2)$ along any system trajectory. Since $dR = 2z^2 R dt$, R is nondecreasing along all trajectories, and hence any state $p' \epsilon \{(x, y, 0); x^2 + y^2 < 1\}$ is not accessible from p. It follows that any accessible point $p' \epsilon D \cap B_\delta(p)$ does not lie in the interior of the accessible set from p which is consequently not a fountain. □

The above examples show that the fountain condition is neither strictly stronger nor strictly weaker than the LARC.

5. Fountains and local accessibility. As has been noted in section 2, the definition of a positive fountain x is not strictly stronger or weaker than the standard definition of local accessibility from x (see [24]). To illustrate this fact, we present in this section two examples in which only one of these properties is satisfied.

Example 5.1.

$$\dot{x} = \sin^2 u + v_1 z, \quad \dot{y} = \sin^4 u + v_2 z, \quad \dot{z} = g(x, y, z) v_3,$$

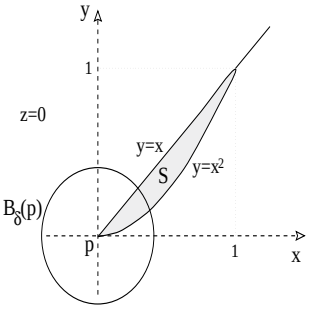


FIG. 5.1. *Example* 5.1: *The local accessibility property fails at the fountain* $p = (0, 0, 0)$.

where $w \underline{\Delta} (u, v_1, v_2, v_3) \epsilon \mathcal{U}$ and the function $g \epsilon C^s(\mathbb{R}^3; \mathbb{R}^1)$, $s \geq 1$, is defined in such a way that $g(x, y, z) = 0$ if $(x, y, z) \epsilon S \underline{\Delta} \{(x, y, 0);\ 0 \leq x \leq 1, x^2 \leq y \leq x\}$ (S is shown in Figure 5.1) and $g(x, y, z) > 0$ otherwise.

For the initial condition $p = (0, 0, 0)$ and an arbitrary control $w \epsilon \mathcal{U}$, denote the corresponding system trajectory as $(x(t), y(t), z(t))$, $t \geq 0$. Then, for any $t \epsilon [0, 1]$,

$$x(t) = \int_0^t \sin^2 u(\tau, x(\tau))d\tau, \quad y(t) = \int_0^t \sin^4 u(\tau, x(\tau))d\tau, \quad z(t) = 0. \tag{5.1}$$

It can be shown (using the Cauchy–Schwarz inequality) that $x(t)$ and $y(t)$ defined by (5.1) satisfy

$$x^2(t) \leq y(t) \leq x(t),$$

and hence $(x(t), y(t), z(t)) \epsilon S$ for any $t \epsilon [0, 1]$.

Hence the system trajectories under various control functions w stay in the set S for any $0 \leq t \leq 1$. This set has an empty interior, and hence the local accessibility condition fails at the origin.

On the other hand, under constant control u and $(v_1, v_2, v_3) = (0, 0, 0)$, the trajectory is such that $y(t) = ax(t)$ (where $a = \sin^2 u$) and $y(1) = x^2(1)$. Hence, for any $\delta > 0$ and any time $T > 1$, there exists a control w under which the trajectory can leave S (in time less than T) without leaving $B_\delta(p)$ and, since $g \neq 0$ outside of S, attain any point in $B_\delta(p)$.

We conclude that the set of accessible from p states (with respect to an arbitrary δ-ball neighborhood $B_\delta(p)$, $\delta > 0$) is equal to the whole $B_\delta(p)$. Hence the fountain condition at the origin is satisfied.

Example 5.2. Consider the system

$$\begin{aligned} \dot{x} &= u^2, \quad x(0) = x_0, \\ \dot{y} &= 1, \quad y(0) = y_0, \end{aligned}$$

where we take $E = \mathbb{R}^2$.

For any $\delta > 0$, the accessibility set is $A_T^\delta(x_0, y_0) = \{(x, y) \epsilon \mathbb{R}^2;\ x \geq x_0; y > y_0; T \geq 0\}$. This set has a nonempty interior, and hence the local accessibility property is satisfied for any state $(x_0, y_0) \epsilon E$. On the other hand, the accessibility set (with the initial state (x_0, y_0) removed) is not open, and thus the positive fountain condition fails.

For nonlinear systems for which the fountain condition is satisfied at some state $x_0 \epsilon E$, it is interesting to investigate under what additional assumptions the local

accessibility property is satisfied at x_0, as well. In this connection, we state the following two lemmas.

LEMMA 5.1 (see [14]). *Assume that, for a system $\mathcal{S}$, there exists a state $x_0 \epsilon E$ at which the fountain condition is satisfied but the local accessibility condition fails. Then*

$$\inf_{u\epsilon\mathcal{U}} \max_{t\epsilon[0,T^*]} \|\phi(t, x_0, u) - x_0\| = 0 \tag{5.2}$$

for some T^*, $0 < T^* < \infty$.

LEMMA 5.2 (see [14]). *Assume that, for a system $\mathcal{S}$, there exists a state $x_0 \epsilon E$ at which the fountain condition is satisfied but the local accessibility condition fails. Then*

$$\inf_{u\epsilon U} \|f(x_0, u)\| = 0, \tag{5.3}$$

where $f(\cdot,\cdot)$ is the vector field of the system $\mathcal{S}$ and $U = \mathbb{R}^m$ is the set of values of admissible control functions.

We conclude that the fountain property, taken together with either of the conditions

$$\text{for all } T^* > 0, \quad \inf_{u\epsilon\mathcal{U}} \max_{t\epsilon[0,T^*]} \|\phi(t, x_0, u) - x_0\| > 0 \tag{5.4}$$

or

$$\inf_{u\epsilon U} \|f(x_0, u)\| > 0, \tag{5.5}$$

implies local accessibility.

6. Algebraic criteria for the fountain condition. In this section, we treat certain classes of nonlinear systems for which it is possible to verify the fountain condition without analyzing the actual geometry of the accessible sets.

6.1. Controllable linearizations. In this section, we take $\mathcal{U}$ to be the set of all admissible time-dependent controls $\mathcal{U}^s(\mathbb{R}; \mathbb{R}^m)$, $s \geq 1$. Let p be an arbitrary state in E, and let a state $q \epsilon E$ be accessible from p. Then there exist a control function $u^0 \epsilon \mathcal{U}$ and a time instant T such that $q = \phi(T, p, u^0)$. Consider the linearization of the system $\mathcal{S}$ along the trajectory $\phi(t, p, u^0)$ on the time interval $[0, T]$ given by

$$\frac{d}{dt} z = A(t)z + B(t)v, \qquad t \epsilon [0, T]. \tag{6.1}$$

The matrices A and B are defined on $[0, T]$ as the derivatives of f with respect to the arguments x and u, respectively:

$$A(t) \underline{\Delta} \left[\frac{\partial f(x,u)}{\partial x}\right] \Big|_{\substack{x=\phi(t,p,u^0)\\ u=u^0(t)}}, \qquad B(t) \underline{\Delta} \left[\frac{\partial f(x,u)}{\partial u}\right] \Big|_{\substack{x=\phi(t,p,u^0)\\ u=u^0(t)}}. \tag{6.2}$$

THEOREM 6.1. *Suppose that a system $\mathcal{S}$ is such that, for each $p \epsilon E \subset \mathbb{R}^n$ and any $q \epsilon A^\rho(p) - \{p\}$ $(\rho > 0)$, the linearization of $\mathcal{S}$ along a trajectory ϕ, joining p to q over a finite time interval $[0, T]$, is controllable on $[0, T]$. Then each state in E is a continuous positive fountain.*

Proof. Take an arbitrary $\rho > 0$ and any $q \epsilon A^\rho(p) - \{p\}$, where $p \epsilon E$. Let $u^0 \epsilon \mathcal{U}$ be such that $q = \phi(T, p, u^0)$ $(0 < T < \infty)$, where the system (6.1) is controllable

along $\phi(t,p,u^0)$ on $[0,T]$. Then there are controls $v_1, v_2, \dots, v_n \in \mathcal{U}$ defined on $[0,T]$ such that the corresponding states $z_i(\cdot)$, $1 \le i \le n$, of (6.1) are linearly independent at time T; i.e.,

$$\frac{d}{dt}z_i = A(t)z_i + B(t)v_i, \qquad z_i(0) = 0, \qquad t \in [0,T]; \quad \text{and}$$
$$\text{for all } \gamma \in \mathbb{R}^n \quad [z_1(T)\cdots z_n(T)]\gamma = 0 \Rightarrow \gamma = 0.$$

Define $u(t,\gamma) = [v_1(t)\cdots v_n(t)]\gamma + u^0(t)$, $\gamma \in \mathbb{R}^n$, $t \in [0,T]$, and consider the state $z(t,p,\gamma)$ of $\mathcal{S}$ corresponding to u. Following the standard arguments (see, e.g., [25]), it can be shown that, for each $x = q + \delta q \in N(q)$, there exist $\gamma = g(x) \in \mathbb{R}^n$ and hence a control $u = u(t,\gamma)$ such that the corresponding trajectory with the initial condition p reaches $q + \delta q$ at time T.

We conclude the proof by noting that

$$\begin{aligned} &\text{for all } \varepsilon > 0 \ \exists\ \delta > 0 \ \text{ such that } \|u_1 - u_2\|_{C_{[0,T]}} < \delta \Rightarrow \\ &\quad \text{for all } t \in [0,T] \ \|\phi(t,x_0,u_1) - \phi(t,x_0,u_2)\| < \varepsilon. \end{aligned} \tag{6.3}$$

Choose $\varepsilon > 0$ to be sufficiently small so that (i) the ε-cylinder of the nominal trajectory $\phi([0,T],p,u^0)$ lies in $B_\rho(p)$, i.e.,

$$\{x;\ \|\phi([0,T],p,u^0) - x\| < \varepsilon\} \subset B_\rho(p),$$

and, furthermore, (ii) $B_\varepsilon(q) \subset N(q)$. For that ε, take $\delta > 0$ for which the property (6.3) is satisfied, and then take α ($\varepsilon > \alpha > 0$) such that, for all $x \in B_\alpha(q)$, $\|u(t,g(x)) - u^0\| = \|[v_1\cdots v_n]g(x)\| < \delta$ (such an α exists since $g(q) = 0$ and $g(\cdot)$ is continuous).

By this construction, the open α-ball neighborhood of q is accessible from p with respect to $B_\rho(p)$. Hence $A^\rho(p) - \{p\}$ is open. Moreover, since this property is satisfied for an arbitrary $\rho > 0$, p is a continuous positive fountain. □

Assume that, for a system $\mathcal{S}: \dot{x} = f(x,u)$ on the state space E, $x_0 \in E$ is a fountain. Let $\Psi : E \to \tilde{E} \triangleq \Psi(E)$ be a diffeomorphism. Consider the system $\tilde{\mathcal{S}}: \dot{z} = \frac{\partial \Psi}{\partial x}(z) f(\Psi^{-1}(z), u)$ on $\tilde{E}$. Since the fountain property is a topological property of the attainable sets, it is preserved under diffeomorphic transformations. Hence the state $\Psi(x_0) \in \tilde{E}$ is also a fountain. In particular, in the case of the existence of a feedback control which, together with a diffeomorphism of the state space, gives rise to a controllable linear system (see [10]), both E and $\tilde{E}$ are fountains, with respect to $\mathcal{S}$ and $\tilde{\mathcal{S}}$, respectively.

6.2. The fountain condition for symmetric systems.

DEFINITION 6.2. *We shall say that a system $\mathcal{S}$ is* symmetric at $x \in E$ *if there exists $\rho > 0$ such that for all $y \in B_\rho(x)$ and $u \in \mathcal{U}$, there exists $u' \in \mathcal{U}$ such that $f(y,u(y)) = -f(y,u'(y))$.*

We note that the definition above is somewhat weaker than the standard definition of a symmetric system (see, for instance, [26]), in which the requirement is that, for all $x \in E$ and $u \in \mathbb{R}^n$, there exists $u' \in \mathbb{R}^n$ such that $f(x,u) = -f(x,u')$.

In [26], a condition for symmetric systems to possess the local controllability property at x is formulated in terms of the Lie algebra of the family $\mathcal{L}_0 \triangleq \{f(\cdot,u);\ u \in \mathbb{R}^n\}$. Namely, for $j = 1, 2, \dots$, let

$$\mathcal{L}_j \triangleq \mathcal{L}_{j-1} \cup \{[f,g](\cdot);\ f(\cdot) \in \mathcal{L}_{j-1}, g(\cdot) \in \mathcal{L}_0 \text{ or } g(\cdot) \in \mathcal{L}_{j-1}, f(\cdot) \in \mathcal{L}_0\}.$$

In addition, define $\mathcal{L}_j(x) \triangleq \{g(x);\ g \in \mathcal{L}_j\}$.

THEOREM 6.3 (see [26]). *Suppose that a system* (2.1) *is symmetric at* x *and the fields* $f(\cdot, u)$, $u \in \mathbb{R}^n$, *are of the class* C^k, $k \geq 2$. *If, for some* $j \leq k$,

$$\dim \mathcal{L}_j(x) = n, \tag{6.4}$$

then the system is locally controllable at x.

THEOREM 6.4. *Suppose that a system* $\mathcal{S}$ *is such that the fields* $f(\cdot, u)$, $u \in \mathcal{U}$, *are of the class* C^k, $k \geq 2$. *Further, let* $x \in E$ *be such that, for some* $\delta > 0$ *and for all states* $y \in A^\delta(x) - \{x\}$,

(i) $\mathcal{S}$ *is symmetric at* y;

(ii) *the condition* (6.4) *is satisfied at* y.

Then x *is a positive fountain.*

Proof. The local controllability at any accessible state $y \in A^\delta(x) - \{x\}$ (which follows from Theorem 6.3) implies that each $y' \in B_\gamma(y)$ (for sufficiently small γ) is accessible from y, and hence from x, with respect to $B_\delta(x)$. Hence the set $A^\delta(x) - \{x\}$ is open, and so x is a positive fountain. ∎

The negative fountain property can be verified by assuming the condition (6.4) and the symmetry property for the reverse time dynamics.

6.3. Full rank condition for systems with time-dependent controls. In this section, the class of admissible control functions, denoted $\mathcal{U}^\omega$, is assumed to be the set of all time-dependent analytic functions taking values in $\mathbb{R}^m$.

Let $f(\cdot, \cdot)$ be an analytic function of the arguments $(x, u) \in E \times \mathbb{R}^m$. Denote the solution trajectory, with the initial condition $x_0 \in E$ under a specified control $u_0(\cdot)$, by $x(\cdot)$. Then $x(\cdot)$ is analytic, and $x(\cdot)$, $u_0(\cdot)$ can be represented locally by the series

$$x(t) = x_0 + \sum_{i=1}^{\infty} x^{(i)}(0)\frac{t^i}{i!}, \qquad u_0(t) = u_0(0) + \sum_{i=1}^{\infty} u_0^{(i)}(0)\frac{t^i}{i!}. \tag{6.5}$$

Denoting $s^i = (x, \lambda_0, \ldots, \lambda_{i-1})$, we define the functions $D_i : \mathbb{R}^n \times \mathbb{R}^{im} \to \mathbb{R}^n$, $i = 1, 2, \cdots$, in the following way:

$$D_1(s^1) = f(x, \lambda_0),$$

$$D_2(s^2) = \frac{\partial D_1}{\partial x}(s^1) f(x, \lambda_0) + \frac{\partial D_1}{\partial \lambda_0}(s^1)\lambda_1$$

$$\cdots$$

$$D_i(s^i) = \frac{\partial D_{i-1}}{\partial x}(s^{i-1}) f(x, \lambda_0) + \frac{\partial D_{i-1}}{\partial \lambda_0}(s^{i-1})\lambda_1 + \cdots + \frac{\partial D_{i-1}}{\partial \lambda_{i-2}}(s^{i-1})\lambda_{i-1} \cdots.$$

THEOREM 6.5. *Let* $x_0 \in E$, *and let* u_0 *be a time-dependent control in* $\mathcal{U}^\omega$. *Suppose that there exist some* $\delta > 0$ *and some integer* $q \geq 0$ *such that, for all* $0 < t < \delta$,

$$\operatorname{rank}\left\{\left[\sum_{i=1}^{\infty} D_i(s_0^i)\frac{t^{i-1}}{(i-1)!} \;\vdots\; \sum_{i=1}^{\infty}\frac{\partial D_i}{\partial \lambda_0}(s_0^i)\frac{t^i}{i!} \;\vdots\; \cdots \;\vdots\; \sum_{i=q+1}^{\infty}\frac{\partial D_i}{\partial \lambda_q}(s_0^i)\frac{t^i}{i!}\right]\right\} = n, \tag{6.6}$$

where we use $u_0 = u_0(0), u_0^{(1)} = \frac{d}{dt}u_0(0), \ldots, u_0^{(q)} = \frac{d^q}{dt^q}u_0(0), \ldots,$ $s_0^i \triangleq (x_0, u_0, u_0^{(1)}, \ldots, u_0^{(i-1)})$, $i \geq 1$, *to simplify the notation. Then* x_0 *is a positive fountain with respect to the trajectory under* u_0 *control.*

Proof. Observe that the functions D_i are defined in such a way that $D_i(s_0^i)$ is equal to the full ith derivative of $x(t)$ with respect to time, evaluated at $t = 0$. Hence the decomposition (6.5) can be rewritten as

$$x(t) = x_0 + \sum_{i=1}^{\infty} D_i(s_0^i)\frac{t^i}{i!}.$$

Next define the function $F : \mathbb{R}^n \times \mathbb{R}^1 \times \mathbb{R}^{(q+1)m} \to \mathbb{R}^n$, $q \geq 0$, as

$$F(x, t, \lambda_0, \lambda_1, \ldots, \lambda_q) \triangleq x - \left\{ x_0 + \sum_{i=1}^{\infty} D_i(p^i)\frac{t^i}{i!} \right\}, \tag{6.7}$$

where $p^i = (\lambda_0, \ldots, \lambda_{i-1}), 1 \leq i \leq q+1$, and $p^i = (\lambda_0, \ldots, \lambda_q, u^{(q+1)}, \ldots, u^{(i-1)}), i > q+1$. Taking an arbitrary $0 < \bar{t} < \delta$ and setting $\bar{x} \triangleq x(\bar{t})$, $p_0 \triangleq (\bar{x}, \bar{t}, u_0, u_0^{(1)}, \ldots, u_0^{(q)})$, we note that

(i) $F(p_0) = \bar{x} - x(\bar{t}) = 0$.

(ii) All first order partial derivatives of F with respect to $x, t, \lambda_0, \ldots, \lambda_q$ are continuous in a neighborhood of the point p_0.

(iii)

$$\left[\frac{\partial F}{\partial t}(p_0) \;\vdots\; \frac{\partial F}{\partial \lambda_0}(p_0) \;\vdots\; \cdots \;\vdots\; \frac{\partial F}{\partial \lambda_q}(p_0) \right]$$

$$= \left[\sum_{i=1}^{\infty} D_i(s_0^i)\frac{\bar{t}^{i-1}}{(i-1)!} \vdots \sum_{i=1}^{\infty} \frac{\partial D_i}{\partial \lambda_0}(s_0^i)\frac{\bar{t}^i}{i!} \vdots \cdots \vdots \sum_{i=q+1}^{\infty} \frac{\partial D_i}{\partial \lambda_q}(s_0^i)\frac{\bar{t}^i}{i!} \right],$$

which, by the condition (6.6), has full rank.

From the implicit function theorem applied to F, it follows that there exist continuous functions $t(r) > 0$ and $\lambda_0(r), \ldots, \lambda_q(r)$ such that

$$F(\bar{x} + r, t(r), \lambda_0(r), \ldots, \lambda_q(r)) = 0$$

for all sufficiently small $r \in \mathbb{R}^n$. Further, define the time-dependent control $v_r \in \mathcal{U}^\omega$ to be

$$\begin{aligned} v_r(t) &= (\lambda_0(r) - u_0) + (\lambda_1(r) - u_0^{(1)})t + \cdots + (\lambda_q(r) - u_0^{(q)})\frac{t^q}{q!} + u_0(t) \\ &= \sum_{i=1}^{q} \lambda_i(r)\frac{t^i}{i!} + \sum_{i=q+1}^{\infty} u_0^{(i)}\frac{t^i}{i!}. \end{aligned} \tag{6.8}$$

Then the solution trajectory under the control v_r drives x_0 to $x(\bar{t}) + r$ at time $t(r)$. Hence each state $x(t)$ (where $t > 0$ is sufficiently small) lies in the interior of the states accessible from x_0, and thus x_0 is a positive fountain with respect to the trajectory under u_0. □

The full rank condition (6.6) can be rewritten (in obvious notation) as

$$\operatorname{rank} \left\{ \sum_{i=0}^{\infty} A_i t^i \right\} = n$$

or, equivalently, as

$$\text{there exists an } n\text{-tuple } k = (k_1, k_2, \dots, k_n),$$
$$1 \le k_i \le n + (q+1)m,\ k_i \ne k_j,\ i \ne j,\ i, j = 1, 2, \dots n, \text{ such that}$$
$$\det\left(\sum_{i=0}^{\infty} P_i^k t^i\right) = \sum_{i=0}^{\infty} p_i^k t^i \not\equiv 0,$$

where $P_i^k \in \mathbb{R}^n \times \mathbb{R}^n$, $i \ge 0$, denotes the submatrix of A_i consisting of the columns $k_1, k_2, \dots, k_n$. We further observe that, for a given k,

$$p_0^k = \det P_0^k, \quad p_\ell^k = \frac{d^{(\ell)}}{dt^\ell}\left\{\det\left(\sum_{i=0}^{\ell} P_i^k t^i\right)\right\}|_{t=0}, \quad \ell \ge 1.$$

Hence a sufficient condition for the fountain property with respect to a specified trajectory to hold can be formulated in terms of the existence of an n-tuple k and an integer $0 \le j < \infty$, such that the coefficient p_j^k (the calculation of which involves only the matrices $D_0, \dots, D_j$) is nonzero.

COROLLARY 6.6. *Let x_0 be an arbitrary state in E. Assume that there exists $\delta > 0$ such that the full rank condition* (6.6) *is satisfied for all $u \in \mathcal{U}^\omega$, for all x in some open neighbourhood $N(x_0)$, and for all t, $t \ne 0$, in the symmetric interval $(-\delta, \delta)$, uniformly in x and u (i.e., δ does not depend on x or u). Then x_0 is a fountain.*

Proof. Consider the set of all states accessible from x_0 with respect to $B_\gamma(x_0)$, where $\gamma > 0$ is chosen so that $B_\gamma(x_0) \subset N(x_0)$. Suppose that there exists a boundary state $y \in A^\gamma(x_0)$. Hence there exists a control $u_0 \in \mathcal{U}^\omega$ and a time instant $0 < T < \infty$ such that $y = \phi(T, x_0, u_0)$. Take $z \triangleq \phi(T - \delta/2, x_0, u_0)$. Then, since $z \in N(x_0)$, any state $y' \in B_\varepsilon(y)$ can be attained from z (and hence from x_0) under the control v_r (defined by (6.8)), $r = y' - y$, in time $t(r)$. Moreover, by the continuity of the functions $t(r), \lambda_0(r), \dots, \lambda_q(r)$, $\varepsilon > 0$ can be chosen so that $\|u_0 - v_r\|$ is sufficiently small and the solution trajectory under v_r does not leave $B_\gamma(x_0)$ on the time interval $[0, t(r)]$. Hence $B_\varepsilon(y) \subset A^\gamma(x_0)$, and x_0 is a positive fountain.

The fact that x_0 is a negative fountain follows from considering the reversed time dynamics. □

Example 6.1. To illustrate the use of Theorem 6.5, we consider the system

$$\dot{x} = 1 + u, \quad \dot{y} = ux^2, \quad u \in \mathbb{R}^1,$$

and test whether the state $p_0 = (1, 0)$ is a fountain with respect to the trajectory under $u_0 \equiv 1$ control. We consecutively compute D_i, $i = 1, 2, 3$:

$$D_1(s^1) = [\lambda_0 + 1\ \ \lambda_0 x^2]^T;\ D_2(s^2) = [0\ \ 2\lambda_0(\lambda_0 + 1)x]^T;\ D_3(s^3) = [0\ \ 2\lambda_0(\lambda_0 + 1)^2]^T;$$

$$\frac{\partial D_1}{\partial \lambda_0} = [1\ \ x^2]^T;\ \frac{\partial D_2}{\partial \lambda_0} = [0\ \ (4\lambda_0 + 2)x]^T.$$

To check the full rank condition, we have to determine the rank of the matrix

$$\left[D_1(s_0^1) + D_2(s_0^2)t + D_3(s_0^3)t^2/2 + o(t) \mid \frac{\partial D_1}{\partial \lambda_0}(s_0^1)t + \frac{\partial D_2}{\partial \lambda_0}(s_0^1)t^2/2 + o(t)\right],$$

where $s_0^1 = (1,1)$, $s^2 = (1,1,0)$, $s^3 = (1,1,0,0)$.

$$\operatorname{rank}\left[\begin{bmatrix}2\\1\end{bmatrix} + \begin{bmatrix}0\\4\end{bmatrix}t + \begin{bmatrix}0\\8\end{bmatrix}t^2/2 \;\middle|\; \begin{bmatrix}1\\1\end{bmatrix}t + \begin{bmatrix}0\\6\end{bmatrix}t^2/2\right] = \operatorname{rank}\begin{bmatrix}2 & t\\ 1+4t+4t^2 & t+3t^2\end{bmatrix} = 2$$

for all sufficiently small $0 < t < \delta$, and hence p_0 is a fountain with respect to the trajectory under the control $u_0 \equiv 1$.

7. Hamiltonian control systems. In this section, we consider an application of the results of section 3 to *affine Hamiltonian control systems* (see [24]), i.e., Hamiltonian control systems with a smooth Hamiltonian of the form

$$H(q,p,u) = H_0(q,p) - \sum_{j=1}^{m} H_j(q,p)u_j,$$

where $H_0(q,p)$ is the *internal Hamiltonian (energy)* and H_j, $j = 1,2,\dots,m$, are the *interaction* or *coupling* Hamiltonians.

For affine Hamiltonian control systems, we have the equations of motion

$$\begin{aligned}\dot{q}_i &= \frac{\partial H_0}{\partial p_i}(q,p) - \sum_{j=1}^{m}\frac{\partial H_j}{\partial p_i}(q,p)u_j,\\ \dot{p}_i &= -\frac{\partial H_0}{\partial q_i}(q,p) + \sum_{j=1}^{m}\frac{\partial H_j}{\partial q_i}(q,p)u_j,\end{aligned} \tag{7.1}$$

where $i = 1,2,\dots,n$.

DEFINITION 7.1. *A u_0 energy slice $ES(H^-,H^+)$ of a Hamiltonian control system is the set of states for which the value of H, under the state-dependent control $u_0 \in \mathcal{U}(\mathbb{R}^n;\mathbb{R}^m)$, lies between some fixed values H^- and H^+,*

$$ES(H^-,H^+) \triangleq \{(p,q);\quad (p,q) \in E \text{ and } H^- < H(p,q,u)_{|_{u=u_0}} < H^+\},$$

where $H(p,q,u)$ is the Hamiltonian function associated with the system (7.1).

We note that, while Definition 7.1 is given in terms of the class $\mathcal{U}(\mathbb{R}^n;\mathbb{R}^m)$, the controllability in the following theorem is with respect to controls in $\mathcal{U} = \mathcal{U}(\mathbb{R};\mathbb{R}^m)$.

THEOREM 7.2. *An affine Hamiltonian control system for which all states are fountains and all equilibrium points under some constant control $u_0 \in \mathbb{R}^m$ are isolated is such that any precompact connected component of a u_0 energy slice is controllable with respect to $\mathcal{U}$.*

Proof. Consider any precompact connected energy slice $ES(H^-,H^+)$, and let $ES_0(H^-,H^+)$ be obtained from $ES(H^-,H^+)$ by removing the (necessarily finite) equilibrium points of the flow under u_0. Since the state space necessarily has dimension greater than 1, we observe that $ES_0(H^-,H^+)$ is also open, precompact, and connected. Now, since $ES_0(H^-,H^+)$ has compact closure, it has finite measure under the density given by H, and, furthermore, the u_0 (i.e., Hamiltonian) flow is Lebesgue measure preserving. It follows by Poincaré's recurrence theorem that almost all states in $ES_0(H^-,H^+)$ are recurrent under the u_0 flow. However, since every state of $ES_0(H^-,H^+)$ is a fountain, and since no state is an equilibrium state under the u_0 flow, Theorem 2.12 implies that $ES_0(H^-,H^+)$ is controllable.

Next consider any equilibrium point $y \in ES(H^-, H^+) - ES_0(H^-, H^+)$ (under u_0). Since all equilibria under u_0 are isolated, there exists a neighborhood $N(y)$ such that $(N(y) - \{y\}) \in ES_0(H^-, H^+)$. Take any $p \in (A^{N(y)}(y) - \{y\})$, which is not empty by the positive fountain condition at y. Then $ES_0(H^-, H^+)$ is a subset $A(p)$ and $p \in A(y)$. Hence $ES_0(H^-, H^+) \subset A(y)$. Finally, consider any equilibrium point $z \in ES(H^-, H^+) - ES_0(H^-, H^+)$ (under u_0). Using analogous arguments and the negative fountain condition at the point z, it can be shown that $ES_0(H^-, H^+) \subset CA(z)$. Since these properties hold for arbitrary equilibria y and z under u_0, it follows that for any $x \in ES(H^-, H^+)$, $ES(H^-, H^+) \subset A(x)$; i.e., $ES(H^-, H^+)$ is controllable. □

We note that the arguments above can be applied to all systems for which there is a measure preserving flow that satisfies the conditions of Theorem 7.2; the theorem is stated for affine Hamiltonian systems because of their importance. We also remark that, in the affine Hamiltonian case, the expression of the linearization condition of section 6.1 for the fountain property takes an interesting symmetric form in terms of the second order partial derivatives of H.

In the case of the existence of a first integral F (which is not necessarily the total energy function) for a dynamical system $\mathcal{S}$, the analogous arguments can be applied to slices based on F or, in the case of the existence of multiple first integrals $F_1, \ldots, F_k$, to intersections of slices based on $F_1, \ldots, F_k$. An example of this is given in [16], where the results of the theory developed in this paper are applied to hybrid systems with disturbances and, in particular, are illustrated on a simplified air traffic system.

8. Applications to hierarchical hybrid control theory (HHCT).

The theory developed in this paper is directly applicable in the analysis of hierarchical hybrid control systems because several of the main results in [5] and [15] employ the HIBC hypothesis (see below). This hypothesis requires that each of the blocks of a given finite analytic partition of the system state space forms a controllable subsystem.

To verify the HIBC condition using the theory of this paper, it is sufficient to establish that the fountain condition holds in E and that one of the recurrence conditions holds for each block of the finite analytic partition under consideration. Furthermore, for the so-called energy slice partitions of affine Hamiltonian systems, the dense recurrence condition under a distinguished constant control is an inherent property which does not need explicit verification whenever each slice is precompact. In this section, we first briefly overview the main concepts of the HHCT developed in [5] and then consider an application of HHCT to a mass-spring system.

8.1. Overview of HHCT.

Consider a system $\mathcal{S}$ of the form (2.1) on some state space $D \in \mathbb{R}^n$. All of the definitions and results here are taken from [5].

DEFINITION 8.1 (see [5]). *A* finite analytic partition *of the state space $D \subset \mathbb{R}^n$ of $\mathcal{S}$ is a pairwise disjoint collection of subsets $\pi = \{X_1, X_2, \ldots, X_{|\pi|}\}$ such that each X_i is nonempty, open, and path-connected and is such that $D = \bigcup_{i=1}^{|\pi|}(X_i \cup \partial X_i)$, where, further, the boundary ∂X_i of every block X_i is a locally finite union of connected components of $n-p$ dimensional, $p \geq 1$, analytic manifolds (possibly with boundary), such that $\partial X_i = \bigcup_{m=1}^{k_i} C_m^i$.*

Henceforth we shall use the following notation for the partition boundary: $\partial\pi = \bigcup_{i=1}^{|\pi|} \partial X_i$.

DEFINITION 8.2 (see [5]). *Let π be a finite analytic partition of D, and let X_i, $1 \le i \le |\pi|$, be any block of π. Then the* interior boundary *∂X_i^{int} of X_i is the set of points $x \in \partial X_i \triangleq \overline{X}_i - \overset{\circ}{X}_i = \overline{X}_i - X_i$ for which there exists a neighborhood N_x meeting only $\overline{X}_i$, and the* exterior boundary *∂X_i^{ext} of X_i is the set complementary to ∂X_i^{int} in ∂X_i; that is, ∂X_i^{ext} is the set of points $x \in \partial X_i$ for which every neighborhood N_x meets $\overline{X}_i^c$.*

For the definition of dynamical consistency, we need the following notion: a state $y \in \partial X_i \cap \partial X_j$ is said to be a *facial (boundary) state of the pair of blocks X_i, X_j* if y lies in the relative interior of the $n-1$ dimensional connected components of the boundaries ∂X_i and ∂X_j. $Facial(\partial X_i \cap \partial X_j)$ shall denote the set of all states that are facial states of the pair X_i, X_j. $NF(\pi)$ shall denote the set of all nonfacial states.

DEFINITION 8.3 (see [5]). *Given a partition π of the set D, $\langle X_i, X_j \rangle \in \pi \times \pi$ is said to be a* dynamically consistent (DC) *pair (with respect to $\mathcal{S}$) if and only if either $i = j$ or, if $i \neq j$, for all x in X_i, there exists $u_x(\cdot) \in \mathcal{U}$, defined upon $[0; T_x]$, $0 < T_x < \infty$, and there exists a facial boundary state $y \in \partial X_i \cap \partial X_j$, such that*

(i) *for all $t \in [0, T_x)$, $\phi(t, x, u_x) \in X_i$, and $\lim_{t \to T_x^-} \phi(t, x, u_x) = y$,*

and, for the state y in (i)*, there exists $u_y \in \mathcal{U}$ defined on $[0, T_y)$, $0 < T_y < \infty$, such that*

(ii) *for all $t \in (0, T_y)$, $\phi(t, y, u_y) \in X_j$,*

where $\phi(\cdot, \cdot, \cdot)$ in (i) *and* (ii) *are the integral curves of the vector field $f(\cdot, \cdot)$ with respect to the controls $u_x, u_y \in \mathcal{U}$ and the initial conditions x, y, respectively.*

DEFINITION 8.4 (see [5]). *Given a partition π of the set D, the* hybrid DC partition machine *$\mathcal{M}^\pi = \langle X^\pi = \{\overline{X}_1, \cdots, \overline{X}_{|\pi|}\}, U = \{\overline{U}_i^j; 1 \le i, j \le |\pi|\}, \Phi^\pi \rangle$, based upon the system $\mathcal{S}$, is the finite state machine defined by $\Phi^\pi(\overline{X}_i, \overline{U}_i^j) = \overline{X}_j$ for all i, j, $1 \le i, j \le |\pi|$, if and only if $\langle X_i, X_j \rangle$ is DC.*

We note that, in the notation introduced in Definition 8.4, if $\langle X_i, X_j \rangle$ is not DC, then the symbol $\overline{U}_i^j$ is not defined.

DEFINITION 8.5 (see [5]). *A hybrid partition machine $\mathcal{M}^\pi$ is called* hybrid in-block controllable *(HIBC) if and only if, for every $X_i \in \pi$ and for all $x, y \in X_i$, the following holds:*

$$\exists u(\cdot) \in \mathcal{U}, \exists T, 0 \le T < \infty, \ (\phi([0; T], x, u) \subset X_i) \wedge (\phi(T, x, u) = y);$$

i.e., each block $X_i \in \pi$ is controllable for the system $\mathcal{S}$.

DEFINITION 8.6. *A hybrid partition machine $\mathcal{M}^\pi$ is called* hybrid between-block controllable *(HBBC) if any block state $\overline{X}_i$ can be reached from any other block state $\overline{X}_j$ by applying a finite number of block transitions.*

We now recall the following result from [5] which characterizes the global controllability of a system $\mathcal{S}$ in terms of the controllability properties of a hierarchical system based upon $\mathcal{S}$. As illustrated in section 8.2, it permits the application of a global controllability analysis to the blocks of a hierarchical system to yield information about the base system and vice versa.

THEOREM 8.7. *An HIBC machine $\mathcal{M}^\pi$ is HBBC if and only if $D^* \triangleq D - \partial\pi$ is controllable for $\mathcal{S}$ with respect to $\tilde{D} \triangleq D - NF(\pi)$, i.e., any two states $x, y \in D^*$ are mutually accessible with respect to $D-$* (nonfacial boundary states).

8.2. Application to a mass-spring system. In this section, we present an illustration of the methodology described in the previous sections for constructing HIBC partitions and the associated partition machines.

Consider N controlled interlinked linear mass-spring systems attached to a moving frame, where the input u determines the velocity of the frame.

Let q_i, $i = 1, \ldots, N$, denote the distance between the ith mass and the frame. Then the Hamiltonian function is given by

$$H(q,p,u) = \frac{1}{2}\sum_{i=1}^{N}\frac{1}{m_i}p_i^2 + \frac{1}{2}\sum_{i=2}^{N}k_i(q_i - q_{i-1})^2 + \frac{1}{2}k_1q_1^2 - \sum_{i=1}^{N}p_iu.$$

The dynamics of the system evolves according to the Hamiltonian system

$$\begin{aligned}\dot{q}_i &= \frac{\partial H}{\partial p_i}(q,p,u) = \frac{1}{m_i}p_i - u,\\ \dot{p}_i &= -\frac{\partial H}{\partial q_i}(q,p,u) + v_i = -k_i(q_i - q_{i-1}) + k_{i+1}(q_{i+1} - q_i) + v_i,\end{aligned} \tag{8.1}$$

where the external forces are interpreted as control variables $v = [v_1 \cdots v_N]^T \in \mathbb{R}^N$ and where we set $q_0 = 0$, $q_{N+1} = 0$, and $k_{N+1} = 0$ to simplify the notation.

We assume that all masses and spring constants of the system are strictly positive. Then the $u^0 = 0, v^0 = 0$ energy slices $ES(H^-, H^+)$ are ellipsoids in the space $\mathbb{R}^{2N}$ (in an appropriate coordinate system $\{\tilde{q}_1, \ldots, \tilde{q}_N, p_1, \ldots, p_n\}$); that is, they have the form

$$\left\{(\tilde{q},p);\ H^- < \frac{1}{2}\sum_{i=1}^{N}\frac{1}{m_i}p_i^2 + \frac{1}{2}\sum_{i=1}^{N}k_i\tilde{q}_i^2 < H^+\right\},$$

where $\{\tilde{q}_1 = q_1, \tilde{q}_i = q_i - q_{i-1}, 2 \le i \le N\}$ is a nonsingular transformation of coordinates.

Let $\{H_1, H_2, \ldots, H_n\}$ be an arbitrary sequence of increasing positive real values. Then π given by the family $\{X_1, X_2, \ldots, X_{n-1}\}$, where X_i is defined to be $ES(H_i, H_{i+1})$, $1 \le i < n$, is a finite analytic partition of the state space $\mathbb{R}^{2N}$.

The fountain property for the system (8.1) can be verified using the technique discussed in section 6.1. The strict positiveness of the masses and spring constants implies that the fountain condition is satisfied for each state $x \in \mathbb{R}^{2N}$. In addition, it can be verified that there exists a unique equilibrium state under the $u^0 = 0$, $v^0 = 0$ control. Consequently, as follows from Theorem 7.2, each block of the partition $\pi \triangleq \{X_1, X_2, \ldots, X_{n-1}\}$ constitutes a controllable set for the linear spring-mass system.

Hence the partition machine $\mathcal{M}^\pi$ based on the energy slice partition π is HIBC. Moreover, by applying the energy slice controllability result of section 7 once more to $E(H)$ defined as the interior of the closure of $\bigcup_{i=1}^{n-1} ES(H_i, H_{i+1})$, i.e., to the interior of the closure of the union of the energy slices, we see that $E(H)$ is controllable.

Further, applying Theorem 8.7, we can conclude that, for the mass-spring system, $\mathcal{M}^\pi$ is HBBC.

Acknowledgments. The authors gratefully acknowledge conversations with Sergej Celikovsky and Robert Hermann concerning parts of this work.

REFERENCES

[1] B. Bonnard, *Controlabilité des systèmes nonlinéaries*, C. R. Acad. Sci. Paris Sér. I Math., 292 (1981), pp. 535–537.

[2] B. Bonnard, *Controlabilité des systèmes mécaniques sur les groupes de Lie*, SIAM J. Control Optim., 22 (1984), pp. 711–722.
[3] R. W. Brockett, *Nonlinear systems and differential geometry*, Proc. IEEE, 64 (1976), pp. 61–72.
[4] P. E. Caines and E. S. Lemch, *On the global controllability of Hamiltonian and other nonlinear systems: Fountains and recurrence*, in Proceedings of the 37th IEEE Conference on Decision and Control, IEEE Control Systems Society, Tampa, FL, 1998, pp. 3575–3580.
[5] P. E. Caines and Y.-J. Wei, *Hierarchical hybrid control systems: A lattice theoretic formulation*, IEEE Trans. Automat. Control, 37 (1998), pp. 501–508.
[6] P. E. Caines and Y.-J. Wei, *The hierarchical lattices of a finite machine*, Systems Control Lett., 25 (1995), pp. 257–263.
[7] P. E. Crouch, *Spacecraft attitude control and stabilisation: Applications of geometric control theory to rigid body models*, IEEE Trans. Automat. Control, 29 (1984), pp. 321–331.
[8] R. Hermann and A. J. Krener, *Nonlinear controllability and observability*, IEEE Trans. Automat. Control, 22 (1977), pp. 728–740.
[9] H. Hermes, *On local and global controllability*, SIAM J. Control, 12 (1974), pp. 252–261.
[10] A. Isidori, *Nonlinear Control Systems*, Springer-Verlag, New York, 1989.
[11] V. Jurdjevic, *Geometric Control Theory*, Cambridge University Press, Cambridge, UK, 1997.
[12] H. Kunita, *Supports of diffusion processes and controllability problems*, in Proceedings of the International Symposium on Stochastic Differential Equations, Kyoto University, Kyoto, Japan, 1976, K. Ito, ed., John Wiley, New York, 1978, pp. 163–185.
[13] H. Kunita, *On the controllability of nonlinear systems with applications to polynomial systems*, Appl. Math. Optim., 5 (1979), pp. 89–99.
[14] E. S. Lemch, *Nonlinear and Hierarchical Hybrid Control Systems*, Ph.D. thesis, Department of Electrical and Computer Engineering, McGill University, Montreal, Canada, 1999.
[15] E. S. Lemch and P. E. Caines, *Partition deformations in hierarchical hybrid control systems*, IMA J. Math. Control Inform., 18 (2001), pp. 531–557.
[16] E. S. Lemch and P. E. Caines, *Hybrid systems with disturbances: Hierarchical control via partition machines*, in Proceedings of the 38th IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 1999, pp. 4909–4915.
[17] E. S. Lemch and P. E. Caines, *Hybrid partition machines with disturbances*, in Proceedings of the AAAI 1999 Spring Symposium Series on Hybrid Systems and AI: Modeling Analysis and Control of Discrete and Continuous Systems, AAAI Technical Report SS-99-05, Stanford University, Stanford, CA, 1999, pp. 109–117.
[18] K.-Y. Lian, L.-S. Wang, and L.-C. Fu, *Controllability of spacecraft systems in a central gravitational field*, IEEE Trans. Automat. Control, 39 (1994), pp. 2426–2441.
[19] C. Lobry, *Controlabilité des systèmes non linéaires*, SIAM J. Control, 8 (1970), pp. 573–605.
[20] C. Lobry, *Controllability of nonlinear systems on compact manifolds*, SIAM J. Control, 12 (1974), pp. 1–4.
[21] K. Lynch, *Controllability of a planar body with unilateral thrusters*, IEEE Trans. Automat. Control, 44 (1999), pp. 1206–1211.
[22] V. Manikonda, *Control and Stabilization of a Class of Nonlinear Systems with Symmetry*, Ph.D. thesis, Department of Electrical Engineering, University of Maryland at College Park, College Park, MD, 1998.
[23] K. Nam and A. Arapostathis, *A sufficient condition for local controllability of nonlinear systems along closed orbits*, IEEE Trans. Automat. Control, 37 (1992), pp. 378–380.
[24] H. Nijmeijer and A. J. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.
[25] E. D. Sontag, *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, 2nd ed., Texts Appl. Math. 6, Springer-Verlag, New York, 1998.
[26] J. Zabczyk, *Mathematical Control Theory: An Introduction*, Birkhäuser Boston, Boston, 1992.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 5, pp. 1554–1566

ESCAPE FUNCTION CONDITIONS FOR THE OBSERVATION, CONTROL, AND STABILIZATION OF THE WAVE EQUATION*

LUC MILLER†

Abstract. For the linear wave equation with time-invariant coefficients on a connected compact Riemannian manifold (Ω, g) with C^3 boundary, the geodesics condition of Bardos, Lebeau, and Rauch [*SIAM J. Control Optim.*, 30 (1992), pp. 1024–1065] is characterized in terms of *escape functions*, which are some Lyapunov functions on the phase space $S^*\bar{\Omega}$ (the unit sphere cotangent bundle). Differentiable escape functions yield a sufficient condition which is slightly less sharp but does not refer to geodesics. The escape function condition yields a straightforward geometric proof that the geodesics condition holds in the situations where first order differential multiplier methods apply. Using microlocal control results, it allows us to generalize some control results (that were obtained by multiplier methods) to variable coefficients and lower order terms. It also allows us to prove, in some class of simple situations (e.g., in $\mathbb{R}^2$ with constant coefficients), that no first order differential multiplier method can reach the optimal control time or control regions.

Key words. wave equation, exact controllability, stabilization, multiplier method, geometric optics, escape function

AMS subject classifications. 35L05, 35L20, 37B25, 49K20, 53D25, 58J47, 93B05, 93B07, 93D20

PII. S036301290139107X

1. Introduction.

This paper is concerned with the widely cited and scarcely used sharp sufficient bicharacteristics condition for the observation, control, and stabilization of the wave equation from the interior or the boundary introduced by Bardos, Lebeau, and Rauch (cf. [2] and the appendix of [21]). Since we shall restrict ourselves to the time-independent coefficients case, this condition can be stated in terms of generalized geodesics (i.e., the rays of geometrical optics) and we shall refer to it as the *geodesics condition*.

While superseding in sharpness and scope earlier results obtained by the "multiplier method" (cf. section 4), the results of [2] are often discarded for reasons already put forward in the introduction of [2]. In the first place, a lot of smoothness is required in [2]. As reviewed in section 1.2, some improvements in this respect have been made thanks to microlocal measures techniques in the last decade. Thus, microlocal techniques apply to the most general geometric situations and have been refined to require less regularity assumptions. The second—more serious—reason is that the explicit computation of the constants appearing in the observation inequalities (which allow us to predict how much energy is needed to control waves of given energy) are out of reach of the closed graph argument in [2] (or the argument by contradiction in the microlocal measures technique). We refer to [23] and [34] for recent significant contributions in this respect using multiplier methods. The third point—to which this article contributes—is that the conditions obtained may not be easy to verify for complicated operators and geometry as acknowledged in [2]. We may add that, even

*Received by the editors June 15, 2001; accepted for publication (in revised form) June 26, 2002; published electronically January 14, 2003.

http://www.siam.org/journals/sicon/41-5/39107.html

†Centre de Mathématiques, U.M.R. 7640 du C.N.R.S., Ecole Polytechnique, 91128 Palaiseau Cedex, France and Équipe Modal'X, Bâtiment G, Université de Paris X - Nanterre, 200, Avenue de la République, 92001 Nanterre Cedex, France (miller@math.polytechnique.fr).

in applications in Euclidean geometry (where geodesics are straight lines) of dimension two or three, we are often in the awkward situation in which it is intuitively clear whether the geodesics condition is satisfied but quite intricate to prove rigorously. Other (sufficient) geometric conditions have been obtained independently by first order multiplier methods in the last decade (e.g., [22], [31], [23], [33]). Thus, multiplier methods apply under the most general regularity assumptions and have been refined to cope with more intricate geometric situations.

This paper bridges the geometric gap between the microlocal results and the multiplier results. In section 3, we introduce a new way of formulating the geodesics condition in terms of *escape functions*. For waves evolving on an n-dimensional space, they can be regarded as functions of $2n$ variables which are Lyapunov functions for the phase space dynamics associated with geometrical optics (technically, on the unit sphere cotangent bundle). The definition of differentiable escape functions, which is enough to formulate a sufficient condition, does not even refer to geodesics (cf. the remarks of section 3). In section 4, we consider the geometric conditions that have been obtained by state-of-the-art multiplier methods and obtain them as direct applications of the microlocal results surveyed in section 1.2 by mere examination of well-chosen escape functions (under stronger smoothness assumptions than in [22], [31], [23] but for variable coefficients). This emphasizes that these geometric conditions provide valuable means to verify the geodesics condition and can be obtained without specific P.D.E. analysis. As a by-product, we extend an exact controllability result due to Yao (in [33] by his innovative Riemann multiplier method) to lower order terms and weaker smoothness assumptions (cf. Remark 4 of section 4).

The interest of our paper in terms of applications is twofold. On the one hand, we give some criteria in section 5 which could save us some time by preventing us from trying to apply first order differential multiplier methods to situations which are out of their scope. On the other hand, the escape function condition is a unified geometric framework, which should help experts at the control of P.D.E. problems make new applications of the microlocal results, or detailed comparisons of their results with the microlocal ones. Since it proves so efficient in the linear case (corresponding to first order multipliers; cf. section 4), we intend it as a track to find sufficient geometric conditions (based on explicit classes of nonlinear escape functions and not referring to geodesics) that would be sharper than those obtained by the multiplier methods and easier to verify—albeit less sharp—than the geodesics condition of [2].

In our time-independent coefficients case, it is well known that the tools of symplectic geometry used in [2] can be formulated in the language of Riemannian geometry (e.g., used in [33]). Since it does not appear explicitly in [2], we have included a concise and thorough presentation of this in section 2 for the reader's convenience. In particular, we emphasize the role of the second fundamental form for gliding geodesics (and give an application in a remark). We hope that this paper will contribute to a better understanding of the geometry involved in the control of P.D.E. problems. For some recent contributions in this direction, we refer to [12] and other papers in the same proceedings.

1.1. P.D.E. problems. Let us recall two simple examples of the P.D.E. problems under consideration.

Our first example is a problem of exact controllability from the boundary. Let Ω be a bounded open connected subset of $\mathbb{R}^n$, with C^1 boundary $\partial\Omega$, inside which waves propagate according to the wave equation $\Box u = 0$, where $\Box = \partial_t^2 - \Delta$ is the speed one d'Alembertian. Let $T > 0$ and $\theta \in C_c^0(]0, T[\times \partial\Omega)$ define the boundary region

$\Sigma = \{(t,x) \in \mathbb{R} \times \partial\Omega |\ \theta(t,x) \neq 0\}$, where the Dirichlet boundary condition is controlled. The function θ is said to *control* Ω *exactly* if for all $(u_0,u_1) \in L^2(\Omega) \times H^{-1}(\Omega)$ and all $(w_0,w_1) \in L^2(\Omega) \times H^{-1}(\Omega)$ there is a control function $v \in L^2(\mathbb{R} \times \partial\Omega)$ such that the solution of the mixed Dirichlet–Cauchy problem,

$$\Box u = 0 \quad \text{in }]0,T[\times \Omega, \qquad u = \theta \times v \quad \text{on }]0,T[\times \partial\Omega, \tag{1.1}$$

with Cauchy data $(u, \partial_t u) = (u_0, u_1)$ at $t = 0$, satisfies $(u, \partial_t u) = (w_0, w_1)$ at $t = T$.

Our second example is a problem of exponential internal stabilization. Let (M,g) be a smooth connected compact Riemannian manifold with boundary ∂M and let $\Box_g$ denote the associated d'Alembertian (cf. section 2 for geometric definitions). Let Ω be the complementary set of the support of a nonnegative $a \in C^\infty(\bar{M})$. The function a defines the damping region $\{x \in \bar{M} |\ a(x) > 0\} = \bar{M} \setminus \bar{\Omega}$ (cf. Figure 1.1). It is said to *stabilize* M *exponentially* if for all $(u_0,u_1) \in H_0^1(M) \times L^2(M)$ the energy of the solution to the mixed Dirichlet–Cauchy problem,

$$\Box_g u + 2a\partial_t u = 0 \quad \text{in } \mathbb{R}_+ \times M, \qquad u = 0 \quad \text{on } \mathbb{R}_+ \times \partial M, \tag{1.2}$$

with Cauchy data $(u, \partial_t u) = (u_0, u_1)$ at $t = 0$, decays exponentially, i.e., there exist $\beta > 0$ and $\beta' \geqslant 1$ such that for all $t \geqslant 0$

$$E(t) = \frac{1}{2} \int_M \left\{ |\partial_t u(t,x)|^2 + |\nabla u(t,x)|_x^2 \right\} d_g x \leqslant \beta' e^{-\beta t} E(0). \tag{1.3}$$

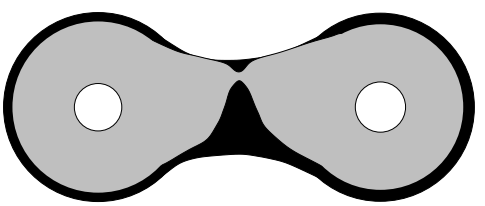


FIG. 1.1. *Ω is light, the damping region $\bar{M} \setminus \bar{\Omega}$ is dark, Γ is the frontier light/dark. The geodesics condition holds.*

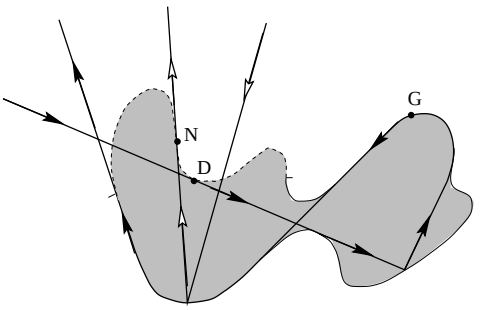


FIG. 1.2. *Two generalized geodesics. Γ is dotted. N: nondiffractive, D: diffractive, G: gliding.*

1.2. Geodesics condition. The common geometric features of these examples are a compact Riemannian manifold (Ω, g) and an open subset Γ of its boundary, if we restrict the first example to a time-independent control region $\Sigma =]0,T[\times \Gamma$, where Γ is an open subset of $\partial\Omega$, and if, in the second example, we denote the part of the boundary of the damping region inside M by $\Gamma = \partial\Omega \cap M$ (cf. Figure 1.1).

In this context, the *generalized geodesics* (cf. Figure 1.2 and Definition 2.1) are continuous trajectories $t \mapsto x(t) \in \bar{\Omega}$ which follow geodesic curves at unit speed in Ω (so that on these intervals $t \mapsto \dot{x}(t)$ is continuous); if they hit $\partial\Omega \setminus \Gamma$ transversely at time t_0, then they reflect as light rays or billiard balls (and $t \mapsto \dot{x}(t)$ is discontinuous at t_0); if they hit Γ transversely at time t_0, then for times $t > t_0$ they have "escaped" from $\bar{\Omega}$; if they hit $\partial\Omega$ tangentially at time t_0, then either there exists a geodesic in Ω which continues $t \mapsto (x(t), \dot{x}(t))$ continuously and they branch onto it, or there is no

such geodesic curve in Ω and—depending on where $\partial\Omega$ was hit—if $x(t_0) \in \Gamma$, then they have "escaped" from $\bar{\Omega}$ at times $t > t_0$; if $x(t_0) \in \Omega \setminus \Gamma$, then they glide at unit speed along the geodesic of $\partial\Omega$, which continues $t \mapsto (x(t), \dot{x}(t))$ continuously until they may branch onto a geodesic in Ω, otherwise they reach $\partial\Gamma$ at a time t_1 and have "escaped" from $\bar{\Omega}$ at times $t > t_1$.

Giving $(x(t_0), \dot{x}(t_0))$ does not always define a unique generalized geodesic (cf. Ex. 24.3.11 in [14] due to M. Taylor). Each of the following assumptions is known to ensure the unique continuation of generalized geodesics (we refer to [4] for (1.6)):

(1.4) g and $\partial\Omega$ are real analytic.

(1.5) g and $\partial\Omega$ are C^∞ and $\partial\Omega$ has no contacts of infinite order with its tangents.

(1.6) g is C^2, $\partial\Omega$ is C^k for some integer $k \geqslant 3$,
and $\partial\Omega$ has no contacts of order $k-1$ with its tangents.

The bicharacteristics condition of Bardos, Lebeau, and Rauch roughly says that every generalized bicharacteristics escapes $\mathbb{R} \times \bar{\Omega}$ through Σ. In our context of time-independent coefficients and region $\Sigma =]0, T[\times \Gamma$, we rephrase the bicharacteristics condition by saying that the time T and the boundary region Γ satisfy the *geodesics condition* if every generalized geodesic starting in $\bar{\Omega}$ has escaped $\bar{\Omega}$ through Γ at time $t = T$; in short, *every generalized geodesic of length T escapes $\bar{\Omega}$ through Γ*. (E.g., in Figure 1.1, this condition is "easily seen" to hold for some T.)

When θ is the characteristic function $\mathbf{1}_\Sigma$ of Σ, it is proved in [2], under (1.4) or (1.5), using microlocal techniques (i.e., the results of [24] and [25] on the propagation of singularities at the boundary and a lifting lemma at nondiffractive points), that the bicharacteristics condition is sufficient and almost necessary for the exact controllability of problem (1.1). Using microlocal measures techniques of Gérard, Tartar, Lions, and Paul (cf. [5] for a survey), Burq obtained the same result in [4] when $\Sigma =]0, T[\times \Gamma$ but $\partial\Omega$ is only C^3 (cf. [11] for propagation results when Ω is convex with $C^{1,1}$ boundary and [6] for results about corners). Moreover, when $\theta \in C_c^0(]0, T[\times \partial\Omega)$, it is proved with the same techniques in [7] under (1.6) that the bicharacteristics condition is both necessary and sufficient for θ to control Ω exactly in problem (1.1), where $\Sigma = \{(t, x) \in \mathbb{R} \times \partial\Omega |\ \theta(t, x) \neq 0\}$. (Note well, if $\mathbf{1}_\Sigma$ controls Ω exactly, then θ does, and if θ does, then so does $\mathbf{1}_{\Sigma'}$ for any open Σ' containing $\bar{\Sigma}$.)

Concerning the stabilization problem in section 1.1 (cf. Figure 1.1), another result of Bardos, Lebeau, and Rauch is that, under (1.4) or (1.5), if there exists a time $T > 0$ such that the geodesics condition holds in Ω with $\Gamma = \partial\Omega \cap M$, then a nonnegative $a \in C^\infty(\bar{M})$ stabilizes M exponentially in problem (1.2) whenever the corresponding damping region $\{x \in \bar{M} |\ a(x) > 0\}$ contains $\bar{M} \setminus \bar{\Omega}$. (Note well, the stabilization of a compact manifold M without boundary corresponds to $\Gamma = \partial\Omega$.) No regularity of $\partial\Omega$ at points of $\bar{\Gamma}$ is required here because there is no boundary condition on Γ. This is the context in which microlocal measures techniques were first applied to control theory, namely by Lebeau in [20], to bound from below the best rate β of exponential decay in (1.3) by some means of a over generalized geodesics.

We refer to [18], [2], [19], [3] for more results allowing observation and stabilization from the boundary, time-dependent coefficients, lower order terms, Neumann and mixed-type boundary conditions, Schrödinger and plate equations, and more. We refer to [27], [26] for results on general boundary conditions and transmission problems and to [8] for general results on systems.

1.3. Escape functions. In this paper, we dwell on pages 1030 and 1031 of [2] which "illustrate the controllability criterion" and end with the following remark: "Our contention is not that we could not do any one of these three [results] with a sufficiently clever differential multiplier. Quite the contrary, the methods of Morawetz, Ralston, and Strauss would surely suffice. However, to create a general result, we would be led inevitably to the same geometric considerations, and avoiding pseudo-differential techniques would only make the task more complicated."

This quote refers to [30], where the decay of solutions of the wave equation outside obstacles with constant coefficients is deduced from some "escape function" which is proved to exist in dimension three under the geodesics condition that the obstacle is "nontrapping." (Melrose later improved on this paper by using the microlocal results of [24] and [25].) Increasingly clever first order differential multipliers had been applied earlier to this problem (radial in [28], gradient of a convex function in [29], "expansive" vector field in [32]) which all correspond to linear escape functions. As Morawetz, Ralston, and Strauss write in [30], "The major point of the present work is that a linear escape function is too special." In the context of resonances for Schrödinger operators with a potential, the same point was made by Helffer and Sjöstrand in [13], where they introduced nonlinear escape functions generalizing the radial case treated earlier by Aguilar and Combes in [1].

Section 5 of this paper makes the same point in the context of the observation, control, and stabilization of waves from the interior or the boundary. In particular, it completes the analysis of [2] by proving that exact controllability cannot be obtained "with a sufficiently clever differential multiplier" of order one in the situation described in Figure 4, p. 1031, of [2]—a disk with some disconnected "minimal" boundary control region which we reproduce in Figure 5.2. This emphasizes the need for new methods to compute explicitly the constants appearing in the observation inequalities in such simple situations.

2. Generalized geodesics. Let (Ω, g) be a connected compact oriented n-dimensional Riemannian manifold with metric g of class C^2 and boundary $\partial\Omega$ of class C^3. Let ν denote the exterior normal vector field and D the Levi–Civita connection of g.

Let $J : X \mapsto \xi$ denote the "flat" isomorphism between the tangent bundle $T\bar{\Omega}$ and cotangent bundle $T^*\bar{\Omega}$ defined by $\xi(Y) = g(X, Y)$, and let a denote the metric on $T^*\bar{\Omega}$ defined by $a(\xi, \eta) = g(X, Y)$, where $\xi = J(X)$ and $\eta = J(Y)$. In local coordinates $(x^1, x^2, \dots, x^n)$, we write a vector field, $X = \sum_i X^i \frac{\partial}{\partial x^i}$, a 1-form, $\xi = \sum_i \xi_i dx^i$, and for each $x \in \bar{\Omega}$ we write the scalar product on the tangent space, $\langle X, Y\rangle_x = \sum_{i,j} g_{i,j}(x) X^i Y^j$, so that $J(X)_j = \sum_i g_{i,j} X^i$ and $J^{-1}(\xi)^j = \sum_i a^{i,j}\xi_i$, where the matrix $A = (a^{i,j})$ is defined by $A^{-1} = (g_{i,j})$. We keep the same notation for the scalar product on the cotangent space $\langle \xi, \eta\rangle_x = {}^t\xi A\eta$ and denote the associated norms by $|\cdot|_x$.

Let $\nabla = J^{-1}d$ denote the gradient operator and Δ_g denote the Laplace–Beltrami operator in (Ω, g). The Sobolev spaces and the energy (1.3) are defined with respect to the measure dx_g. In local coordinates $(x^1, x^2, \dots, x^n)$, $dx_g = \sqrt{\det g}\, dx^1 \cdots dx^n$ and $\Delta_g f = (\sqrt{\det g})^{-1} \sum_{i,j} \partial_{x_j}(a^{i,j}\sqrt{\det g}\,\partial_{x_i} f)$. Let $p \in C^2(T^*(\mathbb{R} \times \bar{\Omega}))$ denote the principal symbol of the d'Alembertian $\Box_g = \partial_t^2 - \Delta_g$. In local coordinates, $p(t, x, \tau, \xi) = \sum_{i,j} a^{i,j}(x)\xi_i\xi_j - \tau^2$ so that p is also the principal symbol of the operator P defined in [2] by $P = \partial_t^2 - \sum_{i,j} a^{i,j}(x)\partial_{x_j}\partial_{x_i} +$ lower order terms.

To link the bicharacteristics of p with the geodesics, it is convenient to consider

the Hamiltonian function $h = p/2$ instead of p. Let H_h denote its Hamiltonian vector field, in local coordinates, $H_h = \partial_{\tau,\xi} h \partial_{t,x} - \partial_{t,x} h \partial_{\tau,\xi}$. The bicharacteristics are integral curves $s \mapsto (t(s), x(s), \tau(s), \xi(s)) = \exp(sH_h)(t(0), x(0), \tau(0), \xi(0))$ of H_h along which $h = 0$. Since p is time-independent, τ is constant along bicharacteristics. By a linear change of parameter, we may restrict ourselves to the bicharacteristics defined on $S = \{(t, x, \tau, \xi) \in h^{-1}(0) | \ -\tau = |\xi|_x = 1\}$. They satisfy $\dot{t}(s) = -\tau(s) = 1$ and $|\dot{x}(s)|_x = |J^{-1}(\xi(s))|_x = |\xi(s)|_x = 1$. Therefore, pushing them through the projections $\pi : T^*(\mathbb{R} \times \bar{\Omega}) \to \bar{\Omega}$ and $\Pi : T^*(\mathbb{R} \times \bar{\Omega}) \to T^*\bar{\Omega}$ (well defined since t is a global coordinate on $\mathbb{R} \times \bar{\Omega}$), we recover the geodesics curves parametrized at unit speed $t \mapsto x(t) = \pi \exp(tH_h)(0, x(0), 0, \xi(0))$ starting from $x(0)$ in the direction $\xi(0)$, which we will sometimes consider as strips $t \to (x(t), J(\dot{x}(t))) = \Pi \exp(tH_h)(0, x(0), 0, \xi(0))$ on the cosphere bundle $S^*\bar{\Omega} = \{(x, \xi) \in T^*\bar{\Omega} | \ |\xi|_x = 1\}$.

Under one of the assumptions (1.4), (1.5), or (1.6), the generalized bicharacteristics introduced in [24] (cf. section 24.3 in [14]) are uniquely defined through each point of $h^{-1}(0)$ (cf. [4] about (1.6)). Let $\mathcal{B}$ denote the second fundamental form, i.e., the symmetric bilinear form $\mathcal{B}_x(X, Y) = -\langle D_X \nu, Y \rangle_x$. Let $\partial\Omega = \phi^{-1}(0)$ be locally defined by a submersion ϕ such that $\Omega = \{\phi(x) > 0\}$. Then $\nabla\phi = -|\nabla\phi|\nu$, $\mathrm{Hess}\phi(X, Y) := D d\phi(X, Y) = -|\nabla\phi| \mathcal{B}_x(X, Y)$, and $H_h^2\phi(x, J(X)) = -|\nabla\phi|_x \mathcal{B}_x(X, X)$. Therefore the *strictly gliding* points of the boundary are the points of $T^*\partial\Omega$ at which the second fundamental quadratic form is positive (cf. Figure 1.2). The Hamiltonian field of h restricted to the symplectic space where $\phi = H_h\phi = 0$ is the gliding vector field $H_h^G = H_h + (H_h^2\phi / H_\phi^2 h) H_\phi$. The gliding bicharacteristics are the trajectories of H_h^G and their image through Π are the geodesic curves of the restriction of g to $\partial\Omega$ parametrized at unit speed.

On a bicharacteristic $t \mapsto (t, x(t), \tau(t), \xi(t))$, the one-sided deleted neighborhoods of the fixed time t_0 are the images of $I \cap \{t > t_0\}$ and the images of $I \cap \{t < t_0\}$ for all neighborhoods I of t_0 in $\mathbb{R}$. Recall that a generalized bicharacteristic pieces together trajectories of H_h in Ω and gliding bicharacteristics in the set $\mathcal{G}_g$ of strictly gliding points: in particular, if at time t it is at $\rho \in T^*\partial\Omega$, then in a one-sided deleted neighborhood of t it coincides with either a bicharacteristic in Ω or a gliding bicharacteristic in $\mathcal{G}_g$. Recall that the *hyperbolic* points of the boundary are the transversal ones (i.e., not in $T^*\partial\Omega$). Recall that $\rho = (x, \xi) \in T^*\bar{\Omega}$ such that $x \in \partial\Omega$ is *nondiffractive* (cf. [2]) if the (nongeneralized) bicharacteristic through ρ at time t is out of $\bar{\Omega}$ at least in one of the one-sided deleted neighborhoods of t (cf. Figure 1.2), i.e., either $\rho \notin T^*\partial\Omega$ is hyperbolic or $\rho \in T^*\partial\Omega$ and the generalized bicharacteristic through ρ at time t is a gliding bicharacteristic in $\mathcal{G}_g$ at least in one of the one-sided deleted neighborhoods of t.

DEFINITION 2.1. *The* generalized geodesic strips *are the images of the generalized bicharacteristics of* $h = (|\xi|_x - \tau^2)/2$ *over* $S = \{-\tau = |\xi|_x = 1\}$ *through the bijection* $\Pi : S \to S^*\bar{\Omega}$. *The* generalized geodesic curves *are the projections of the generalized geodesic strips on* $\bar{\Omega}$.

The generalized geodesic curves are described in section 1.2 for readers not familiar with generalized bicharacteristics. In section 3, it will be convenient to think of generalized geodesics hitting Γ at a nondiffractive point at time $t = t_0$ as having escaped $\bar{\Omega}$ at times $t > t_0$. This interpretation is natural when Ω is an open subset of a larger manifold (M, g): $\partial\Omega \setminus \Gamma$ is a border "obstacle" which confines geodesics inside $\bar{\Omega}$ and Γ is a border "hole" through which the geodesics may "escape" out of $\bar{\Omega}$.

DEFINITION 2.2. *The* geodesics condition $G(T, \Gamma)$ *for the time* $T > 0$ *and the open region* $\Gamma \subset \partial\Omega$ *holds if every generalized geodesic of length greater than* T *passes*

through Γ *at a nondiffractive point (i.e., every generalized geodesic of length greater than* T *escapes* $\bar\Omega$ *through* Γ*).*

As recalled in section 1, $G(T,\Gamma)$ implies, for instance, that all $T'>T$ has the following control property: for all $(u_0,u_1)\in L^2(\Omega)\times H^{-1}(\Omega)$ there exists a control function $v\in L^2(]0,T'[\times\Gamma)$ such that the solution of the mixed Dirichlet–Cauchy problem,

$$\Box_g u=0 \text{ in }]0,T'[\times\Omega,\quad u=v \text{ on }]0,T'[\times\Gamma,\quad u=0 \text{ on }]0,T'[\times\partial\Omega\setminus\Gamma, \tag{2.1}$$

with Cauchy data $(u,\partial_t u)=(u_0,u_1)$ at $t=0$, satisfies $u=\partial_t u=0$ at $t=T'$. Moreover, if T' satisfies this property, then $G(T',\Gamma)$ holds. When $k<\infty$, these results are implicit in [4].

Remark 1. If we also assume Ω to be convex, then the second fundamental form $\mathcal{B}$ is nonnegative on $T^*\partial\Omega$, so that all generalized geodesics starting from $T^*\partial\Omega$ keep gliding on $\partial\Omega$ forever. This answers the question raised in Remark 4.7 of [22]: in a convex open $\Omega\subset\mathbb{R}^n$ satisfying (1.4), (1.5), or (1.6), there is never internal exact controllability for the wave equation with Dirichlet condition on $\partial\Omega$ from a control region G such that $\bar G\subset\Omega$.

3. Escape function condition. In [30], a notion of "escape function" was introduced to characterize the "nontrapping" geodesics condition of Lax and Phillips for exterior problems in $\mathbb{R}^3$ with the Euclidean metric (roughly, $\Gamma=\emptyset$ and there is no specific time, as for stabilization). We adapt the notion of escape function to the geodesics condition of Bardos, Lebeau, and Rauch on a connected compact oriented n-dimensional Riemannian manifold (Ω,g) (with metric g of class C^2 and boundary $\partial\Omega$ of class C^3) by taking T into account.

Definition 3.1. *An* escape function *adapted to* $T>0$ *and* $\Gamma\subset\partial\Omega$ *is a real function* f *defined on* $S^*\bar\Omega$ *such that*

(i) *for all* (x,ξ) *and* (y,η) *in* $S^*\bar\Omega$, $|f(x,\xi)-f(y,\eta)|\leqslant T$;

(ii) f *increases at least as fast as the distance along the closure of any finite interval of geodesic strip in* Ω;

(iii) *for all* $(x,\xi)\in S^*\bar\Omega$ *such that* $x\in\partial\Omega\setminus\Gamma$ *and* $\xi(\nu)>0$, $f(x,\xi')\geqslant f(x,\xi)$ *for* $\xi'=\xi-2\xi(\nu)J^{-1}(\nu)$;

(iv) f *increases at least as fast as the distance along the closure of any finite interval of geodesic strip in* $\partial\Omega\setminus\Gamma$ *on which the second fundamental quadratic form is positive.*

The escape function condition $E(T,\Gamma)$ *holds if there is an escape function adapted to* T *and* Γ.

Note that (i) says that f takes its values in an interval of length less than or equal to T and that (iii) says that f is nondecreasing at reflections on $\partial\Omega\setminus\Gamma$. As such, this definition may appear more intricate than the geodesics condition. If one is only interested in it as a sufficient condition, then it can be simplified by considering only differentiable functions. In the following remarks, the definition of differentiable escape functions is expressed in terms of the values of f and its derivatives and does not refer to geodesics.

Remark 2. When $f\in C^1(T^*\bar\Omega)$, then (ii) says that for all $(x,\xi)\in S^*\Omega$, $H_hf(x,\xi)\geqslant 1$, and (iv) says that for all $(x,\xi)\in S^*\bar\Omega$ such that $x\in\partial\Omega\setminus\Gamma$, $\xi(\nu)=0$, and $\mathcal{B}_x(J^{-1}(\xi),J^{-1}(\xi))>0$, $H_h^Gf(x,\xi)\geqslant 1$. Moreover, (ii) and (iii) imply (iv) by taking a limit as $\xi(\nu)$ tends to 0. Lastly, (iii) is implied by $\nu.\partial_\xi f(x,\xi)\leqslant 0$ for all $x\in\partial\Omega\setminus\Gamma$.

Remark 3. When g is the Euclidean metric, we may consider the unit sphere cotangent bundle as a subspace of $\mathbb{R}^{2n}$, i.e., $S^*\bar{\Omega} = \bar{\Omega} \times \{\xi \in \mathbb{R}^n \mid \sum_{j=1}^n \xi_j^2 = 1\}$. If $f \in C^1(\bar{\Omega} \times \mathbb{R}^n)$, then the four conditions in Definition 3.1 are equivalent to

(i) $f(S^*\bar{\Omega})$ is an interval of length less than or equal to T;

(ii) for all $(x,\xi) \in S^*\bar{\Omega}$, $\xi.\partial_x f(x,\xi) \geqslant 1$;

(iii) for all $(x,\xi) \in S^*\bar{\Omega}$ such that $x \in \partial\Omega \setminus \Gamma$ and $\xi.\nu > 0$, $f(x,\xi') \geqslant f(x,\xi)$ for $\xi' = \xi - 2(\xi.\nu)\nu$.

THEOREM 3.2. *$E(T,\Gamma)$ implies $G(T,\Gamma)$. Under the additional assumption that the generalized geodesics have the uniqueness property (e.g., under one of the assumptions* (1.4), (1.5), *or* (1.6)*), $G(T,\Gamma)$ implies $E(T,\Gamma)$; moreover, f can be chosen continuous outside $\bar{\Gamma}$ and at hyperbolic and strictly gliding points of Γ.*

Proof. We first prove $E(T,\Gamma) \Rightarrow G(T,\Gamma)$. Assume $E(T,\Gamma)$ holds. Let $x : [0,T'] \to \bar{\Omega}$ be a generalized geodesic of length $T' > T$ which does not pass through Γ at a nondiffractive point. From (ii), (iii), and (iv) we deduce that $t \mapsto f(x(t), J(\dot{x}(t)))$ increases at least as fast as t. Hence, $f(x(T'), J(\dot{x}(T'))) - f(x(0), J(\dot{x}(0))) \geqslant T' > T$, which contradicts (i). This proves that such an x does not exist, and therefore $G(T,\Gamma)$ also holds.

We now prove $G(T,\Gamma) \Rightarrow E(T,\Gamma)$. Assume $G(T,\Gamma)$ holds. Let $T' > 0$ and consider a generalized geodesic curve x which does not pass through Γ at a nondiffractive point for $t \in]0,T'[$, such that $x(t) \in \bar{\Gamma}$ for $t \in \{0,T'\}$ and there exists $\epsilon > 0$ satisfying the following properties. If $x(0) \in \Gamma$, then we assume $x(t) \in \Gamma$ for $t \in]-\epsilon,0[$ (in particular, $x(0)$ is nondiffractive) and $x(t) \in \Omega$ for $t \in]0,\epsilon[$. If $x(0) \in \partial\Gamma$, then we assume $x(t) \in \Gamma$ for $t \in]-\epsilon,0[$ and $x(t) \notin \Gamma$ for $t \in]0,\epsilon[$. If $x(T') \in \Gamma$, then we assume $x(t) \in \Gamma$ for $t \in]T',T'+\epsilon[$ (in particular, $x(T')$ is nondiffractive) and $x(t) \in \Omega$ for $t \in]T'-\epsilon,T'[$. If $x(T') \in \partial\Gamma$, then we assume $x(t) \in \Gamma$ for $t \in]T',T'+\epsilon[$ and $x(t) \notin \Gamma$ for $t \in]T'-\epsilon,T'[$.

For each such x, we set $f(x(t), J(\dot{x}(t))) = t$ for $t \in [0,T']$, where the equality is understood as valid for derivatives from both sides at points of reflection except when $t \in \{0,T'\}$. $G(T,\Gamma)$ ensures that $T' < T$ and therefore this f satisfies (i). By definition, this f also satisfies (ii), (iii), and (iv) (with equalities instead of inequalities). The only points of $S^*\bar{\Omega}$ where f has not been yet defined are the points $\rho \in S^*\bar{\Gamma}$ such that the generalized bicharacteristic through ρ at time t is a gliding bicharacteristic in $\mathcal{G}_g$ in both one-sided deleted neighborhoods of t. We set $f = 0$ at those points and recall that strictly gliding points of Γ have this property. The uniqueness property of generalized geodesics ensures that f is well defined.

The continuity of compressed generalized bicharacteristics ensures that f is continuous outside $\bar{\Gamma}$ and at hyperbolic points of Γ. For any point $\rho \in \mathcal{G}_g \cap S^*\Gamma$ and for $\delta > 0$ small enough, there is a neighborhood of ρ in Γ included in the union of $\mathcal{G}_g \cap S^*\Gamma$ and hyperbolic points of Γ which are endpoints of a generalized geodesic x of the preceding type with $T' < \delta$. Therefore f is also continuous at strictly gliding points. □

4. Linear escape functions and the multiplier methods. We discuss the geometrical relationship between the geodesics condition and the situations where first order multiplier techniques apply (cf. the books [21] and [15]). In the framework of escape functions, first order multipliers correspond to escape functions f which are linear with respect to the cotangent variable ξ.

DEFINITION 4.1. *Consider a time $T > 0$ and an open region $\Gamma \subset \partial\Omega$. The* escape vector field condition *$EV(T,\Gamma)$ holds if there is a C^1 section L of $T\bar{\Omega}$ such that*

(i) *for all $x \in \bar{\Omega}$, $|L(x)|_x \leqslant T/2$;*

(ii) *for all* $(x,X)\in S\bar\Omega$, $\langle D_XL,X\rangle_x\geqslant 1$;
(iii) $\{x\in\partial\Omega\,|\,\langle L(x),\nu\rangle_x>0\}\subset\Gamma$.

The escape potential condition $EP(T,\Gamma)$ *holds if there is a function* $\varphi\in C^2(\bar\Omega)$ *such that*

(i) *for all* $x\in\bar\Omega$, $|d\varphi|_x\leqslant T/2$;
(ii) *for all* $(x,X)\in S\bar\Omega$, $\mathrm{Hess}\varphi(X,X):=D d\varphi(X,X)\geqslant 1$;
(iii) $\{x\in\partial\Omega\,|\,\frac{\partial\varphi}{\partial\nu}(x):=d_x\varphi(\nu)>0\}\subset\Gamma$.

When Ω *is a submanifold of* $\mathbb{R}^n$ *with the Euclidean metric, the* radial condition $R(T,\Gamma)$ *holds if there is a point* $x_0\in\mathbb{R}^n$ *such that*

(i) $R(x_0):=\sup\{|x-x_0|\,|\,x\in\bar\Omega\}\leqslant T/2$;
(ii) $\{x\in\partial\Omega\,|\,\langle x-x_0,\nu\rangle>0\}\subset\Gamma$.

Thanks to Theorem 3.2, a straightforward computation is enough to prove that these conditions are sufficient to apply the microlocal results surveyed in section 1.2.

PROPOSITION 4.2. $R(T,\Gamma)\Rightarrow EP(T,\Gamma)\Rightarrow EV(T,\Gamma)\Rightarrow E(T,\Gamma)$.

Proof. To prove $R(T,\Gamma)\Rightarrow EP(T,\Gamma)$, take $\varphi(x)=|x-x_0|^2/2$: $d\varphi(x)=x-x_0$ and $\mathrm{Hess}\varphi$ is the identity matrix. To prove $EP(T,\Gamma)\Rightarrow EV(T,\Gamma)$, take $L(x)=\nabla\varphi$: $\langle L,X\rangle_x=d\varphi(X)$ and $\langle D_XL,X\rangle_x=\mathrm{Hess}\varphi(X,X)$.

Assume $EV(T,\Gamma)$ holds. To prove $E(T,\Gamma)$, take $f(x,\xi)=\xi(L(x))$ and, since this function is C^1, use Remark 2 after Definition 3.1. For all $(x,\xi)\in S^*\bar\Omega$, we have $|f(x,\xi)|\leqslant|\xi|_x|L(x)|_x=|L(x)|_x\leqslant T/2$, so that (i) in $E(T,\Gamma)$ holds. Let $t\mapsto x(t)$ be a geodesic curve, i.e., $D_{\dot x}\dot x=0$. Since D is the Levi–Civita connection of g, $D_X\langle L,X\rangle_x=\langle D_XL,X\rangle_x+\langle L,D_XX\rangle_x$. In particular, $D_{\dot x}f=\langle D_{\dot x}L,\dot x\rangle_x\geqslant 1$, so that (ii) in $E(T,\Gamma)$ holds. By linearity, $f(x,\xi-2\xi(\nu)J^{-1}(\nu))-f(x,\xi)=-2\xi(\nu)f(x,J^{-1}(\nu))=-2\xi(\nu)\langle L(x),\nu\rangle_x<0$ whenever $\xi(\nu)>0$, so that (iii) in $E(T,\Gamma)$ holds. □

The radial multiplier was introduced by Morawetz in [28] for exterior problems and condition $R(T,\Gamma)$ is a variation on her "star-shape" condition. Using this radial multiplier, sufficient conditions for exact controllability from the boundary were obtained by Chen (1979) and Ho (1986) and condition $R(T,\Gamma)$ is their sharper form due to Lions (cf. [21] and [15]). The condition $EP(T,\Gamma)$ is adapted from the convex function condition of Morawetz in [29] for exterior problems with Euclidean metric. There are interesting remarks about global conditions for the existence of potential escape functions in [33]. The condition $EV(T,\Gamma)$ is adapted from the condition of Strauss in [32] for exterior problems with Euclidean metric. The condition of Strauss was used by Chen in [9] for boundary stabilization with Euclidean metric, and later by Lagnese in [16] with less restrictive assumptions and hints for general metrics.

Remark 4. In [33], Yao introduced a "Riemann multiplier method" under the condition $EV(T,\Gamma)$ for smooth $\partial\Omega$ and g. His Theorem 1.1 proves exact controllability under these geometric and regularity assumptions, i.e., for all $(u_0,u_1)\in L^2(\Omega)\times H^{-1}(\Omega)$ there is a control function $v\in L^2(\mathbb{R}\times\Gamma)$ such that the solution of the problem, $\Box_g u=0$ in $]0,T[\times\Omega$, $u=0$ on $]0,T[\times(\partial\Omega\setminus\Gamma)$, $u=v$ on $]0,T[\times\Gamma$, with Cauchy data $(u,\partial_t u)=(u_0,u_1)$ at $t=0$, satisfies $u=\partial_t u=0$ at $t=T$. Under the stronger geometric condition $EP(T,\Gamma)$, the same result has been proved in [17] when $\partial\Omega$ is only C^2 and the operator is $\partial_t^2-\sum_{i,j}a^{i,j}(x)\partial_{x_j}\partial_{x_i}+$ lower order terms, where $a^{i,j}$ are C^1 coefficients and lower order terms have bounded coefficients. On the contrary, it is mentioned on p. 19 of [17] that "in its original form [33], this approach also cannot handle genuine first-order" terms. On the one hand, Theorem 3.2 and Proposition 4.2 prove that Yao's theorem is still true with first order terms, thanks to [2]. On the other hand, Theorem 3.2 and Proposition 4.2 prove that Yao's theorem

is still true when $\partial\Omega$ is of class C^3 and g is of class C^2, thanks to [4]. Moreover, the microlocal measure method of [4] extends to lower order terms (cf. [8]) and should allow us to prove Yao's theorem with lower order terms when $\partial\Omega$ is of class C^3 and g is of class C^2.

Remark 5. If L_2 is a vector field satisfying $\langle D_X L_2, X\rangle_x = 0$ for all $(x,X) \in S\bar{\Omega}$, then adding L_2 to an escape vector field L_1 modifies the boundary condition (iii) without modifying the interior condition (ii). Hence the escape vector field $L = L_1 + L_2$ yields control regions which could not be obtained with L_1. The "multipliers with rotated direction" introduced by Osses (e.g., $L(x) = M(\theta)(x - x_0)$ where $M(\theta)$ is the rotation of angle θ) build on this remark: in [31], Ω is a submanifold of $\mathbb{R}^n$ with the Euclidean metric, $L_1(x) = x - x_0$, and $L_2(x) = A(x - x_0)$, where A is a skew-symmetric matrix.

Remark 6. In [22], Liu introduced a "piecewise multiplier method" for internal exact controllability in a bounded connected open $M \subset \mathbb{R}^n$ with Euclidean metric, with Dirichlet condition on ∂M, from a control region $G \subset M$ under the following geometric condition: there exist open sets $\Omega_j \subset M$ and points $x_j \in \mathbb{R}^n$ (for $j = 1, \dots, J$) such that $\Omega_i \cap \Omega_j = \emptyset$ for $i \neq j$ and $G \supset M \cap \mathcal{N}_\epsilon[(\cup_j \Gamma_j) \cup (M \setminus \cup_j \Omega_j)]$ for some $\epsilon > 0$, where $\mathcal{N}_\epsilon[S]$ denotes an ϵ neighborhood of the set S and $\Gamma_j = \{x \in \partial\Omega_j \mid \langle x - x_j, \nu_j\rangle > 0\}$, where ν_j is the unit exterior normal to $\partial\Omega_j$. Remark 4.7 of [22] calls for a geometric argument proving that this condition implies the geodesics condition when ∂M is sufficiently smooth. Let Ω denote a connected component of $\bar{M}\setminus G$ and $\Gamma = \partial\Omega \cap \partial G \subset \partial G \cap M$. It is included in one of the Ω_j only and we fix this j henceforth. $(T_j, \partial\Omega_j \cap G)$ satisfies the radial condition in Ω_j with $T_j = 2R(x_j)$ since $\Gamma_j \subset G$, hence it also satisfies the geodesics condition. Every generalized geodesic of length T_j starting in $\bar{\Omega}$ and reflecting on $\partial\Omega \setminus \partial G \subset \partial\Omega_j \setminus (\partial\Omega_j \cap G) \subset \partial M$ reaches $\partial\Omega_j \cap G$ and a fortiori escapes $\bar{\Omega}$ through Γ. Therefore every generalized geodesics in $\bar{M}$ of length greater than $\max_j T_j$ reaches G, which proves that the condition of Liu implies the geodesics condition. (No regularity of $\partial\Omega_j$ outside ∂M is needed since it carries no boundary condition.) In the framework of escape functions, the corresponding idea is to consider $f(x,\xi) = (x - \sum_j \mathbf{1}_{\Omega_j}(x) x_j).\xi$ on $S^*(\bar{M} \setminus G)$. Note that the lower bound $\max_j 2R(x_j)$ on the control time could be improved by the microlocal results, e.g., if Ω_1 is star-shaped with respect to x_1, then we can replace $2R(x_1)$ by the length $\mathrm{diam}(\bar{\Omega}_1)$ of the longest segment in $\bar{\Omega}_1$. ($\mathrm{diam}(\bar{\Omega}_1) \leqslant 2R(x_1)$ always holds, and the equality does not hold for any x_1 in Figure 5.1, for example.)

Remark 7. In [23], Martinez introduced the "almost star-shaped" condition for boundary stabilization when g is the Euclidean metric: there exists $\phi \in C^2(\bar{\Omega})$ and $c > 0$ such that $\Delta\phi = 1$ in Ω; $\lambda_1(x) \geqslant c$ in Ω; $\frac{\partial\phi}{\partial\nu}(x) \geqslant 0$ on $\partial\Omega \setminus \Gamma$; $\frac{\partial\phi}{\partial\nu}(x) \leqslant 0$ on Γ, where $\lambda_1(x)$ is the smallest eigenvalue of the matrix $\mathrm{Hess}\phi$ of second order partial derivatives of ϕ. Taking $\varphi = \phi/c$, the second and third requirements are equivalent to the requirements (ii) and (iii) in condition $EP(T,\Gamma)$. This emphasizes that the first requirement $\Delta\phi = 1$ is only useful for explicit computation of the stabilization rate.

5. Linear escape functions are too special. We describe situations where linear escape functions (and therefore all the multiplier methods discussed in section 4) are too special to reach the optimal control time (cf. Proposition 5.1 and Figure 5.1) or the optimal control region (cf. Proposition 5.2 and Figure 5.2). This contrasts with the context of exterior problems where the geodesics condition implies the existence of a linear escape function in dimension $n = 2$ (cf. section 4 of [30]).

Let $\mathrm{diam}_g(\bar{\Omega})$ denote the supremum of the lengths of the geodesics in Ω. A diameter of $\bar{\Omega}$ is a geodesic in Ω whose closure is of length $\mathrm{diam}_g(\bar{\Omega})$. (If $\mathrm{diam}_g(\bar{\Omega})$ is

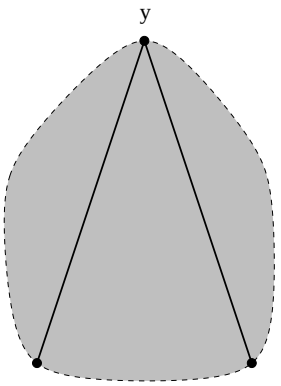


FIG. 5.1. *Segments are diameters of length T. By Proposition* 5.1, *$G(T,\partial\Omega)$ holds but $EV(T,\partial\Omega)$ does not. Here g is Euclidean.*

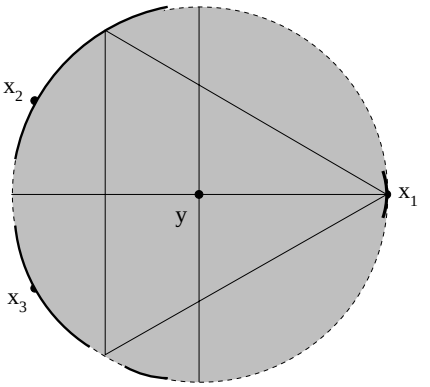


FIG. 5.2. *Segments explain why $G(T,\Gamma)$ holds for some T. $EV(T,\Gamma)$ does not hold for any T by Proposition* 5.2, *with $J=3$ and g Euclidean.*

finite, then there are no closed geodesics and there is at least one diameter of $\bar{\Omega}$ since it is compact.)

PROPOSITION 5.1. *If there exist $y\in\bar{\Omega}$ and two distinct geodesics in Ω of length $T>0$ issued from y, then $EV(T,\partial\Omega)$ does not hold. In particular, if $T=\operatorname{diam}_g(\bar{\Omega})$ and there are two diameters of $\bar{\Omega}$ issued from the same point $y\in\bar{\Omega}$, then $E(T,\partial\Omega)$ holds but $EV(T,\partial\Omega)$ does not.*

Proof. Assume L satisfies $EV(T,\partial\Omega)$. Denote by $[0,T]\ni t\mapsto x(t)$ the geodesic from $x(0)=z$ to $x(T)=y$. By (ii), $\langle L(y),\dot{x}(T)\rangle_y-\langle L(z),\dot{x}(0)\rangle_z\geqslant T$, and by (i), we know that $\langle L(y),\dot{x}(T)\rangle_y\leqslant|L(y)|_y\leqslant T/2$ and $-\langle L(z),\dot{x}(0)\rangle_z\leqslant|L(z)|_z\leqslant T/2$, so that $\langle L(y),\dot{x}(T)\rangle_y=|L(y)|_y=T/2=|L(z)|_z=-\langle L(z),\dot{x}(T)\rangle_z$ and therefore $L(y)=\dot{x}(T)$. The same argument with the other geodesic proves both geodesics have the direction $L(y)$ at y which contradicts that they are distinct. □

PROPOSITION 5.2. *If there are J geodesics in Ω which are issued from points $x_j\in\partial\Omega\setminus\Gamma$ $(j=1,\dots,J)$ with directions normal to the boundary and cross at $y\in\Omega$ with directions ξ_j such that the complementary set in $T_y\Omega$ of their polar cone $C=\{Y\in T_y\Omega\,|\,\forall j,\xi_j(Y)\leqslant 0\}$ is empty, then $EV(T,\Gamma)$ does not hold for any $T>0$.*

Proof. Assume L satisfies $EV(T,\Gamma)$. Denote by $[0,T]\ni t\to x(t)$ the geodesic from $x(0)=x_j$ to $x(T)=y$. Since $\dot{x}(0)=-\nu(x_j)$, (iii) implies $\langle L(x_j),\dot{x}(0)\rangle_{x_j}\geqslant 0$. But by (ii), we know that $\langle L(y),\dot{x}(T)\rangle_y>\langle L(x_j),\dot{x}(0)\rangle_{x_j}$, so that $\langle L(y),\xi_j\rangle_y>0$ since $\dot{x}(T)=\xi_j$. Repeating the argument with all j proves $L(y)\notin C$, which is a contradiction. □

Note that the proofs of these propositions also apply to escape functions of the escape vector field form $f(x,\xi)=\xi(L(x))$ without assuming any regularity on the section L of $T\bar{\Omega}$. Also note that Figure 5.1 is a counterexample to the conjecture (cf. Remark 3.2 in [17]) that the control time in the escape potential condition is optimal.

Remark 8. In the introduction to chapter four of their book *Controllability of Evolution Equations* [10], Fursikov and Imanuvilov mention "a very interesting (and still open to the author's knowledge) question: Does the fulfillment of non-trapping condition imply the existence of pseudoconvex function?". Proposition 5.2 and Figure 5.2 answer this question negatively once it is translated in our terms. On the one hand, the *nontrapping condition* for the control region Γ means that the geodesics

condition $G(T,\Gamma)$ holds for some time $T>0$, or equivalently (cf. Theorem 3.2) that $E(T,\Gamma)$ holds for some time $T>0$. On the other hand, the *existence of a pseudoconvex function* adapted to Γ means that the escape potential condition $EP(T,\Gamma)$ holds for some time $T>0$, and implies (cf. Proposition 4.2) that $EV(T,\Gamma)$ holds for some time $T>0$. Therefore, the situation described in Figure 5.2 (a disk with some disconnected boundary control region as in Figure 4, p. 1031, of [2]) satisfies the nontrapping condition but precludes any pseudoconvex function.

Acknowledgments. I am thankful to N. Burq, G. Lebeau, and C. Margerin for stimulating discussions. N. Burq triggered this investigation by mentioning the function $f(x,\xi)=\xi.(x-x_0)$ as I was asking him whether he knew of a straightforward proof that the radial multiplier condition as found in [21] implies the bicharacteristics condition of [2]. References [23] and [34] were pointed out to me by one of the referees.

REFERENCES

[1] J. Aguilar and J. M. Combes, *A class of analytic perturbations for one-body Schrödinger Hamiltonians*, Comm. Math. Phys., 22 (1971), pp. 269–279.

[2] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[3] N. Burq, *Contrôle de l'équation des plaques en présence d'obstacles strictement convexes*, Mém. Soc. Math. France (N.S.), 55 (1993), 126 pp.

[4] N. Burq, *Contrôlabilité exacte de l'équation des ondes dans des ouverts peu réguliers*, Asymptot. Anal., 14 (1997), pp. 157–191.

[5] N. Burq, *Mesures semi-classiques et mesures de défaut*, Astérisque, 245 (1997), pp. 167–195.

[6] N. Burq, *Contrôle de l'équation des ondes dans des ouverts comportant des coins*, Bull. Soc. Math. France, 126 (1998), pp. 601–637.

[7] N. Burq and P. Gérard, *Condition nécessaire et suffisante pour la contrôlabilité exacte des ondes*, C. R. Acad. Sci. Paris Sér. I Math., 325 (1997), pp. 749–752.

[8] N. Burq and G. Lebeau, *Mesures de défaut de compacité, application au système de Lamé*, Ann. Sci. École Norm. Sup. (4), 34 (2001), pp. 817–870.

[9] G. Chen, *A note on the boundary stabilization of the wave equation*, SIAM J. Control Optim., 19 (1981), pp. 106–113.

[10] A. V. Fursikov and O. Y. Imanuvilov, *Controllability of Evolution Equations*, Lecture Notes Ser. 34, Seoul Nat. Univ., Seoul, 1996.

[11] P. Gérard and E. Leichtnam, *Ergodic properties of eigenfunctions for the Dirichlet problem*, Duke Math. J., 71 (1993), pp. 559–607.

[12] R. Gulliver and W. Littman, *Chord uniqueness and controllability: The view from the boundary.* I, in Differential Geometric Methods in the Control of Partial Differential Equations (Boulder, CO, 1999), Amer. Math. Soc., Providence, RI, 2000, pp. 145–175.

[13] B. Helffer and J. Sjöstrand, *Résonances en limite semi-classique*, Mém. Soc. Math. France (N.S.), 24–25 (1986), iv+228 pp.

[14] L. Hörmander, *The Analysis of Linear Partial Differential Operators*, Vol. III, Springer–Verlag, Berlin, 1985.

[15] V. Komornik, *Exact Controllability and Stabilization: The Multiplier Method*, Masson, Paris, 1994.

[16] J. Lagnese, *Decay of solutions of wave equations in a bounded region with boundary dissipation*, J. Differential Equations, 50 (1983), pp. 163–182.

[17] I. Lasiecka, R. Triggiani, and P.-F. Yao, *Inverse/observability estimates for second-order hyperbolic equations with variable coefficients*, J. Math. Anal. Appl., 235 (1999), pp. 13–57.

[18] G. Lebeau, *Control for Hyperbolic Equations*, in Journées "Équations aux Dérivées Partielles" (Saint-Jean-de-Monts, 1992), École Polytech., Palaiseau, 1992, 24 pp..

[19] G. Lebeau, *Contrôle de l'équation de Schrödinger*, J. Math. Pures Appl. (9), 71 (1992), pp. 267–291.

[20] G. Lebeau, *Équation des ondes amorties*, in Algebraic and Geometric Methods in Mathematical Physics, A. Boutet de Monvel and V. Marchenko, eds., Kluwer Academic, 1996, pp. 73–109.

[21] J.-L. Lions, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués*, Tome 1, Masson, Paris, 1988.
[22] K. Liu, *Locally distributed control and damping for the conservative systems*, SIAM J. Control Optim., 35 (1997), pp. 1574–1590.
[23] P. Martinez, *Boundary stabilization of the wave equation in almost star-shaped domains*, SIAM J. Control Optim., 37 (1999), pp. 673–694.
[24] R. B. Melrose and J. Sjöstrand, *Singularities of boundary value problems.* I, Comm. Pure Appl. Math., 31 (1978), pp. 593–617.
[25] R. B. Melrose and J. Sjöstrand, *Singularities of boundary value problems.* II, Comm. Pure Appl. Math., 35 (1982), pp. 129–168.
[26] L. Miller, *Bicharacteristics Conditions for Exact Controllability from the Boundary in Problems of Transmission*, manuscript.
[27] L. Miller, *Refraction of high-frequency waves density by sharp interfaces and semiclassical measures at the boundary*, J. Math. Pures Appl. (9), 79 (2000), pp. 227–269.
[28] C. S. Morawetz, *The decay of solutions of the exterior initial-boundary value problem for the wave equation*, Comm. Pure Appl. Math., 14 (1961), pp. 561–568.
[29] C. S. Morawetz, *Decay for solutions of the exterior problem for the wave equation*, Comm. Pure Appl. Math., 28 (1975), pp. 229–264.
[30] C. S. Morawetz, J. V. Ralston, and W. A. Strauss, *Decay of solutions of the wave equation outside nontrapping obstacles*, Comm. Pure Appl. Math., 30 (1977), pp. 447–508.
[31] A. Osses, *Une nouvelle famille de multiplicateurs et applications à la contrôlabilité exacte de l'équation d'ondes*, C. R. Acad. Sci. Paris Sér. I Math., 326 (1998), pp. 1099–1104.
[32] W. A. Strauss, *Dispersal of waves vanishing on the boundary of an exterior domain*, Comm. Pure Appl. Math., 28 (1975), pp. 265–278.
[33] P.-F. Yao, *On the observability inequalities for exact controllability of wave equations with variable coefficients*, SIAM J. Control Optim., 37 (1999), pp. 1568–1599.
[34] X. Zhang, *Explicit observability inequalities for the wave equation with lower order terms by means of Carleman inequalities*, SIAM J. Control Optim., 39 (2000), pp. 812–834.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 5, pp. 1567–1585

POSITIVE DEFINITENESS OF FORMS: NUMERICAL IDENTIFICATION*

ARAM V. ARUTYUNOV† AND ALEXEY F. IZMAILOV‡

Abstract. The question about positive definiteness or semidefiniteness of quadratic forms (or, more generally, polynomial homogeneous forms of an even degree) arises in numerous fields of mathematics and its applications. This is certainly the case for optimization theory, including calculus of variations and optimal control. Effective methods intended to obtain a reliable answer to this question for a given form are of doubtless theoretical and practical interest. For that purpose, we propose to use the traditional unconstrained optimization technique, namely, the steepest descent and the conjugate gradient methods. The effectiveness of this approach is justified by theoretical analysis and computational experiments.

Key words. quadratic form, polynomial homogeneous form, positive definiteness, positive semidefiniteness, steepest descent, conjugate gradients, optimal control, calculus of variations, finite-dimensional approximation

AMS subject classifications. 49K15, 49N10, 65K05

PII. S0363012901389925

1. Introduction.

The need to verify positive definiteness or semidefiniteness of (real) polynomial homogeneous forms arises in numerous fields of mathematics and its applications. In particular, this is certainly the case for optimization theory, including calculus of variations and optimal control. For instance, second-order optimality conditions are normally stated in terms of positive semidefiniteness or definiteness of particular quadratic forms. Moreover, sometimes one has to deal with forms of an even degree higher than 2. As will be shown below, for the quadratic form arising from calculus of variations on an infinite time interval, the analysis can be reduced to the corresponding considerations for the particular finite-dimensional form of degree 4. Efficient methods intended to obtain a reliable answer to the question about the "sign" for a given form are of doubtless theoretical and practical interest. Unfortunately, in most cases it is not possible to find the answer by analytical considerations only, and numerical methods have to be involved.

Procedures for verification of second-order optimality conditions can be used as (part of) a stopping test in numerical optimization, but this seems too costly. Probably, main applications of such procedures arise in the context of "postoptimal analysis": they can help when one needs to check whether the candidate for a solution (found by analytic or numerical methods) is an actual solution.

Note that in the optimal control theory, it is often needed to test the definiteness of a certain quadratic form on a subspace rather than on the entire space. However, this problem can be reduced to the case of the entire space using, e.g., Finsler's theorem [3], or by "reducing" the Hessian. In [8, 9, 17], the definiteness of the reduced Hessian for

*Received by the editors May 29, 2001; accepted for publication (in revised form) June 10, 2002; published electronically January 14, 2003.

http://www.siam.org/journals/sicon/41-5/38992.html

†Russian Peoples Friendship University, Miklukho-Maklaya Str. 6, 117198 Moscow, Russia (arutun@orc.ru). The research of this author was supported by Russian Foundation for Basic Research grant 02-01-00334.

‡Moscow State University, Department of Computational Mathematics and Cybernetics, Vorob'yovi Gori, 119899 Moscow, Russia (izmaf@ccas.ru). The research of this author was supported by Russian Foundation for Basic Research grant 01-01-00810.

discretized optimal control problems is analyzed by means of numerical approximation of the eigenvalues.

First let us give a few words about basic assumptions and notation. All matrices in this paper are supposed to have real entries. Let $\mathcal{M}(n, m)$ stand for the set of all $n \times m$ matrices, let $\mathcal{S}(n)$ stand for the set of symmetric $n \times n$ matrices, and let $\mathcal{A}(n)$ stand for the set of antisymmetric $n \times n$ matrices. A real homogeneous form defined on a linear space is referred to as positively semidefinite if it is nonnegative everywhere on this space, and positively definite if it is positive everywhere on this space except the zero point. Recall that finite-dimensional quadratic forms are generated by symmetric matrices, and it is commonly accepted to use the term "positive definiteness" or "semidefiniteness" with respect to such matrices rather than only with respect to forms. In what follows, we use the same symbol for the (polynomial homogeneous) form of degree l and the corresponding symmetric l-linear form.

In what follows (excluding the last section), we basically consider two main objects. The first one is an abstract continuous quadratic form $q : U \to \mathbf{R}$ defined on a Hilbert space U. Let $Q : U \to U$ stand for the (uniquely defined) continuous self-adjoint linear operator such that

$$q(u) = \langle Qu, u\rangle, \quad u \in U. \tag{1.1}$$

We assume that q is at least weakly lower semicontinuous. In particular, special attention will be paid to the case of finite-dimensional U.

Another object is the quadratic form arising from optimal control, and it is much more specific:

$$q(u) = \int_0^1 (\langle A(t)u(t), u(t)\rangle + 2\langle C(t)u(t), x(t)\rangle + \langle B(t)x(t), x(t)\rangle)\, dt, \quad u \in U, \tag{1.2}$$

with $U = L_{2,\,r}[0, 1]$. Here $x(\cdot) \in W^1_{2,\,m}[0, 1]$ is defined by the initial value problem

$$\dot{x} = G(t)x + \Gamma(t)u(t), \quad x(0) = 0. \tag{1.3}$$

The matrix functions $A : [0, 1] \to \mathcal{S}(r)$, $B : [0, 1] \to \mathcal{S}(m)$, $C : [0, 1] \to \mathcal{M}(m, r)$, $G : [0, 1] \to \mathcal{M}(m, m)$, and $\Gamma : [0, 1] \to \mathcal{M}(m, r)$ are supposed to be piecewise continuous. It is well known that this quadratic form is weakly lower semicontinuous on U if and only if the so-called *Legendre condition* is satisfied, i.e., the matrix $A(t)$ is positively semidefinite for almost all $t \in [0, 1]$. The Legendre condition is also necessary (but certainly not sufficient) for q to have a finite index and, in particular, to be positively semidefinite (for the detailed discussion of these questions, see [13, 15]).

For the quadratic form q on the finite-dimensional U, several traditional algebraic approaches to the problem under consideration are well known, such as the Sylvester criterion, numerical methods for computing or approximating the eigenvalues of Q, etc. However, these approaches are too costly, especially when the dimension n is large. Note that the methods for computing the eigenvalues are iterative by nature [21], while the methods we discuss below are finite.

Apparently, the most effective known finite method appropriate for our purpose is the Cholesky (square root) algorithm. For the case of positively definite q, application of this algorithm to Q results in its LL^{T} decomposition, and this requires approximately $n^3/6$ multiplication and the same number of additions [21]. (Here L is a lower triangular matrix with positive diagonal entries.) If q is not positively definite, then

the algorithm will be terminated at some stage because of the necessity to compute the square root from a negative number or to divide by zero.

However, all approaches mentioned above are completely finite-dimensional and, in particular, cannot be applied directly to the quadratic form q given by (1.2), (1.3). Moreover, these approaches assume that the matrix Q is given explicitly, while determination of this matrix can be a rather complicated problem. In particular, this is normally the case for finite-dimensional approximations of q given by (1.2), (1.3) (see section 4). Finally, all these approaches are applicable to quadratic forms only and cannot be applied to forms of an even degree higher than 2.

The alternative approach (free from the disadvantages mentioned above) can be based on optimization methods for the following problem:

$$\begin{array}{ll} \text{minimize} & q(u) \\ \text{subject to} & u \in D, \end{array} \tag{1.4}$$

where $D \subset U$ is a closed convex set such that $0 \in \operatorname{int} D$ or is the boundary of such a set. For instance, one can take the unite ball or the unite sphere in U as D. With this choice, each local solution to problem (1.4) is actually global. Hence, every method which is (globally) convergent to a (local) minimizer will yield the needed result. Note, however, that for forms of degree 4, this is already not the case: such a form can have local minimizers on the unite sphere with positive values of q and, at the same time, can be even not positively semidefinite; see section 5.

This observation is one of the reasons why we propose in this paper to take $D = U$, i.e., to consider the unconstrained optimization problem, q being the objective function. Note that if q is not positively semidefinite, then problem (1.4) with $D = U$ has no (even local) solutions. Nevertheless, as we are going to show, application of the standard unconstrained optimization technique can help to distinguish between cases when q is positively definite, positively semidefinite, or not positively semidefinite, and can do so quite effectively.

Specifically, we consider the steepest descent and the conjugate gradient methods (see, e.g., [18, 20, 12, 5, 7]), though there is a wide range of different unconstrained optimization algorithms applicable in this context. Note that both methods use gradients of q at particular points only, and computation of Q is not necessary for that purpose. Moreover, both methods are correctly defined in the infinite-dimensional setting, at least formally, and both are applicable not only to quadratic forms. Both methods will be described in the next section in the framework of the basic algorithm.

2. The basic algorithm. We introduce our *basic algorithm* for an abstract quadratic form $q : U \to \mathbf{R}$, U being a Hilbert space. The initial *step* 0 of the algorithm consists of fixing two index sets K_1 and K_2 such that $K_1 \cup K_2 = \{0, 1, \ldots\}$, $0 \in K_1$, with an arbitrary initial point $u^0 \in U \setminus \{0\}$.

After k steps of the algorithm, the following elements are supposed to be computed: $u^0, u^1, \ldots, u^k \in U$ and $g^0, g^1, \ldots, g^{k-1} \in U$, and if $k \geq 1$, then

$$q(u^i) > 0, \quad q(g^i) > 0 \quad \forall\, i = 0, 1, \ldots, k-1, \tag{2.1}$$

and

$$\langle q'(u^k), g^{k-1} \rangle = 0. \tag{2.2}$$

Step $k+1$. Compute $q(u^k)$. If $q(u^k) < 0$, then stop. Let $q(u^k) \geq 0$, then compute $q'(u^k) = 2Qu^k$. If $q'(u^k) = 0$, then stop. Let $q'(u^k) \neq 0$. If $q(u^k) = 0$, then stop. Let

$q(u^k) > 0$, then compute

$$g^k = -q'(u^k) + \beta_{k-1} g^{k-1}, \tag{2.3}$$

where

$$\beta_{k-1} = \begin{cases} 0 & \text{if } k \in K_1, \\ \frac{\langle q'(g^{k-1}), q'(u^k)\rangle}{2q(g^{k-1})} & \text{if } k \in K_2. \end{cases} \tag{2.4}$$

(It is known that $\beta_{k-1} > 0$ for $k \in K_2$; see (5.4) below.) Compute $q(g^k)$. If $q(g^k) \le 0$, then stop. Let $q(g^k) > 0$, then take

$$u^{k+1} = u^k + \alpha_k g^k,$$

with $\alpha_k \ge 0$ being defined by the condition

$$q(u^k + \alpha_k g^k) = \min_{\alpha \ge 0} q(u^k + \alpha g^k),$$

and proceed to the next step.

Note that α_k can be given by the explicit formula

$$\alpha_k = -\frac{\langle q'(u^k),\, g^k\rangle}{2q(g^k)} > 0,$$

where the inequality follows from (2.2) and (2.3), as $\langle q'(u^k),\, g^k\rangle = \langle q'(u^k),\, -q'(u^k) + \beta_{k-1} g^{k-1}\rangle = -\|q'(u^k)\|^2$.

Obviously, if $K_2 = \emptyset$, then the underlying iteration of the algorithm described above is the pure steepest descent iteration. We refer to this variant as the *SD-algorithm*. In the opposite case $K_1 = \{0\}$, the algorithm is based on the pure conjugate gradient method. This variant will be referred to as the *CG-algorithm*.

Suppose that the algorithm was stopped at step $k+1$. The following cases are possible. If $q(u^k) < 0$, then q is not positively semidefinite. Let $q(u^k) \ge 0$. If $q'(u^k) = 0$, then with $u^k = 0$ we claim q to be positively definite, and with $u^k \ne 0$ we claim it to be positively semidefinite but not definite. Let $q'(u^k) \ne 0$. If $q(u^k) = 0$, then q is not positively semidefinite. Indeed, in this case $q(u^k + \alpha g^k) = q(g^k)\alpha^2 - \|q'(u^k)\|^2 \alpha < 0$ for each $\alpha > 0$ sufficiently small. Let $q(u^k) > 0$. If $q(g^k) \le 0$, then again q is not positively semidefinite. Indeed, if $q(g^k) = 0$, then $q(u^k + \alpha g^k) = -\|q'(u^k)\|^2 \alpha + q(x^k) < 0$ for each $\alpha > 0$ sufficiently large.

To apply the algorithm to quadratic form q given by (1.2), (1.3), one needs a formula for the gradient of q. This formula is well known [18, 20]; we provide it here for the sake of completeness:

$$q'(u) = 2(A(\cdot)u(\cdot) + (C(\cdot))^{\mathrm{T}} x(\cdot)) - (\Gamma(\cdot))^{\mathrm{T}} \psi(\cdot), \quad u \in U,$$

where $\psi(\cdot) \in W^1_{2,m}[0, 1]$ is defined by the initial value problem

$$\dot\psi = -(G(t))^{\mathrm{T}} \psi + 2(C(t)u(t) + B(t)x(t)), \quad \psi(1) = 0. \tag{2.5}$$

Of course, implementation of the algorithm in this case will require solving the (linear) initial value problems (1.3) and (2.5) and evaluating the integrals (in particular, in (1.2)) on each step. This can hardly be done precisely, except in the most simple cases; hence, some approximation technique will have to be involved.

We proceed with justification of the algorithm. In the rest of this section, we deal mainly with the SD-algorithm, even though some results for the basic algorithm will also be provided. In the next section, the CG-algorithm will be justified in the finite-dimensional setting, and computational results will be reported. Section 4 is devoted to the application of the algorithms to quadratic form defined by (1.2), (1.3) via the direct finite-dimensional approximation. Finally, in section 5, we present the application of the basic algorithm to forms of degree 4.

The following lemma will play a central role in justification of the algorithm.

LEMMA 2.1. *Let $q : U \to \mathbf{R}$ be a continuous quadratic form, and let $Q : U \to U$ be the continuous self-adjoint linear operator associated with q according to* (1.1). *Assume that q is weakly lower semicontinuous and is not positively semidefinite. Then*

(i) *Q has an eigenvalue $\lambda < 0$ with the associated eigenelement $v \in U$;*

(ii) *if the basic algorithm was not stopped on the first k steps, then there exist numbers $\mu_i > 0$ such that*

$$\langle u^i, v\rangle = \left(1 + \sum_{j=1}^{i} \mu_j\right) \langle u^0, v\rangle \quad \forall\, i = 1, \ldots, k.$$

Proof. Consider the optimization problem

$$\begin{array}{ll} \text{minimize} & q(u) \\ \text{subject to} & \|u\|^2 \le 1. \end{array}$$

By weak semicontinuity of q, this problem has a solution v. Obviously, $\|v\| = 1$, and from the Lagrange principle it easily follows that v is an eigenelement of Q associated with some eigenvalue λ. Moreover,

$$0 > q(v) = \langle Qv, v\rangle = \lambda \|v\|^2 = \lambda.$$

This completes the proof of (i).

Assertion (ii) can be proved by induction. For $k = 1$,

$$\langle u^1, \xi\rangle = \langle u^0 - 2\alpha_0 Q u^0, v\rangle = \langle u^0, v\rangle - 2\alpha_0 \lambda \langle u^0, v\rangle = (1 + \mu_1)\langle u^0, v\rangle$$

with $\mu_1 = -2\alpha_0\lambda > 0$.

Suppose that (ii) holds true $\forall\, k \in \{1, \ldots, s\}$ with some s, and let $k = s+1$. Then there exist numbers $\mu_i > 0$ such that

$$\langle u^i, v\rangle = \left(1 + \sum_{j=1}^{i} \mu_j\right) \langle u^0, v\rangle \quad \forall\, i = 1, \ldots, s.$$

Hence

$$\begin{aligned} \langle u^{s+1}, v\rangle &= \langle u^s + \alpha_s g^s, v\rangle \\ &= \left\langle u^s + \alpha_s \left(-2Qu^s + \frac{\beta_{s-1}}{\alpha_{s-1}}(u^s - u^{s-1})\right), v \right\rangle \\ &= \langle u^s, v\rangle - 2\alpha_s\lambda\langle u^s, v\rangle + \frac{\alpha_s}{\alpha_{s-1}}\beta_{s-1}(\langle u^s, v\rangle - \langle u^{s-1}, v\rangle) \\ &= (1 - 2\alpha_s\lambda)\left(1 + \sum_{i=1}^{s} \mu_i\right)\langle u^0, v\rangle + \frac{\alpha_s}{\alpha_{s-1}}\beta_{s-1}\mu_s\langle u^0, v\rangle \\ &= \left(1 + \sum_{i=1}^{s+1} \mu_i\right)\langle u^0, v\rangle, \end{aligned}$$

where

$$\mu_{s+1} = -2\alpha_s\lambda\left(1 + \sum_{i=1}^{s}\mu_i\right) + \frac{\alpha_s}{\alpha_{s-1}}\beta_{s-1}\mu_s > 0.$$

This completes the proof of (ii). □

The following theorem shows that if q is not positively semidefinite, this fact will be identified by the SD-algorithm after a finite number of steps for any choice of the initial point u^0, except the points from a proper closed linear subspace in U.

THEOREM 2.2. *Let $q : U \to \mathbf{R}$ be a continuous quadratic form, and let $Q : U \to U$ be the continuous self-adjoint linear operator associated with q according to* (1.1). *Then*

(i) *for every $u^0 \in U$, either the SD-algorithm will be stopped on some step or $\{q'(u^k)\} = \{2Qu^k\} \to 0$ as $k \to \infty$;*

(ii) *if q is weakly lower semicontinuous and is not positively semidefinite, then the set of initial points $u^0 \in U$—such that either the basic algorithm will be stopped at some point in $\ker Q$ or $\{Qu^k\} \to 0$ as $k \to \infty$—is contained in a proper closed linear subspace in U. Specifically, this subspace is the orthogonal complement to the subspace spanned by the eigenelements of Q associated with the negative eigenvalues.*

Proof. Let us prove (i). Suppose that the algorithm generates the infinite sequence $\{u^k\}$. Then

$$q(u^k) - q(u^{k+1}) \geq \|Q\|^{-1}\|Qu^k\|^2 \quad \forall\, k = 0, 1, \ldots. \tag{2.6}$$

This estimate for the steepest descent trajectories is well known. (One should take into account that $q'(\cdot)$ is Lipschitz-continuous on U with the constant $2\|Q\|$.) On the other hand, this estimate can be easily derived by direct evaluation.

The sequence $\{q(u^k)\}$ is bounded from below, because in the opposite case there would exist some k such that $q(u^k) < 0$, and the algorithm would be stopped. Hence, taking into account the monotonicity of this sequence, it converges, and from (2.6) it follows that $\{Qu^k\} \to 0$ as $k \to \infty$.

To prove (ii), suppose again that the algorithm generates the infinite sequence $\{u^k\}$. Fix an arbitrary eigenelement $v \in U$ of Q associated with a negative eigenvalue λ (which exists, according to assertion (i) of Lemma 2.1). According to assertion (ii) of Lemma 2.1, there exist numbers $\nu_k \geq 0$ such that

$$\langle Qu^k, v\rangle = \lambda\langle u^k, v\rangle = \lambda(1 + \nu_k)\langle u^0, v\rangle \quad \forall\, k = 0, 1, \ldots.$$

Hence, $\{Qu^k\} \to 0$ as $k \to \infty$ takes place if and only if $\langle u^0, v\rangle = 0$. This is certainly also true for the case when $Qx^k = 0$ for some k. This completes the proof. □

When q is positively definite or semidefinite, the behavior of the steepest descent is well known (see, e.g., [18, 20, 12, 5, 7]). First, let U be finite-dimensional. If q is positively definite, then $\forall\, u^0 \in U$ the SD-algorithm will either be stopped on some step at the zero point 0 or will generate the sequence $\{u^k\}$ such that $\{u^k\} \to 0$ as $k \to \infty$, and the rate of convergence is geometric. If q is positively semidefinite but not definite, then $\forall\, u^0 \in U$ the SD-algorithm will either be stopped on some step at the point $\bar{u} = \bar{u}(u^0)$, which is a unique intersection point of $\ker Q$ and $u^0 + (\ker Q)^\perp$, or generate the sequence $\{u^k\}$ converging geometrically to $\bar{u}$. Here we take into account the following observation: $q'(u) = 2Qu \in (\ker Q)^\perp\ \forall\, u \in U$; hence, all the points generated by the SD-algorithm (and the CG-algorithm as well) belong to $u^0 + (\ker Q)^\perp$. Note that $\bar{u}(u^0) = 0$ if and only if $u^0 \in (\ker Q)^\perp$; hence, the set of

"bad" initial points u^0 (such that the SD-algorithm will incorrectly identify the form q as positively definite) is a proper subspace in U.

Hence, in the finite-dimensional setting, the SD-algorithm can be used for our purposes as soon as it is equipped with an additional procedure intended to analyze convergence properties of the sequence being generated. The main problem here is that the number of steps needed to identify the absence of positive semidefiniteness can be arbitrary large, even though it is always finite. This problem will be resolved in the next section in the context of the CG-algorithm.

In the case of infinite-dimensional U, the situation is somewhat more involved, of course. Namely, for our justification above to hold true, one should additionally assume the following property: if q is positively semidefinite, then it is strongly positive on $(\ker Q)^\perp$, i.e., there exists $\gamma > 0$ such that

$$q(u) \geq \gamma \|u\|^2 \quad \forall\, u \in (\ker Q)^\perp.$$

Without this condition, the behavior of the algorithm in the positively semidefinite case can be quite unpredictable.

3. Conjugate gradient method for the finite-dimensional case. In this section, let $U = \mathbf{R}^n$. Let $q : \mathbf{R}^n \to \mathbf{R}$ be an abstract quadratic form, and let $Q \in \mathcal{S}(n)$ be the matrix associated with q. The following simple lemma will be needed.

LEMMA 3.1. *Let $q : \mathbf{R}^n \to \mathbf{R}$ be a quadratic form. If q is nonnegative everywhere on a linear manifold (plane) $V \subset \mathbf{R}^n$, then it is positively semidefinite on* $\operatorname{span} V$.

Proof. For an arbitrary integer s, fix arbitrary points $v^1, \ldots, v^s \in V$ and numbers $\lambda_1, \ldots, \lambda_s$ such that $\mu = \sum_{i=1}^s \lambda_i \neq 0$. Since $\sum_{i=1}^s (\lambda_i/\mu) v^i \in V$ (as V is a linear manifold), for the point $u = \sum_{i=1}^s \lambda_i v^i$ we have

$$q(u) = q\left(\mu \sum_{i=1}^s \frac{\lambda_i}{\mu} v^i\right) = \mu^2 q\left(\sum_{i=1}^s \frac{\lambda_i}{\mu} v^i\right) \geq 0.$$

Obviously, $\operatorname{span} V$ is the closure of the set of such points u. Positive semidefiniteness of q on $\operatorname{span} V$ follows now from continuity of q. ∎

Recall that the index $\operatorname{ind} q$ of a quadratic form q is the maximum dimension of a subspace in $\mathbf{R}^n$ such that q is negatively definite on this subspace. The equivalent definition for $\operatorname{ind} q$ is the minimum codimension of the subspace such that q is positively semidefinite on this subspace.

The following theorem coupled with assertion (ii) of Theorem 2.2 shows that if q is not positively semidefinite, this fact will be identified by the CG-algorithm after not more than $n - \operatorname{ind} q + 1$ steps for almost any choice of the initial point u^0.

THEOREM 3.2. *Let $q : \mathbf{R}^n \to \mathbf{R}$ be a quadratic form. Then for every $u^0 \in U$, the CG-algorithm will be stopped not later than on step $n - \operatorname{ind} q + 1$.*

Proof. Suppose that the CG-algorithm was not stopped on the first k steps, and the elements $u^0, u^1, \ldots, u^k \in \mathbf{R}^n$ and $g^0, g^1, \ldots, g^{k-1} \in \mathbf{R}^n$ were generated, satisfying (2.1). We can consider the case $k \geq 2$ only, because $n - \operatorname{ind} q + 1 \geq 1$, and if the inequality holds as equality, the algorithm would necessarily be stopped on step 1.

By the standard argument for the conjugate gradient method for quadratic functions (see, e.g., [20]), the following can be shown: the vectors $g^0, g^1, \ldots, g^{k-1}$ are linearly independent, and for the linear manifolds $V_i = u^0 + \operatorname{span}\{g^0, g^1, \ldots, g^{i-1}\}$ it holds that

$$q(u^i) = \min_{u \in V_i} q(u) \quad \forall\, i = 1, \ldots, k.$$

In particular, according to (2.1),

$$0 < q(u^{k-1}) = \min_{u \in V_{k-1}} q(u); \tag{3.1}$$

hence

$$q(u) > 0 \quad \forall\, u \in V_{k-1}.$$

By Lemma 3.1, it follows that

$$q(u) \geq 0 \quad \forall\, x \in \operatorname{span} V_{k-1}. \tag{3.2}$$

Suppose for a moment that $u^0 \in \operatorname{span}\{g^0, g^1, \ldots, g^{k-2}\}$. Then V_{k-1} is a linear subspace in $\mathbf{R}^n$, which is in contradiction with (3.1), as $q(0) = 0$. Hence, $u^0 \notin \operatorname{span}\{g^0, g^1, \ldots, g^{k-2}\}$, and $\dim \operatorname{span} V_{k-1} = k$. Thus, according to (3.2), q is positively semidefinite on the subspace of dimension k. Hence, $k \leq n - \operatorname{ind} q$. This completes the proof. □

For example, if $n = 2$, then the absence of positive semidefiniteness will be identified by the algorithm for almost any initial point after, at worst, one step (actually, this will be the steepest descent step). This assertion can also be easily verified directly, using the canonical representation of a quadratic form.

If q is positively definite, then $\forall\, u^0 \in \mathbf{R}^n$ the CG-algorithm will be stopped at 0 not later than on step $n+1$, and positive definiteness will be identified. If q is positively semidefinite but not definite, then $\forall\, u^0 \in U$ the CG-algorithm will be stopped at $\bar{u}$ not later than on step n (recall that $\bar{u} = \bar{u}(u^0)$ is the intersection point of $\ker Q$ and $u^0 + (\ker Q)^\perp$). Similarly to the SD-algorithm, the set of "bad" initial points u^0 (such that the CG-algorithm will incorrectly identify the form q as positively definite) is contained in a proper subspace $(\ker Q)^\perp$ in $\mathbf{R}^n$.

Summarizing, we have proved that for almost any choice of the initial point u^0, the CG-algorithm will completely identify the presence or the absence of positive definiteness or semidefiniteness of a quadratic form after not more than n steps. (If the algorithm was not stopped after n steps, it is not necessary to make step $n + 1$, as by necessity $x^n = 0$, and the form should be claimed positively definite.)

Note that the conjugate gradient step is not much more costly than the steepest descent step, and the CG-algorithm seems to be definitely preferable when compared with the SD-algorithm, at least in the finite-dimensional setting.

Another point is that our justification of the algorithm deals with the idealized situation: it is assumed that all the computations are precise. The question about the influence of computational errors and perturbations of another kind is beyond the scope of this paper, even though this question is certainly important, especially for those forms which are degenerate or nearly degenerate. Here, we mention only that in our computational experiments, any failure evidently caused by inexactness of computations never occurred.

We complete this section with a brief report on computational experiments which were carried out in order to confirm the theoretically justified properties of the algorithm and to compare the two variants of the algorithm with each other and with the Cholesky method (for the latter, the standard realization in Maple V Release 5 was used).

Dimensions in the experiments being reported were up to $n = 100$. The following scheme was used to generate the matrix $Q \in \mathcal{S}(n)$: $Q = S^{\mathrm{T}} \hat{Q} S$, where $\hat{Q}$ is a fixed diagonal $n \times n$ matrix, and S is a nondegenerate $n \times n$ matrix. The effectiveness of

our algorithms can depend drastically on the choice of the initial point, which is why a comparison of our algorithms with the Cholesky method has restricted reliability, of course.

In our experiments, various $\hat{Q}$ were taken, though special attention was paid to matrices with a small number of negative diagonal elements. The algorithms were multiply started with random initial points, and the average characteristics were computed. General conclusions are as follows. For any dimension, both the SD-algorithm and the CG-algorithm are quite competitive in comparison with the Cholesky method, and the CG-algorithm normally requires substantially less CPU time than the Cholesky method. This can be explained by the observation that in large dimensions, the number of steps required by the CG-algorithm is normally substantially less than one guaranteed by the estimate in Theorem 3.2. Moreover, the number of steps needed for the SD-algorithm is usually also substantially less than n. This compensates the necessity of two costly matrix-vector multiplications on every step of the algorithm.

As an example, let us report the typical results for one of the series of experiments. One diagonal entry of $\hat{Q}$ was taken equal to -1, and the others were taken as random numbers in $(0, 10)$. Clearly, this choice of $\hat{Q}$ is quite unfavorable for identification of the absence of positive semidefiniteness. Next, S was taken as the random matrix with entries in $(-1, 1)$. These two objects where used to compute Q as suggested above, and the algorithms were multiply started with random initial points with components in $(-10, 10)$. For $n = 100$, the minimum number of steps for the SD-algorithm is 17, the maximum number is 93, and the average number is 46. For the CG-algorithm, the corresponding numbers are 8, 24, and 14.

4. Finite-dimensional approximation for the optimal control quadratic form. In this section, we deal with the quadratic form $q : U \to \mathbf{R}$ given by (1.2), (1.3), $U = L_{2,r}[0, 1]$. Recall that the matrix functions $A(\cdot)$, $B(\cdot)$, $C(\cdot)$, $G(\cdot)$, and $\Gamma(\cdot)$ in (1.2), (1.3) are supposed to be piecewise continuous.

Consider the simplest finite-dimensional approximation of q. Namely, for every integer n, take a mesh $\{t_0^n, t_1^n, \ldots, t_n^n\}$ on $[0, 1]$ such that $0 = t_0^n < t_1^n < \ldots t_n^n = 1$ and the following hypotheses are satisfied:

(H1) All points of discontinuity of $A(\cdot)$, $B(\cdot)$, $C(\cdot)$, $G(\cdot)$, and $\Gamma(\cdot)$ belong to the mesh for n sufficiently large. (We emphasize that this requirement is the only serious reason why we consider the mesh which is not uniform, in general.)

(H2) There exists a constant $c > 0$ such that $\Delta_n \le c/n \ \forall\, n$, where

$$\Delta_n = \max_{i=0,\,1,\,\ldots,\,n-1} \Delta_i^n, \quad \Delta_i^n = t_{i+1}^n - t_i^n, \quad i = 0, 1, \ldots, n-1.$$

Take $U_n = \prod_{i=0}^{n-1} \mathbf{R}^r$, $X_n = \prod_{i=0}^{n} \mathbf{R}^m$, $q_n : U_n \to \mathbf{R}$,

$$q_n(u^n) = \sum_{i=0}^{n-1} \Delta_i^n (\langle A_i^n u_i^n, u_i^n \rangle + 2\langle C_i^n u_i^n, x_i^n \rangle + \langle B_i^n x_i^n, x_i^n \rangle), \quad u^n = (u_0^n, u_1^n, \ldots, u_{n-1}^n) \in U_n, \tag{4.1}$$

where $x^n = (x_0^n, x_1^n, \ldots, x_n^n) \in X_n$ is defined by the discrete initial value problem given by the explicit Euler scheme for (1.3):

$$x_{i+1}^n = x_i^n + \Delta_i^n (G_i^n x_i^n + \Gamma_i^n u_i^n), \quad i = 0, 1, \ldots, n-1, \quad x_0^n = 0. \tag{4.2}$$

Here $A_i^n = A(t_i^n + 0)$, $i = 0, 1, \ldots, n-1$, and the same notation is used for matrix functions $B(\cdot)$, $C(\cdot)$, $G(\cdot)$, and $\Gamma(\cdot)$.

THEOREM 4.1. *Under the assumptions imposed, the following hold:*

(i) *If q is not positively semidefinite, then q_n is not positively semidefinite for every n sufficiently large.*

(ii) *If q is strongly positive, i.e., there exists $\gamma > 0$ such that*

$$q(u) \geq \gamma \|u\|^2 \quad \forall\, u \in U, \tag{4.3}$$

then for any $\varepsilon > 0$,

$$q_n(u^n) \geq (\gamma - \varepsilon)\|u^n\|^2 \quad \forall\, u^n \in U_n$$

for every n sufficiently large.

As is well known, condition (4.3) is equivalent to saying that q is positively definite, and the *strengthened Legendre condition* is satisfied, i.e., there exists $\tilde{\gamma} > 0$ such that the matrix $A(t) - \tilde{\gamma}E$ is positively semidefinite for almost all $t \in [0, 1]$.

Results similar to Theorem 4.1 can be found in the literature (see, e.g., [11, 16]). In particular, Lemma 11 in [11] contains our assertion (ii), assuming the continuity of the matrix functions involved. However, we include the proof for the sake of completeness.

Proof. In order to prove (i), let us fix $u \in U$ such that $q(u) < 0$. Since continuous functions comprise a dense subset in $L_{2,\,r}[0, 1]$ (see [19]) and q is continuous, we can suppose $u(\cdot)$ to be continuous on $[0, 1]$.

Next, let $x(\cdot) \in W^1_{2,\,m}[0, 1]$ be defined by (1.3) for the given $u(\cdot)$. For each n, take $u^n = (u(t_0^n), u(t_1^n), \ldots, u(t_{n-1}^n)) \in U_n$ and define $x^n \in X_n$ according to (4.2). Clearly, it suffices to show that

$$q_n(u^n) \to q(u) \quad \text{as } n \to \infty.$$

According to (1.2) and (4.1), $\forall\, n$

$$\begin{aligned}
q_n(u^n) - q(u) = \sum_{i=0}^{n-1} \int_{t_i^n}^{t_{i+1}^n} \big(& \langle A_i^n u_i^n,\, u_i^n \rangle - \langle A(t)u(t),\, u(t) \rangle \\
& + 2(\langle C_i^n u_i^n,\, x_i^n \rangle - \langle C(t)u(t),\, x(t) \rangle) \\
& + \langle B_i^n x_i^n,\, x_i^n \rangle - \langle B(t)x(t),\, x(t) \rangle \big)\, dt \\
= \sum_{i=0}^{n-1} \int_{t_i^n}^{t_{i+1}^n} \big(& \langle A_i^n u(t_i^n),\, u(t_i^n) \rangle - \langle A(t)u(t),\, u(t) \rangle \\
& + 2(\langle C_i^n u(t_i^n),\, x(t_i^n) \rangle - \langle C(t)u(t),\, x(t) \rangle) \\
& + \langle B_i^n x(t_i^n),\, x(t_i^n) \rangle - \langle B(t)x(t),\, x(t) \rangle \big)\, dt \\
+ \sum_{i=0}^{n-1} \int_{t_i^n}^{t_{i+1}^n} \big(& 2\langle C_i^n u(t_i^n),\, x_i^n - x(t_i^n) \rangle \\
& + \langle B_i^n x_i^n,\, x_i^n \rangle - \langle B_i^n x(t_i^n),\, x(t_i^n) \rangle \big)\, dt.
\end{aligned}$$

From continuity of $u(\cdot)$ and $x(\cdot)$, piecewise continuity of $A(\cdot)$, $B(\cdot)$, $C(\cdot)$, the definition of A_i^n, B_i^n, C_i^n, $i = 0, 1, \ldots, n$, and the hypotheses (H1) and (H2), it follows that

the first sum on the right-hand side of the last equality tends to zero as $n \to \infty$. To prove that the second sum also tends to zero, it suffices to show that

$$\max_{i=0,\,1,\,\ldots,\,n} |x_i^n - x(t_i^n)| \to 0 \quad \text{as } n \to \infty.$$

But under our assumptions, this is the standard convergence result for the Euler scheme (see, e.g., [11] or [20, p. 358]). This completes the proof of (i).

We proceed with the proof of (ii). Suppose that for every n sufficiently large, there exists $u^n \in U_n \setminus \{0\}$ such that $q_n(u^n) \le 0$. We may suppose that u^n is normalized in the following sense:

$$\|u^n\| = \sqrt{\sum_{i=0}^{n-1} \Delta_i^n |u_i^n|^2} = 1.$$

Define the piecewise constant function $u_n(\cdot) : [0,\,1] \to \mathbf{R}^r$, $u_n(t) = u_i^n$, $t \in [t_i^n,\,t_{i+1}^n)$, $i = 0,\,1,\,\ldots,\,n-1$. Note that

$$\Delta_i^n |u_i^n| = \sqrt{\Delta_i^n}\sqrt{\Delta_i^n |u_i^n|^2} \le \sqrt{\Delta_i^n}\|u^n\| = \sqrt{\Delta_n} \quad \forall\, i = 0,\,1,\,\ldots,\,n-1, \tag{4.4}$$

$$\sum_{i=0}^{n-1} \Delta_i^n |u_i^n| = \|u_n\|_{L_1} \le \|u_n\|_{L_2} = \|u^n\| = 1. \tag{4.5}$$

In particular, from (4.3) it follows that $q(u_n) \ge \gamma$. Now it is clear that in order to come to a contradiction, it suffices to show that

$$q_n(u^n) - q(u_n) \to 0 \quad \text{as } n \to \infty.$$

For every n sufficiently large, let $x^n \in X_n$ be defined according to (4.2), and let $x_n(\cdot) \in W_{2,\,m}^1[0,\,1]$ be defined by (1.3) for $u(\cdot) = u_n(\cdot)$. Let $\bar G = \sup_{t\in[0,\,1]} \|G(t)\|$, $\bar\Gamma = \sup_{t\in[0,\,1]} \|\Gamma(t)\|$. Now, from the Gronwall inequality (in its discrete and continuous forms; see [20]), hypothesis (H2), and (4.5), the following estimates can be derived:

$$\max_{i=0,\,1,\,\ldots,\,n} |x_i^n| \le \bar\Gamma e^{c\bar G}, \quad \|x_n\|_C \le \bar\Gamma e^{\bar G}. \tag{4.6}$$

According to (1.2) and (4.1), for every n sufficiently large

$$\begin{aligned}
q_n(u^n) - q(u_n) = \sum_{i=0}^{n-1} \int_{t_i^n}^{t_{i+1}^n} &\big(\langle A_i^n u_i^n,\, u_i^n\rangle - \langle A(t)u_n(t),\, u_n(t)\rangle \\
&+ 2(\langle C_i^n u_i^n,\, x_i^n\rangle - \langle C(t)u_n(t),\, x_n(t)\rangle) \\
&+ \langle B_i^n x_i^n,\, x_i^n\rangle - \langle B(t)x_n(t),\, x_n(t)\rangle\big)\, dt \\
= \sum_{i=0}^{n-1} \int_{t_i^n}^{t_{i+1}^n} &\big(\langle (A_i^n - A(t))u_i^n,\, u_i^n\rangle \\
&+ 2\langle (C_i^n - C(t))u_i^n,\, x_i^n\rangle + \langle (B_i^n - B(t))x_i^n,\, x_i^n\rangle\big)\, dt \\
+ \sum_{i=0}^{n-1} \int_{t_i^n}^{t_{i+1}^n} &\big(\langle C(t)u_i^n,\, x_i^n - x_n(t)\rangle \\
&+ \langle B(t)(x_i^n - x_n(t)),\, x_i^n\rangle + \langle B(t)x_n(t),\, x_i^n - x_n(t)\rangle\big) dt.
\end{aligned}$$

The first sum on the right-hand side of the last equality tends to zero as $n \to \infty$. This follows from piecewise continuity of $A(\cdot)$, $B(\cdot)$, $C(\cdot)$, the definition of A_i^n, B_i^n, C_i^n, $i = 0, 1, \ldots, n$, hypotheses (H1) and (H2), and (4.5), (4.6). To prove that the second sum also tends to zero, it suffices to show that

$$\max_{i=0,\,1,\,\ldots,\,n-1} \max_{t\in[t_i^n,\,t_{i+1}^n]} |x_i^n - x_n(t)| \to 0 \quad \text{as } n \to \infty.$$

For every n sufficiently large, and for arbitrary $t \in [t_i^n, t_{i+1}^n)$, $i = 0, 1, \ldots, n-1$, set $\varphi_n(t) = |x_i^n - x_n(t)|$. Then, according to (1.3) and (4.2),

$$\begin{aligned}
\varphi_n(t) &\le \sum_{j=0}^{i-1} \int_{t_j^n}^{t_{j+1}^n} (|G_j^n x_j^n - G(\tau)x_n(\tau)| + |\Gamma_j^n u_j^n - \Gamma(\tau)u_j^n|)\,d\tau \\
&\quad + \int_{t_i^n}^{t} (|G(\tau)x_n(\tau)| + |\Gamma(\tau)u_i^n|)\,d\tau \\
&\le \sum_{j=0}^{i-1} \int_{t_j^n}^{t_{j+1}^n} (\|G_j^n - G(\tau)\||x_j^n| + \|\Gamma_j^n - \Gamma(\tau)\||u_j^n|)\,d\tau \\
&\quad + \sum_{j=0}^{i-1} \int_{t_j^n}^{t_{j+1}^n} \bar{G}\varphi_n(\tau)\,d\tau + \int_{t_i^n}^{t} \bar{G}\varphi_n(\tau)\,d\tau \\
&\quad + \int_{t_i^n}^{t} (\bar{G}|x_i^n| + \bar{\Gamma}|u_i^n|)\,d\tau \\
&\le \sum_{j=0}^{i-1} |x_j^n| \int_{t_j^n}^{t_{j+1}^n} \|G_j^n - G(\tau)\|\,d\tau + \sum_{j=0}^{i-1} |u_j^n| \int_{t_j^n}^{t_{j+1}^n} \|\Gamma_j^n - \Gamma(\tau)\|\,d\tau \\
&\quad + \Delta_i^n(\bar{G}|x_i^n| + \bar{\Gamma}|u_i^n|) + \int_0^t \bar{G}\varphi_n(\tau)\,d\tau \\
&= \omega_n + \int_0^t \bar{G}\varphi_n(\tau)\,d\tau,
\end{aligned}$$

where $\omega_n \to 0$ as $n \to \infty$. This follows from piecewise continuity of $G(\cdot)$, $\Gamma(\cdot)$, the definition of G_i^n, Γ_i^n, $i = 0, 1, \ldots, n$, hypotheses (H1) and (H2), and (4.4)–(4.6). By the Gronwall inequality,

$$\varphi_n(t) \le \omega_n e^{\bar{G}} \to 0 \quad \text{as } n \to \infty.$$

This completes the proof of (ii). □

By Theorem 4.1, the question about the "sign" of q is reduced to the same question for the finite-dimensional form q_n for n sufficiently large. More precisely, by analyzing the "sign" of q_n, one can distinguish between the case when q is strongly positive and the case when it is not positively semidefinite.

For finite-dimensional q_n, the SD-algorithm and the CG-algorithm are applicable. For that purpose, a formula for the gradient of q_n is needed:

$$(q_n'(u^n))_i = 2\Delta_i^n(A_i^n u_i^n + (C_i^n)^{\mathrm{T}} x_i^n) - \Delta_i^n(\Gamma_i^n)^{\mathrm{T}}\psi_{i+1}^n, \quad i = 0, 1, \ldots, n-1, \quad u^n \in U_n,$$

where $\psi^n \in X_n$ is defined by the discrete initial value problem

$$\psi_{i-1}^n = \psi_i^n + \Delta_{i-1}^n (G_{i-1}^n)^{\mathrm{T}}\psi_i^n - 2\Delta_i^n(C_{i-1}^n u_{i-1}^n + B_{i-1}^n x_{i-1}^n), \;\; i = 1, \ldots, n-1, \;\; \psi_n^n = 0$$

(see [18, 20]). Again we emphasize that we do not need the symmetric matrix associated with q_n and that it would by quite costly to compute this matrix, especially for large n.

5. Forms of degree 4. Consider the following quadratic form, arising in various applications, particularly in the study of degenerate quadratic forms in calculus of variations [22, 1]: $q : W \to \mathbf{R}$,

$$q(x) = \int_0^{+\infty} (\langle A\dot{x}(t),\, \dot{x}(t)\rangle + 2\langle C\dot{x}(t),\, x(t)\rangle + \langle Bx(t),\, x(t)\rangle)\, dt, \quad x \in W,$$

where W is the space comprised by absolutely continuous functions $x(\cdot) : [0, +\infty) \to \mathbf{R}^m$ such that $x(\cdot)$ and $\dot{x}(\cdot)$ belong to $L_{2,\,m}[0,\, 1]$, $A,\, B \in \mathcal{S}(m)$, $C \in \mathcal{A}(m)$, and A is assumed to be positively semidefinite (Legendre condition). It follows from the results in [22, 1, 10] that q is positively definite (semidefinite) if and only if the form $\tilde{q} : U \to \mathbf{R}$ is positively definite (semidefinite), where $U = \mathbf{R}^m \times \mathbf{R}^m$,

$$\begin{aligned} \tilde{q}(u) &= \tilde{q}_{A,\,B,\,C}(u) \\ &= (\langle A\xi,\, \xi\rangle + \langle A\eta,\, \eta\rangle)(\langle B\xi,\, \xi\rangle + \langle B\eta,\, \eta\rangle) \\ &\quad -4\langle C\xi,\, \eta\rangle^2, \quad u = (\xi,\, \eta) \in U. \end{aligned} \tag{5.1}$$

Thus, the question about the "sign" for the infinite-dimensional quadratic form q is reduced to the same question for the form $\tilde{q}$ of degree 4 on the space U of dimension $2m$.

Clearly, for $\tilde{q}$ to be positively definite, the matrices A and B should be positively definite. For $m = 2$, the simple criterion for positive definiteness of $\tilde{q}$ is known.

PROPOSITION 5.1. *If $m = 2$ and matrices A and B are positively definite, then the form $\tilde{q}$ defined by* (5.1) *is positively semidefinite if and only if*

$$4 \det C \le \det A \left(\operatorname{tr} \sqrt{A^{-1/2} B A^{-1/2}}\right)^2, \tag{5.2}$$

and positively definite if and only if (5.2) *holds as a strict inequality.*

Inequality (5.2) can be rewritten in the form

$$4 \det C \le \det B \operatorname{tr}(B^{-1}A) + 2\sqrt{\det A \det B}.$$

Note that the right-hand side of the last inequality is actually symmetric with respect to A and B, as for 2×2 matrices the following holds: $\det B \operatorname{tr}(B^{-1}A) = \det A \operatorname{tr}(A^{-1}B)$.

The proof of Proposition 5.1 is quite long and involved; see [2]. Moreover, to our knowledge, for $m > 2$ no analytical criterion is available. That is why it would be desirable to develop a numerical technique appropriate for identification of positive definiteness of $\tilde{q}$.

From now on, let $U = \mathbf{R}^n$, and let $q : \mathbf{R}^n \to \mathbf{R}$ be a (polynomial homogeneous) form of degree 4. The basic algorithm introduced in section 2 can be formally extended to this case with the following natural modifications.

First, $q'(\cdot)$ is now defined by the relation

$$\langle q'(u),\, v\rangle = 4q(u,\, u,\, u,\, v), \quad u,\, v \in \mathbf{R}^n.$$

Note that normally it is not that difficult to obtain explicit formulas for the gradient and higher differentials of a given form.

Next, if the algorithm will be stopped on step $k+1$ because of the equality $q'(u^k) = 0$, one should be more careful with conclusions (the same is true for the case when $\{q'(u^k)\} \to 0$ as $k \to \infty$). For example, the form $q : \mathbf{R}^2 \to \mathbf{R}$, $q(u) = u_1 u_2 (u_1 - u_2)^2$ is not positively semidefinite, though at the same time all the points on the line $u_1 = u_2$ are zeros and local minimizers for q. If u^k belongs to this line, then $q'(u^k) = 0$, and the algorithm will be stopped with the wrong conclusion that q is positively semidefinite. Moreover, it is easy to see that in this example, the set of initial points such that the algorithm will be stopped on the mentioned line after one step is quite wide. The need to consider such situations can be avoided by the assumption that q is *nondegenerate*, namely,

$$q'(u) \neq 0 \quad \forall\, u \in \mathbf{R}^n \setminus \{0\}. \tag{5.3}$$

Note that nondegeneracy is a generic condition in the corresponding space of forms; this follows from the parametric transversality theorem [14]. On the other hand, if one assumes this condition to be satisfied, then the cases of semidefiniteness, but not definiteness, are immediately excluded from consideration. Under the nondegeneracy condition, equality $q'(u^k) = 0$ means that $u^k = 0$ (respectively, relation $\{q'(u^k)\} \to 0$ as $k \to \infty$ means that $\{u^k\} \to 0$ as $k \to \infty$), and q is declared positively definite in this case.

If $k \in K_2$, then one of the following formulas can be used for β_{k-1} instead of (2.4):

$$\beta_{k-1} = \frac{\langle q'(u^k),\, q'(u^k) - q'(u^{k-1})\rangle}{|q'(u^{k-1})|^2},$$

$$\beta_{k-1} = \frac{|q'(u^k)|^2}{|q'(u^{k-1})|^2}, \tag{5.4}$$

or

$$\beta_{k-1} = \frac{\langle q''(u^k) g^{k-1},\, q'(u^k)\rangle}{\langle q''(u^k) g^{k-1},\, g^{k-1}\rangle}.$$

(In the quadratic case, all these representations of β_{k-1} are actually the same and are equivalent to (2.4).) Furthermore, in the literature on the conjugate gradient method, it is usually recommended to take an infinite K_1 in the nonquadratic case. The most customary choice is $K_1 = \{0,\, n,\, 2n,\, \ldots\}$, i.e., the method is "renewed" after every n steps, and we accept this choice in what follows.

If $q(u^k) > 0$, $q(g^k) = 0$, the algorithm should not be stopped automatically on step $k+1$. One should compute

$$c_2 = \frac{1}{3!}(q'''(u^k))(g^k,\, g^k,\, g^k) = 4q(u^k,\, g^k,\, g^k,\, g^k).$$

If $c_2 \neq 0$, then q is not positively semidefinite, because $q(x^k + \alpha g^k)$ is a polynomial of degree 3 with respect to α, and it always takes negative values. Let $c_2 = 0$, then compute

$$c_1 = \frac{1}{2!}\langle q''(u^k) g^k,\, g^k\rangle = 6q(u^k,\, u^k,\, g^k,\, g^k).$$

If $c_1 \le 0$, then q is evidently not positively semidefinite. On the other hand, if $c_1 > 0$, then the algorithm should not be stopped. Note that in this case,

$$\alpha_k = \frac{c_0}{2c_1},$$

where

$$c_0 = -\langle q'(u^k),\, g^k\rangle = |q'(u^k)|^2.$$

Finally, if $q(u^k) > 0$, $q(g^k) > 0$, then α_k should be computed as the nonnegative minimizer to the polynomial

$$q(u^k + \alpha g^k) = q(g^k)\alpha^4 + c_2\alpha^3 + c_1\alpha^2 + c_0\alpha + q(u^k)$$

of degree 4.

Of course, the algorithm can be applied to forms of an even degree higher than 4, but in that case, in order to compute the step-length parameter, one would have to deal with the polynomial equation of degree 5 or higher.

A form q of degree 4 satisfying the nondegeneracy condition (5.3) has a unique critical point at zero. This point is a global minimizer, provided q is positively definite, and is not even a local minimizer, provided q is not positively definite. Because of this observation, one can expect the reasonable behavior of the algorithm.

Note that the gradient of a nontrivial form of degree 4 cannot be Lipschitzian on the entire space. This fact results in additional difficulties in theoretical analysis of the algorithm for such forms; for instance, results from [6] cannot be applied here.

THEOREM 5.2. *Let $q : \mathbf{R}^n \to \mathbf{R}$ be a nondegenerate form of degree 4.*

Then for every $u^0 \in \mathbf{R}^n$, either the SD-algorithm will be stopped on some step or $\{u^k\} \to 0$ as $k \to \infty$.

Proof. Suppose that the SD-algorithm generates an infinite sequence $\{u^k\}$. If this sequence is bounded, then from the results in [4] it follows that $\{q'(u^k)\} \to 0$ as $k \to \infty$. Taking into account nondegeneracy of q, $\{u^k\}$ converges to zero.

Thus, we have only to prove that $\{u^k\}$ is bounded. Suppose that this is not the case, i.e., there exists a subsequence $\{u^{k_i}\}$ such that $|u^{k_i}| \to \infty$ as $i \to \infty$.

From (5.3) it follows that

$$|q'(u)| \ge \gamma |u|^3 \quad \forall\, u \in \mathbf{R}^n$$

with some $\gamma > 0$. It is easy to see now that there exist constants C_1, C_2, $C_3 > 0$ such that $\forall\, \alpha \ge 0$

$$\begin{aligned} q(u^{k_i} - \alpha q'(u^{k_i})) &\le q(u^{k_i}) - |q'(u^{k_i})|^2\alpha + C_1|u^{k_i}|^8\alpha^2 \\ &\quad + C_2|u^{k_i}|^{10}\alpha^3 + C_3|u^{k_i}|^{12}\alpha^4 \\ &\le |u^{k_i}|^4 \left(\frac{q(u^{k_i})}{|u^{k_i}|^4} - \gamma^2\alpha|u^{k_i}|^2 + C_1\alpha^2|u^{k_i}|^4 \right. \\ &\quad \left. + C_2\alpha^3|u^{k_i}|^6 + C_3\alpha^4|u^{k_i}|^8 \right). \end{aligned} \tag{5.5}$$

Choose $t > 0$ satisfying the inequality $-\gamma^2 + C_1 t + C_2 t^2 + C_3 t^3 < 0$, and set $\alpha = t|u^{k_i}|^{-2}$; then

$$q(u^{k_i} - \alpha q'(u^{k_i})) \le |u^{k_i}|^4 \left(\frac{q(u^{k_i})}{|u^{k_i}|^4} - \gamma^2 t + C_1 t^2 + C_2 t^3 + C_3 t^4 \right) < 0 \tag{5.6}$$

for i large enough. Here we take into account that $|u^{k_i}|^{-4}q(u^{k_i}) \to 0$ as $i \to \infty$, because $\{q(u^k)\}$ is a bounded sequence (as it is monotonically decreasing and composed of positive numbers).

Inequality (5.6) means that the algorithm would be stopped on step $k_i + 1$, which contradicts the assumption that $\{u^k\}$ is infinite. This completes the proof. □

Unfortunately, for the CG-algorithm, a complete analogue of Theorem 5.2 was not proved by the authors. It can be shown that for a nondegenerate form, if the CG-algorithm generates an infinite sequence, then the latter can have a limit point only at zero. This follows from results in [4], and it is important here that the algorithm is "renewed" on steps kn, $k = 0, 1, \ldots$. Moreover, the argument used in the proof of Theorem 5.2 actually shows that $\{u^{kn}\} \to 0$ as $k \to \infty$, and hence 0 is indeed a limit point for $\{u^k\}$. At the same time, we cannot prove that the entire sequence $\{u^k\}$ is bounded, even though in our computational experiments, the case of unbounded $\{u^k\}$ has never occurred.

Generally speaking, the advantages of the conjugate gradient method in the nonquadratic case are normally theoretically justified only locally, near the solution being sought. Reasonable global behavior (which is of primary interest in the context of this paper) is usually ensured precisely by the "renewing" gradient steps; see [4]. That is why theoretical questions are discussed below mainly for the SD-algorithm, even though, by the evidence of computational experiments, the CG-algorithm is preferable here, as well as in the quadratic case.

Unfortunately, however, even for the SD-algorithm, in the nonquadratic case one cannot guarantee the correct identification with almost any choice of the initial point. If q is positively definite, then for every initial point, the SD-algorithm will either be stopped at zero on some step or generate the sequence convergent to zero. Note that the rate of convergence is normally rather low, as zero is a degenerate critical point for q. However, if q is not positively semidefinite, computational experiments show that there can exist "wide" sets (of nonzero measure) such that starting from them, the SD-algorithm will generate a sequence convergent to zero, and the form will be incorrectly identified as positively definite. We proceed with the example for which this "bad" behavior can be detected.

First, consider the form $q : \mathbf{R}^2 \to \mathbf{R}$, $q(u) = (au_1^2 + u_2^2)(u_1^2 - u_2^2)$, with $0 < a < 1/3$. This form is not positively semidefinite. Note that for problem (1.4) with D equal to the unit sphere in $\mathbf{R}^2$, the points $\pm(1, 0)$ are local solutions, and the value of q at these points is positive (recall that this is not possible for quadratic forms). To construct the example we need, let us now increase the dimension by 1, and take $q : \mathbf{R}^3 \to \mathbf{R}$, $q(u) = (au_1^2 + u_2^2 + bu_3^2)(u_1^2 - u_2^2 + u_3^2)$, with $b > 0$ being large enough. Geometrical analysis of level surfaces for this form, combined with numerical experiments, shows that the point (1, 0, 0), e.g., belongs to the closure of an open set of "bad" initial points.

According to the discussion above, the estimates for sets of appropriate initial points can be of interest. The next proposition contains the estimates for the sets of initial points such that the SD-algorithm will give a correct answer not later than steps 2 and 3, respectively. Set

$$\theta = \max_{t \geq 0} t(\gamma^2 - C_1 t - C_2 t^2 - C_3 t^3) > 0,$$

where γ, C_1, C_2, and C_3 are taken from the proof of Theorem 5.2.

PROPOSITION 5.3. *Assume that a nondegenerate form $q : \mathbf{R}^n \to \mathbf{R}$ of degree 4 is not positively semidefinite.*

Then for every $u^0 \in \mathbf{R}^n \setminus \{0\}$ *such that*

$$q(u^0/|u^0|) + 32\theta\varphi^2(u^0)(1 - 8\varphi^2(u^0)) < 2\theta \tag{5.7}$$

with

$$\varphi(u^0) = \frac{q(u^0/|u^0|)}{|q'(u^0/|u^0|)|},$$

the SD-algorithm will be stopped at some point distinct from zero not later than step 3. *In particular, if*

$$q(u^0/|u^0|) < \theta, \tag{5.8}$$

then the algorithm will be stopped not later than step 2.

It is easy to see that the second term on the left-hand side of (5.7) is less than or equal to θ, and hence (5.8) implies (5.7).

Proof. Suppose that the SD-algorithm was not stopped on the first $k+1$ steps, and the following points were generated: $u^0, u^1, \dots, u^k \in \mathbf{R}^n$ (distinct from zero) and $u^{k+1} \in \mathbf{R}^n$. Replace u^{k_i} in (5.5) by x^i and set $t = \alpha|u^{k_i}|^2$; then

$$\begin{aligned} q(u^{i+1}) &= \min_{\alpha \ge 0} q(u^i - \alpha q'(u^i)) \\ &\le q(u^i) - |u^i|^4 \max_{t \ge 0}(\gamma^2 t - C_1 t^2 - C_2 t^3 - C_3 t^4) \\ &= q(u^i) - \theta|u^i|^4 \\ &\le \cdots \\ &\le q(u^0) - \theta \sum_{j=0}^{i} |u^j|^4 \quad \forall\, i = 0, 1, \dots, k. \end{aligned}$$

Hence, provided

$$q(u^0) - \theta \sum_{i=0}^{k} |u^i|^4 < 0, \tag{5.9}$$

on step $k+2$ the algorithm will necessarily be stopped at a point distinct from zero. In particular, under the condition (5.8), (5.9) holds with $k = 0$; i.e., the algorithm will be stopped on step 2, unless it was stopped on step 1.

Furthermore,

$$\begin{aligned} |u^{i+1}|^2 &= |u^i|^2 - 2\langle q'(u^i), u^i\rangle\alpha_i + |q'(u^i)|^2\alpha_i^2 \\ &= |u^i|^2 - 8q(u^i)\alpha_i + |q'(u^i)|^2\alpha_i^2 \\ &\ge |u^i|^2 - \max_{\alpha \ge 0} \alpha(8q(u^i) - |q'(u^i)|^2\alpha) \\ &= |u^i|^2 - 16\left(\frac{q(u^i)}{|q'(u^i)|}\right)^2 \quad \forall\, i = 0, 1, \dots, k. \end{aligned}$$

It follows that

$$\begin{aligned} q(u^0) - \theta(|u^0|^4 + |u^1|^4) &\le q(u^0) - \theta\left(2|u^0|^4 - 32\left(\frac{q(u^0)}{|q'(u^0)|}\right)^2 |u^0|^2 \right. \\ &\qquad \left. + 16^2\left(\frac{q(u^0)}{|q'(u^0)|}\right)^4\right) \\ &= q(u^0) + 32\theta\varphi^2(u^0)(1 - 8\varphi^2(u^0))|u^0|^4 - 2\theta|u^0|^4, \end{aligned}$$

and hence (5.7) implies (5.9) with $k = 1$; i.e., the algorithm will be stopped on step 3, unless it was stopped earlier. This completes the proof. □

In several series of computational experiments, the forms of the following type were considered:

$$q(u) = \langle Q_1 u,\, u\rangle \langle Q_2 u,\, u\rangle, \quad u \in \mathbf{R}^n,$$

where matrices $Q_1,\, Q_2 \in \mathcal{S}(n)$ were generated in the same manner as Q in the experiments with quadratic forms. For instance, let $\hat{Q}_1$ and $\hat{Q}_2$ be diagonal $n \times n$ matrices. Let the diagonal entries of the first matrix be random numbers in (0, 10), and for the second matrix, let it be the same, except for one diagonal entry equal to -1. Fix random $n \times n$ matrices S_1 and S_2 with entries in $(-1,\, 1)$, and then set $Q_1 = S_1^{\mathrm{T}} \hat{Q}_1 S_1$, $Q_2 = S_2^{\mathrm{T}} \hat{Q}_2 S_2$. Further, the algorithms were multiply started with random initial points with components in $(-1,\, 1)$. For $n = 10$, computations were terminated if the number of steps exceeded 50. (It was considered that in this case, the algorithm failed to identify the absence of positive semidefiniteness.) Typical results are as follows. The SD-algorithm failed for 36% of trials, and when it did not fail, identification required 8 steps on average. For the CG-algorithm, the last number is 11, though the algorithm failed for 8% of trials only. Thus, the CG-algorithm turns out to be more robust with respect to the choice of the initial point, though the required number of steps for the CG-algorithm can even exceed the corresponding number for the SD-algorithm.

Recall the form $\tilde{q} = \tilde{q}_{A,\,B,\,C} : U \to \mathbf{R}$ defined in (5.1). Let $\mathcal{P}(m) \subset \mathcal{S}(m)$ stand for the cone of positively definite matrices, equipped with the metrics induced from $\mathcal{S}(m)$.

PROPOSITION 5.4. *The set of triples* $(A,\, B,\, C) \in \mathcal{P}(m) \times \mathcal{P}(m) \times \mathcal{A}(m)$ *such that the form* $\tilde{q}_{A,\,B,\,C}$ *of degree* 4 *defined in* (5.1) *is nondegenerate is open and dense in* $\mathcal{P}(m) \times \mathcal{P}(m) \times \mathcal{A}(m)$.

The proof easily follows from the parametric transversality theorem [14].

In the computational experiments, matrices A, B, and C were generated, e.g., in the following way. Fix diagonal $m \times m$ matrices $\hat{A}$ and $\hat{B}$, the diagonal entries of both being random numbers in (0, 10), except for the first two diagonal entries. Let the first two diagonal entries of $\hat{A}$ be equal to 1, and let the corresponding diagonal entries b_1 and b_2 of $\hat{B}$ be considered as parameters. Further, define a matrix $\hat{C} \in \mathcal{A}(m)$ such that the upper left 2×2 submatrix of $\hat{C}$ is

$$\begin{pmatrix} 0 & c \\ -c & 0 \end{pmatrix},$$

c being a parameter. Fix random $m \times m$ matrix S with elements in $(-1,\, 1)$, and set $A = S^{\mathrm{T}} \hat{A} S$, $B = S^{\mathrm{T}} \hat{B} S$, $C = S^{\mathrm{T}} \hat{C} S$. Choosing parameters b_1, b_2, and c in accordance with criteria from Proposition 5.1, one can obtain the form with the known answer to the question about the "sign" for any m.

The algorithms were multiply started with random initial points with components in $(-1,\, 1)$. For $n = 10$, computations were terminated if the number of steps exceeded 100. Let us report the typical results for $b_1 = 1$, $b_2 = 4$, $c = 2$ (the form is not positively semidefinite with this choice). The SD-algorithm failed for 35% of trials, and when it did not fail, identification required 28 steps on average. The CG-algorithm never failed, and the required number of steps was 16 on average. Note that for the same b_1 and b_2, and $c = 1.6$, the SD-algorithm failed for 87% of trials, though the CG-algorithm failed for 3% of trials only. (The form is still not positively semidefinite with this choice, but c is close to the "critical" value 1.5.)

Acknowledgments. The authors thank the anonymous referees, whose valuable comments significantly improved the presentation of the paper.

REFERENCES

[1] A. V. Arutyunov, *Optimality conditions: Abnormal and degenerate problems*, Kluwer Academic, Dordrecht, Boston, London, 2000.
[2] A. V. Arutyunov and A. F. Izmailov, *On identification of semidefiniteness of forms*, Comput. Math. and Math. Phys., 42 (2002), pp. 767–780.
[3] R. Bellman, *Introduction to Matrix Analysis*, McGraw–Hill, New York, Toronto, London, 1960.
[4] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, London, 1982.
[5] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1995.
[6] D. P. Bertsekas and J. N. Tsitsiklis, *Gradient convergence in gradient methods with errors*, SIAM J. Optim., 10 (2000), pp. 627–642.
[7] J. F. Bonnans, J. Ch. Gilbert, C. Lemaréchal, and C. Sagastizábal, *Optimisation numérique: Aspects théoriques et pratiques*, Springer-Verlag, Berlin, 1997.
[8] C. Büskens and H. Maurer, *SQP-methods for solving optimal control problems with control and state constraints: Adjoint variables, sensitivity analysis and real-time control*, J. Comput. Appl. Math., 120 (2000), pp. 85–108.
[9] C. Büskens and H. Maurer, *Nonlinear programming methods for real-time control of an industrial robot*, J. Optim. Theory Appl., 107 (2000), pp. 505–527.
[10] W. A. Coppel, *Linear-quadratic optimal control*, Proc. Roy. Soc. Edinburgh Sect. A, 73 (1975), pp. 271–289.
[11] A. L. Dontchev and W. W. Hager, *Lipschitzian stability in nonlinear control and optimization*, SIAM J. Control Optim., 31 (1992), pp. 569–603.
[12] R. Fletcher, *Practical Methods of Optimization. Vol.* 1. *Unconstrained Optimization*, John Wiley, Chichester, 1980.
[13] M. R. Hestenes, *Application of the theory of quadratic forms in Hilbert spaces to the calculus of variations*, Pacific J. Math., 1 (1951), pp. 525–580.
[14] M. W. Hirsh, *Differential topology*, Springer-Verlag, New York, Heidelberg, Berlin, 2000.
[15] A. D. Ioffe and V. M. Tikhomirov, *Theory of Extremal Problems*, North–Holland, Amsterdam, Holland, 1974.
[16] K. Malanowski, C. Büskens, and H. Maurer, *Convergence of approximations to nonlinear optimal control problems*, in Mathematical Programming with Data Perturbations, Lecture Notes in Pure and Appl. Math. 195, A. V. Fiacco, ed., Marcel Dekker, New York, 1998, pp. 253–284.
[17] H. D. Mittelmann, *Verification of second-order sufficient optimality conditions for semilinear and parabolic control problems*, Comput. Optim. Appl., 20 (2001), pp. 93–100.
[18] E. Polak, *Computational Methods in Optimization*, Academic Press, New York, London, 1971.
[19] W. Rudin, *Principles of Mathematical Analysis*, McGraw–Hill, New York, 1964.
[20] F. P. Vasil'yev, *Lectures on Methods for Solving Extremum Problems*, Moscow State University, Moscow, 1974 (in Russian).
[21] V. V. Voyevodin and Yu. A. Kuznetsov, *Matrices and Computations*, Nauka, Moscow, 1984 (in Russian).
[22] V. A. Yakubovich, *The frequency theorem in the control theory*, Siberian Math. J., 14 (1973), pp. 384–420.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 5, pp. 1586–1608

IDENTIFICATION FOR CONTROL: OPTIMAL INPUT DESIGN WITH RESPECT TO A WORST-CASE ν-GAP COST FUNCTION*

ROLAND HILDEBRAND† AND MICHEL GEVERS‡

Abstract. Parameter identification experiments deliver an identified model together with an ellipsoidal uncertainty region in parameter space. The objective of robust controller design is thus to stabilize all plants in the identified uncertainty region. The subject of the present contribution is to design an identification experiment such that the worst-case ν-gap over all plants in the resulting uncertainty region between the identified plant and plants in this region is as small as possible. The experiment design is performed via input power spectrum optimization. Two cost functions are investigated, which represent different levels of trade-off between accuracy and computational complexity. It is shown that the input optimization problem with respect to these cost functions is amenable to standard numerical algorithms used in convex analysis.

Key words. identification for control, worst-case ν-gap, parametric uncertainty region

AMS subject classifications. Primary 93E12; Secondary 93D21, 49M05

PII. S0363012901399866

1. Introduction.

This paper continues the line of research that aims at connecting prediction error identification methods with robust control theory ([2], [3], [4], [10]). Subject to investigation are discrete time SISO real-rational stable linear time invariant (LTI) plants, which are to be identified in open loop within an autoregressive with exogeneous input (ARX) model structure. We assume the true plant to lie in the model set. Hence the model error is determined only by the covariance of the estimated parameter vector.

Since the aim of the identification experiment is control design, it is desirable to obtain an uncertainty region with good stability robustness properties. By this is meant that the set of controllers that stabilize all models in the uncertainty set should be as large as possible. A suitable measure of robust stability that allows one to connect the "size" of an uncertainty set with a set of robustly stabilizing controllers is the worst-case ν-gap $\delta_{WC}(\hat{G}, \mathcal{D})$ introduced in [10]. It is the supremum of the Vinnicombe ν-gap (see, e.g., [28]) between the identified model $\hat{G}$ and all plants in the uncertainty set $\mathcal{D}$. Specifically, if $\delta_{WC}(\hat{G}, \mathcal{D}) = \beta$, then all controllers C that stabilize the model $\hat{G}$ with a stability margin $b_{\hat{G},C} > \beta$ stabilize all plants in $\mathcal{D}$.

In previous papers ([3], [4], [10]) a special type of uncertainty set $\mathcal{D}$ of transfer functions, which emerges from prediction error identification experiments, was described and investigated. It is given by an ellipsoid in parameter space and is determined by the covariance matrix of the parameter vector and the prespecified confidence level. The latter is defined to be the probability with which the true plant is lying inside the considered uncertainty set.

*Received by the editors December 19, 2001; accepted for publication (in revised form) July 5, 2002; published electronically January 14, 2003. The European Commission is herewith acknowledged for its financial support in part to the research reported on in this paper. The support is provided via the Program Training and Mobility of Researchers (TMR) and Project System Identification (ERB FMRX CT98 0206) to the European Research Network System Identification (ERNSI).
http://www.siam.org/journals/sicon/41-5/39986.html

†CORE, Université Catholique de Louvain, 34 voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium (hildebrand@core.ucl.ac.be).

‡CESAME, Université Catholique de Louvain, Bâtiment EULER, 4 av. Georges Lemaitre, 1348 Louvain-la-Neuve, Belgium (Gevers@auto.ucl.ac.be).

The goal of this paper is to minimize the worst-case ν-gap of such uncertainty regions $\mathcal{D}$ by choosing a suitable input $u(t)$ for the identification experiment. To restrict the class of admissible inputs we assume the total input energy to be bounded.

The problem setting of experiment design first arose in statistics and was extensively studied throughout the last century. Important results were obtained by Kiefer [15], Kiefer and Wolfowitz [16], Fedorov (e.g., [9]), Mehra (e.g., [20], [21]), Goodwin, Zarrop, and Payne [12], Zarrop [30], and others.

We shall adopt the most common viewpoint and study input optimization in the frequency domain, i.e., optimize the input power spectrum with respect to a cost function that depends on the average per data sample information matrix $\bar{M}$ of the experiment. This matrix is defined as the limit of the ratio between the information matrix and the number of data as the number of data tends to infinity (see, e.g., [30]). For typical number of data, this leads to a sufficiently good approximation of the optimal input. The latter can be obtained only by computationally expensive time domain optimization (see, e.g., [11], [23], [26], [6]). Thus we will essentially regard the average information matrix instead of the input power spectrum as the quantity that is going to be optimized. Once the optimal average information matrix, i.e., the one that minimizes the considered cost function, is found, we proceed by construction of an input power spectrum that produces this information matrix.

For different classes of cost functions iterative procedures were designed to find the optimal input power spectrum up to a prespecified precision. Most common cost functions are $\ln(\det \bar{M}^{-1})$ (D-optimality), $\mathrm{tr}\bar{M}^{-1}$ (A-optimality), $\mathrm{tr}W\bar{M}^{-1}$, where $W \geq 0$ (L-optimality), $\lambda_{\max}(\bar{M}^{-1})$ (E-optimality), $\Phi_s = (p^{-1}\mathrm{tr}\bar{M}^{-s})^{1/s}$, where p is the dimension of the parameter vector and $s = 0, 1, \ldots, \infty$ (Φ-optimality). All mentioned cost functions except $\Phi_\infty = \lambda_{\max}(\bar{M}^{-1})$ depend analytically on the entries of $\bar{M}$ and Kiefer–Wolfowitz theory can effectively be applied to them (see [15]). All above-mentioned criteria are convex and monotonic with respect to $\bar{M}$ (see [30, p. 39]).

In this paper, we optimize the input power spectrum with respect to the worst-case ν-gap of the uncertainty region $\mathcal{D}$. This is a nonstandard cost function, which is nonsmooth and thus more difficult to treat than the common above-mentioned criteria. We shall also introduce another cost function, which approximates the worst-case ν-gap, but is somewhat simpler. Nevertheless, both cost functions are compound criteria (see [15, section 4G]) and application of Kiefer–Wolfowitz theory does not make them more tractable. However, the proposed criteria satisfy the natural condition of monotonicity with respect to $\bar{M}$, as well as the condition of quasiconvexity, which is slightly weaker than convexity.

It follows from a classical result on trigonometric moment spaces (see [14, Chapter VI, Theorem 4.1]) that the set of possible average information matrices $\bar{M}$ can be represented as the feasible set of a linear matrix inequality (LMI). For a survey on LMIs, see, e.g., [5]. Since the worst-case ν-gap and the other proposed criteria are quasiconvex with respect to the input power spectrum, the apparatus of convex analysis and the theory of LMIs can be applied to solve this optimization problem. For recent results in convex optimization, see, e.g., [22].

In the last years several authors successfully treated input design problems arising in identification for control with convex optimization methods. In [18], the input spectrum for an open loop identification experiment was designed to minimize the closed loop system performance. By a Taylor series truncation, the cost function reduced to the weighted-trace criterion (L-optimality). However, the input spectra

were restricted to those which can be realized by white noise filtered through a finite impulse response (FIR) filter. An LMI description of the corresponding set of information matrices can be derived from the positive-real lemma [5], [29].

In this paper we optimize over the whole set of nonnegative input power spectra. It can be shown [30] that under the assumptions made above the corresponding set of admissible average information matrices, over which the optimization is performed, represents a moment space of a trigonometric Tchebycheff system. The foundations of the theory of moment spaces are classical. In the last century important contributions were made by Krein (see, e.g., [17]), Karlin and Shapley [13], and others. For a comprehensive treatment, see the textbook [14] by Karlin and Studden. It follows from a well-known fact of Tchebycheff system theory (see, e.g., [14]) that any admissible average information matrix $\bar{M}$ can be obtained by applying an input with discrete power spectrum, and that there exist admissible $\bar{M}$ which can be realized only by discrete power spectra. A restatement of this assertion is provided in Theorem 3.6 in this paper. In view of this, we propose an algorithm that yields optimal input power spectra which are discrete. Given the result just quoted, this is in no way a restriction. There are different ways to choose an input sequence with a desired power spectrum. We can choose the input, e.g., as a multisine function. However, in many cases one could use also binary signals (see, e.g., [30, p. 29]) or other functions.

Another approach, which leads to a suboptimal discrete input power spectrum, was proposed by Schoukens, Guillaume, and Pintelon [25] and van den Eijnde and Schoukens [27]. Here a finite subset of frequencies is prespecified and the optimal input power spectrum is sought within this subset. Advantages of this suboptimal method are less computational effort and an easier way to generate an input signal with the desired spectrum.

Let us also mention the paper [7], where identification in the ν-gap metric was treated outside the context of input design. The identification of a model was performed from a set of frequency response measurements in a way that aimed at minimizing the ν-gap between the true plant and the model.

We stress that the assumption of an ARX model structure and an input energy constraint are in no way restrictive. The ideas and methods proposed in the present paper easily carry over to other model structures and to input power or output power/energy constraints.

The remainder of the paper is structured as follows. In the next section the considered identification problem as well as the cost functions will be formally defined. In section 3 we will show that the set over which the optimization takes place is amenable to an LMI formulation. In section 4 we prove that the optimization problem is quasiconvex. In section 5 we show how to construct cutting planes to the different cost functions. Sections 3 to 5 are the key parts of the paper. The results obtained therein allow the problem to be treated with standard convex analysis methods. In section 6 we provide some results that are useful for designing stopping criteria for iterative search algorithms and quality assessment of the solution. Since the optimization takes place in an abstract parameter space, it is necessary to convert values in this space into power spectra and input sequences. This task is accomplished in section 7. In section 8 we present a simulation example, which demonstrates the superiority of the proposed cost functions over the classical design criteria D- and E-optimality. Finally, in section 9 we draw some conclusions.

2. Problem setting. Let us consider an ARX model structure

$$y(t) + a_1 y(t-1) + \cdots + a_{n_a} y(t-n_a) = b_1 u(t-n_k) + \cdots + b_{n_b} u(t-n_k-n_b+1) + e(t),$$

where $u(t)$ is the input signal, $y(t)$ is the output signal, both one-dimensional, $\theta = (a_1, \dots, a_{n_a}, b_1, \dots, b_{n_b})^T$ is the parameter vector, and $e(t)$ is normally distributed white noise with covariance λ_0. Let us assume that the true system dynamics can be described within this structure and corresponds to a parameter value $\theta = \theta_0$. Assume further that the true system is stable. Denote by z^{-1} the delay operator. Then we can write

$$\begin{aligned} y &= z^{-n_k+1}\frac{b_1 z^{-1} + \cdots + b_{n_b} z^{-n_b}}{1 + a_1 z^{-1} + \cdots + a_{n_a} z^{-n_a}} u + \frac{1}{1 + a_1 z^{-1} + \cdots + a_{n_a} z^{-n_a}} e \\ &= z^{-n_k+1}\frac{B(\theta)}{A(\theta)} u + \frac{1}{A(\theta)} e = G(\theta) u + \frac{1}{A(\theta)} e, \end{aligned}$$

where A, B are obviously defined polynomials in the delay operator with coefficients depending on the parameter vector. Note that by our stability assumption A has no zeros on the unit circle and hence $|A|^2$ is strictly positive there.

Suppose an identification experiment with input $(u(1), \dots, u(N))$ is performed, leading to an observed output $(y(1), \dots, y(N))$ with N data samples, where $u(t)$ is a realization of a quasi-stationary stochastic process with power spectrum Φ_u. Suppose a parameter estimate $\hat{\theta}$ is obtained by least squares prediction error minimization. Then it is well known (see [19]) that the estimate $\hat{\theta}$ is asymptotically unbiased as $N \to \infty$ and its covariance for large N is given by $E(\theta_0 - \hat{\theta})(\theta_0 - \hat{\theta})^T \approx \frac{\lambda_0}{N}(\bar{E}\psi\psi^T)^{-1}$, where $\psi^T = (-z^{-1}y, \dots, -z^{-n_a}y, z^{-n_k}u, \dots, z^{-n_k-n_b+1}u)$ is the gradient of the predictor with respect to θ at $\theta = \theta_0$. The power spectrum of ψ is given by

$$\Phi_\psi = \begin{pmatrix} -z^{-n_k}\frac{B}{A} \\ \vdots \\ -z^{-n_a-n_k+1}\frac{B}{A} \\ z^{-n_k} \\ \vdots \\ z^{-n_k-n_b+1} \end{pmatrix} \Phi_u \left(-z^{n_k}\frac{\bar{B}}{\bar{A}} \cdots z^{n_k+n_b-1} \right) + \begin{pmatrix} -\frac{z^{-1}}{A} \\ \vdots \\ -\frac{z^{-n_a}}{A} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \lambda_0 \left(-\frac{z}{\bar{A}} \cdots 0 \right).$$

This yields the following asymptotic expression for the parameter covariance:

$$E(\theta_0 - \hat{\theta})(\theta_0 - \hat{\theta})^T \approx \left(\frac{N}{2\pi} \int_{-\pi}^{\pi} \frac{1}{|A|^2} \left(\frac{\Phi_u}{\lambda_0} \begin{pmatrix} -z^{-1}B \\ \vdots \\ -z^{-n_a}B \\ z^{-1}A \\ \vdots \\ z^{-n_b}A \end{pmatrix} \begin{pmatrix} -z^{-1}B \\ \vdots \\ -z^{-n_a}B \\ z^{-1}A \\ \vdots \\ z^{-n_b}A \end{pmatrix}^* + \begin{pmatrix} -z^{-1} \\ \vdots \\ -z^{-n_a} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{pmatrix} -z^{-1} \\ \vdots \\ -z^{-n_a} \\ 0 \\ \vdots \\ 0 \end{pmatrix}^* \right) d\omega \right)^{-1} = M^{-1},$$

where $z = e^{j\omega}$. The inverse of the parameter covariance matrix is the Fisher information matrix. Let us denote the asymptotic expression for the information matrix by

M and the average information matrix per data sample (see, e.g., [30, p. 24]) by $\bar{M}$, $\bar{M} = \frac{1}{N}M$. We obtain

$$\bar{M} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\Phi_u}{\lambda_0 |A|^2} \begin{pmatrix} |B|^2 \begin{pmatrix} 1 & \cdots & z^{n_a-1} \\ \vdots & \ddots & \vdots \\ z^{-n_a+1} & \cdots & 1 \end{pmatrix} & -B\bar{A} \begin{pmatrix} 1 & \cdots & z^{n_b-1} \\ \vdots & \ddots & \vdots \\ z^{-n_a+1} & \cdots & z^{n_b-n_a} \end{pmatrix} \\ -\bar{B}A \begin{pmatrix} 1 & \cdots & z^{n_a-1} \\ \vdots & \ddots & \vdots \\ z^{-n_b+1} & \cdots & z^{n_a-n_b} \end{pmatrix} & |A|^2 \begin{pmatrix} 1 & \cdots & z^{n_b-1} \\ \vdots & \ddots & \vdots \\ z^{-n_b+1} & \cdots & 1 \end{pmatrix} \end{pmatrix}$$

$$+ \frac{1}{|A|^2} \begin{pmatrix} 1 & \cdots & z^{n_a-1} & \\ \vdots & \ddots & \vdots & 0 \\ z^{-n_a+1} & \cdots & 1 & \\ & 0 & & 0 \end{pmatrix} d\omega. \tag{2.1}$$

Note that in the expansion of $\bar{M}$, we have $A = A(\theta_0)$, $B = B(\theta_0)$. Since the parameter estimate $\hat{\theta}$ is asymptotically normally distributed (see [19]), we can assume, following [10], that the true parameter vector θ_0 lies with a prespecified probability $\alpha \in (0,1)$ in the uncertainty ellipsoid

$$U = \left\{ \theta | \frac{N}{\chi^2_{n_a+n_b}(\alpha)} (\theta - \hat{\theta})^T \bar{M} (\theta - \hat{\theta}) < 1 \right\}, \tag{2.2}$$

where χ^2_l is the χ^2 probability distribution with l degrees of freedom.

The uncertainty ellipsoid U corresponds to an uncertainty set

$$\mathcal{D} = \left\{ G(z,\theta) = z^{-n_k+1} \frac{B(\theta)}{A(\theta)} | \theta \in U \right\} = \left\{ G(z,\theta) = \frac{Z_N(z)\theta}{1 + Z_D(z)\theta} | \theta \in U \right\}$$

in the space of transfer functions. Here

$$Z_N = z^{-n_k+1}(0 \ \cdots \ 0 \ z^{-1} \ \cdots \ z^{-n_b}), \qquad Z_D = (z^{-1} \ \cdots \ z^{-n_a} \ 0 \ \cdots \ 0) \tag{2.3}$$

are row vectors of dimension $n_a + n_b$. The set $\mathcal{D}$ belongs to the class of generic prediction error model uncertainty sets as defined in [10].

The worst-case ν-gap between the identified model $G(\hat{\theta})$ and the uncertainty region $\mathcal{D}$ is defined by

$$\delta_{WC}(G(\hat{\theta}), \mathcal{D}) = \sup_{\theta \in U} \delta_\nu(G(\hat{\theta}), G(\theta)), \tag{2.4}$$

where δ_ν denotes the Vinnicombe ν-gap between two plants (see [28]). Since $G(\hat{\theta})$ belongs to $\mathcal{D}$, the worst-case ν-gap can be expressed in the following way (see [10, Lemma 5.1]):

$$\delta_{WC}(G(\hat{\theta}), \mathcal{D}) = \sup_{\omega \in [0,\pi]} \kappa_{WC}(G(e^{j\omega}, \hat{\theta}), \mathcal{D}), \tag{2.5}$$

where $\kappa_{WC}(G(e^{j\omega}, \hat{\theta}), \mathcal{D})$ is called the worst-case chordal distance between $G(\hat{\theta})$ and $\mathcal{D}$ at frequency ω and is defined by

$$\kappa_{WC}(G(e^{j\omega}, \hat{\theta}), \mathcal{D}) = \sup_{\theta \in U} \frac{|G(e^{j\omega}, \hat{\theta}) - G(e^{j\omega}, \theta)|}{\sqrt{(1 + |G(e^{j\omega}, \hat{\theta})|^2)(1 + |G(e^{j\omega}, \theta)|^2)}}. \tag{2.6}$$

The worst-case ν-gap is directly related to the robustness properties of the uncertainty region $\mathcal{D}$. The smaller it is, the larger is the set of controllers stabilizing simultaneously all plants in $\mathcal{D}$. Therefore our primary goal shall be to minimize the quantity $\delta_{WC}(G(\hat{\theta}), \mathcal{D}) = \max_{\omega\in[0,\pi]} \kappa_{WC}(G(e^{j\omega}, \hat{\theta}), \mathcal{D})$ by choosing an input with an appropriate power spectrum.

To be more precise, by input spectrum we mean a nonnegative measure on $[-\pi, \pi]$ such that the equality $\int_{-\pi}^{\pi} \Phi_u \varphi(\omega)\, d\omega = \int_{-\pi}^{\pi} \Phi_u \varphi(-\omega)\, d\omega$ holds for all functions $\varphi(\omega) \in C^\infty([-\pi, \pi])$. To any such measure Φ_u on $[-\pi, \pi]$ corresponds a unique nonnegative measure $\bar{\Phi}_u$ on $[0, \pi]$ such that $\int_{-\pi}^{\pi} \Phi_u \varphi(\omega)\, d\omega = \int_0^{\pi} \bar{\Phi}_u \frac{\varphi(\omega)+\varphi(-\omega)}{2}\, d\omega$ for all $\varphi \in C^\infty([-\pi, \pi])$. For details on constructing $\bar{\Phi}_u$ from Φ_u, see, e.g., [30, p. 23]. In what follows we will denote the single-sided measure $\bar{\Phi}_u$ also by Φ_u. Since the measures are defined on different intervals, confusion is excluded.

To restrict the class of admissible power spectra we impose an input energy constraint

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \Phi_u(\omega) d\omega \le c, \tag{2.7}$$

where $c > 0$ is a prespecified positive constant.

The worst-case ν-gap depends on Φ_u via the average per data sample information matrix $\bar{M}$, which enters in the expression for the set U. Furthermore, it depends via $\bar{M}$ on the unknown true parameter value θ_0 and noise covariance λ_0. In addition it depends on the identified parameter value $\hat{\theta}$, which is naturally not available before the identification experiment is performed. All these three quantities have to be approximated with values derived from previous knowledge about the system, for instance from a preliminary identification experiment. Since the expectation of $\hat{\theta}$ equals θ_0, these two quantities can be approximated by the same value. Denote this value by $\bar{\theta}$, and denote the approximation of λ_0 by $\bar{\lambda}$.

We can now formulate our main problems.

PROBLEM 1. *Find Φ_u satisfying* (2.7) *such that $\bar{M}(\Phi_u)$ defined by* (2.1) *minimizes the cost function $\mathcal{J}_1 = \delta_{WC}(G(\hat{\theta}), \mathcal{D})$ defined by* (2.5), (2.6).

Along with the worst-case ν-gap of the uncertainty region $\mathcal{D}$, we will consider another cost function, which is easier to compute and is an approximation of δ_{WC}.

Let us approximate cost function $\mathcal{J}_1 = \mathcal{J}_1(\bar{M})$ by its asymptotic expression for large information matrices. For a fixed positive definite matrix $\bar{M}_0$ the size of the parameter ellipsoid U defined by any multiple $\bar{M} = \beta \bar{M}_0$ of $\bar{M}_0$, where $\beta > 0$, is proportional to $\beta^{-1/2}$. Since for small ellipsoids the worst-case ν-gap is asymptotically proportional to the size of the former, it follows that for large β the value of $\mathcal{J}_1(\bar{M})$ diminishes asymptotically proportionately to $\beta^{-1/2}$. Thus we can approximate $\mathcal{J}_1$ by

$$\mathcal{J}_2 = \lim_{\varepsilon\to 0} \frac{\mathcal{J}_1(\varepsilon^{-2}\bar{M})}{\varepsilon}. \tag{2.8}$$

PROBLEM 2. *Find Φ_u satisfying* (2.7) *such that $\bar{M}(\Phi_u)$ defined by* (2.1) *minimizes cost function $\mathcal{J}_2$ defined by* (2.8).

The goal of the present paper is the development of numerical algorithms for solving both Problems 1 and 2. There is a twofold reason for introducing cost function $\mathcal{J}_2$. Beside its much lower computational complexity, it turns out that identification with an input power spectrum minimizing $\mathcal{J}_2$ in many cases gives better results than one with an input power spectrum minimizing $\mathcal{J}_1$. This apparently counterintuitive observation has the following reason. Both cost functions depend on the identified

parameter value $\hat{\theta}$, the true parameter value θ_0, and the noise covariance λ_0. As mentioned above, these quantities are unknown and must be replaced by estimates obtained, e.g., from a preliminary identification experiment. This approximation introduces an error to the argument of the minimum of the cost functions $\mathcal{J}_1$ and $\mathcal{J}_2$, i.e., to the solutions of Problems 1 and 2. Now simulations show that the impact of this effect on $\arg\min \mathcal{J}_2$ is lower than that on $\arg\min \mathcal{J}_1$ and that this difference as a rule outweighs the error introduced by approximating cost function $\mathcal{J}_1$ by $\mathcal{J}_2$. We will address this issue again in the simulation section.

3. LMI description of the search space. In this section we shall describe the set of possible average information matrices $\bar{M}$, over which the optimization takes place, as the feasible set of an LMI.

The following fact is due to Payne and Goodwin [24].

PROPOSITION 3.1. *The average information matrix $\bar{M}$ is contained in a (n_a+n_b)-dimensional affine subspace of the space of symmetric $(n_a+n_b)\times(n_a+n_b)$-matrices.*

We find it convenient to give a proof here in order to provide explicit expressions that clarify the structure of $\bar{M}$.

Proof. Define $a_0 = 1$ and $n = n_a + n_b - 1$. Then we have

$$
\begin{aligned}
|B|^2 &= \sum_{k=-(n_b-1)}^{n_b-1} \left(\sum_{j=\max(1,1-k)}^{\min(n_b,n_b-k)} b_{j+k} b_j \right) z^k, \\
-B\bar{A} &= \sum_{k=-n_b}^{n_a-1} \left(\sum_{j=\max(1,-k)}^{\min(n_b,n_a-k)} -a_{j+k} b_j \right) z^k, \\
|A|^2 &= \sum_{k=-n_a}^{n_a} \left(\sum_{j=\max(0,-k)}^{\min(n_a,n_a-k)} a_{j+k} a_j \right) z^k. \qquad (3.1)
\end{aligned}
$$

Using (3.1) in (2.1) and ordering by powers of z, we can rewrite (2.1) as $\bar{M} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\Phi_u}{\lambda_0|A|^2}(\sum_{i=-n}^{n} \tilde{M}_i z^i)\, d\omega + \tilde{M}$. The matrices $\tilde{M}_i$ are constant and depend only on the coefficients of A and B. By $\tilde{M}$ the integral over the second term in (2.1) is denoted. It is a constant matrix and independent of Φ_u. $\tilde{M}$ is most easily computed using the method proposed in [19, p. 50]. Note that $\tilde{M}_i = \tilde{M}_{-i}^T$. Hence we obtain

$$
\begin{aligned}
\bar{M} &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\Phi_u}{\lambda_0|A|^2}\, d\omega \tilde{M}_0 + \sum_{i=1}^{n} \left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\Phi_u}{\lambda_0|A|^2} z^i\, d\omega (\tilde{M}_i + \tilde{M}_i^T) \right) + \tilde{M} \\
&= \frac{1}{\pi}\int_{0}^{\pi} \frac{\Phi_u}{\lambda_0|A|^2}\, d\omega \frac{\tilde{M}_0}{2} + \sum_{i=1}^{n} \left(\frac{1}{\pi}\int_{0}^{\pi} \frac{\Phi_u}{\lambda_0|A|^2} \cos(i\omega)\, d\omega \frac{\tilde{M}_i + \tilde{M}_i^T}{2} \right) + \tilde{M}. \qquad (3.2)
\end{aligned}
$$

Thus $\bar{M}$ is contained in the $(n+1)$-dimensional affine subspace that is spanned by $\tilde{M}_0, \tilde{M}_i + \tilde{M}_i^T$, $i = 1, \ldots, n$, and shifted by $\tilde{M}$. This completes the proof. ∎

Let us compose a vector $\tilde{x} \in \mathbf{R}^{n+1}$ of real numbers $\tilde{x}_i$, $i = 0, \ldots, n$, defined by $\tilde{x}_i = \frac{1}{\pi}\int_0^{\pi} \frac{\Phi_u}{\lambda_0|A|^2} \cos(i\omega)\, d\omega$.

DEFINITION 3.2. *The quantities $\tilde{x}_k = \frac{1}{\pi}\int_0^{\pi} \frac{\Phi_u}{\lambda_0|A|^2} \cos(k\omega)\, d\omega$, $k \in \mathbf{N}$, are called* trigonometric moments *of the measure $\frac{\Phi_u}{\pi\lambda_0|A|^2}$.*

Since $\frac{1}{\pi\lambda_0|A|^2}$ is strictly positive on $\omega \in [0, \pi]$, we have the following result [30].

PROPOSITION 3.3. *The set* $\{\tilde{x}(\Phi_u) \,|\, \Phi_u$ *is a nonnegative measure on* $[0,\pi]\}$ *equals the moment space* $\mathcal{M}^{(n+1)}$ *of the Tchebycheff system* $\{1, \cos\omega, \ldots, \cos n\omega\}$ *on* $[0,\pi]$.

For definition and properties of moment spaces, see, e.g., [14].

The characterization of the space $\mathcal{M}^{(n+1)}$ is a special case of the extensively studied classical trigonometric moment problem. The following theorem is a consequence of the general result [14, Chapter VI, Theorem 4.1]. It asserts that $\mathcal{M}^{(n+1)}$ can be characterized as the feasible set of an LMI.

THEOREM 3.4. *A point* $\tilde{x} \in \mathbf{R}^{n+1}$ *belongs to the space* $\mathcal{M}^{(n+1)}$ *if and only if the Töplitz matrix composed of the elements of* $\tilde{x}$ *is positive semidefinite, i.e.,*

$$T(\tilde{x}) = \begin{pmatrix} \tilde{x}_0 & \tilde{x}_1 & \ddots & \tilde{x}_n \\ \tilde{x}_1 & \tilde{x}_0 & \ddots & \tilde{x}_{n-1} \\ \ddots & \ddots & \ddots & \ddots \\ \tilde{x}_n & \tilde{x}_{n-1} & \ddots & \tilde{x}_0 \end{pmatrix} \geq 0. \qquad \square \tag{3.3}$$

Since $n+1 \geq 2$, the strict LMI $T(\tilde{x}) > 0$ is feasible. Hence the feasible set of the strict version is the interior of $\mathcal{M}^{(n+1)}$ (see [5, section 2.5]). By $\mathcal{M}$ denote the set of average information matrices corresponding to the interior of $\mathcal{M}^{(n+1)}$. From (3.2) we have

$$\mathcal{M} = \left\{ \bar{M}(\tilde{x}) = \tilde{x}_0 \frac{\tilde{M}_0}{2} + \sum_{i=1}^{n} \tilde{x}_i \frac{\tilde{M}_i + \tilde{M}_i^T}{2} + \tilde{M} \,|\, T(\tilde{x}) > 0 \right\}.$$

DEFINITION 3.5. *Let* Φ_u *be a discrete double-sided power spectrum with support* $\operatorname{supp}\Phi_u \subset [-\pi,\pi]$. *The number of points in the intersection* $\operatorname{supp}\Phi_u \cap [-\pi,\pi)$, *divided by two, is called the index of* Φ_u: *index*$(\Phi_u) = \frac{1}{2}\#(\operatorname{supp}\Phi_u \cap [-\pi,\pi))$. *The index of a single-sided nonnegative discrete measure on* $[0,\pi]$ *is defined as the index of the corresponding double-sided power spectrum.*

Remark. This definition of the index is consistent with its definition for nonnegative discrete measures on the interval $[0,\pi]$ (see, e.g., [14]).

The notion of the index also allows us to characterize the interior of the moment space $\mathcal{M}^{(n+1)}$. The following theorem is a standard result on moment spaces.

THEOREM 3.6 (see, e.g., [14]). *Let* $\tilde{x}$ *be a point in* $\mathcal{M}^{(n+1)}$. *Then the following conditions hold:*

(i) $\tilde{x} \in \mathrm{Bd}(\mathcal{M}^{(n+1)})$ *if and only if there exists a discrete nonnegative measure on* $[0,\pi]$ *with index less than* $\frac{n+1}{2}$ *that induces* $\tilde{x}$. *This measure is unique.*

(ii) $\tilde{x} \in \mathrm{Int}(\mathcal{M}^{(n+1)})$ *if and only if there exists a discrete nonnegative measure on* $[0,\pi]$ *with index* $\frac{n+1}{2}$ *that induces* $\tilde{x}$. *There are exactly two such measures. Exactly one of them contains the frequency* π.

(iii) *Let* $\tilde{x} \in \mathrm{Int}(\mathcal{M}^{(n+1)})$ *and* $\omega \in [0,\pi]$. *Then there exists a unique discrete nonnegative measure on* $[0,\pi]$ *which induces* $\tilde{x}$, *has index not exceeding* $\frac{n+2}{2}$, *and contains the frequency* ω. $\square$

Remark. Measures with index $\frac{n+1}{2}$ which induce $\tilde{x}$ are called *principal realizations* of $\tilde{x}$. The one containing π is called *upper principal*, the other *lower principal*, realization. Measures with index not exceeding $\frac{n+2}{2}$ are called *canonical*.

We see that the interior of $\mathcal{M}^{(n+1)}$ is characterized by those points $\tilde{x}$ which can be represented by a discrete measure with index not less than $\frac{n+1}{2}$.

Now we shall characterize the set of input power spectra Φ_u that lead to nonsingular average information matrices $\bar{M}$.

PROPOSITION 3.7. *Let Φ_u be a power spectrum and $\bar{M}$ the corresponding average information matrix. Then $\bar{M}$ is singular if and only if Φ_u is discrete and its index is less than $\frac{n_b}{2}$.*

Proof. "⇒": Suppose $\bar{M}(\Phi_u)$ is singular. Then there exists a nonzero vector $v = (v_1, \dots, v_{n_a+n_b})^T \in \mathbf{R}^{n_a+n_b}$ such that $v^T \bar{M} v = 0$. Expanding $\bar{M}$, we obtain

$$v^T \bar{M} v = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\frac{\Phi_u}{\lambda_0 |A|^2} | -v_1 z^{-1} B - \dots - v_{n_a} z^{-n_a} B + v_{n_a+1} z^{-1} A + \dots + v_{n_a+n_b} z^{-n_b} A|^2 \right.$$
$$\left. + \frac{1}{|A|^2} | - v_1 z^{-1} - \dots - v_{n_a} z^{-n_a}|^2 \right) d\omega = 0$$

with $z = e^{j\omega}$. This yields $-v_1 z^{-1} - \dots - v_{n_a} z^{-n_a} = 0$ for all z on the unit circle and $-v_1 z^{-1} B - \dots - v_{n_a} z^{-n_a} B + v_{n_a+1} z^{-1} A + \dots + v_{n_a+n_b} z^{-n_b} A = 0$ for all $z = e^{j\omega}$ such that $\omega \in \operatorname{supp} \Phi_u$. From the first identity we obtain $v_1 = \dots = v_{n_a} = 0$. Inserting this in the second equality, we get $v_{n_a+1} + \dots + v_{n_a+n_b} z^{-n_b+1} = 0$. Since $v \neq 0$, this equation can have at most $n_b - 1$ different roots. Since Φ_u has to be concentrated at these roots, it is discrete and its index cannot exceed $\frac{n_b-1}{2}$.

"⇐": Suppose Φ_u is discrete with index less than $\frac{n_b}{2}$. Denote the frequencies of Φ_u by $\omega_1, \dots, \omega_{k'}$. They correspond to k different points $z_1, \dots, z_k$ on the unit circle, where $k < n_b$. We have

$$\bar{M} = \frac{1}{2\pi} \sum_{i=1}^{k} \frac{\alpha_i}{\lambda_0 |A(z_i)|^2} \begin{pmatrix} -z_i^{-1} B(z_i) \\ \vdots \\ -z_i^{-n_a} B(z_i) \\ z_i^{-1} A(z_i) \\ \vdots \\ z_i^{-n_b} A(z_i) \end{pmatrix} \begin{pmatrix} -z_i^{-1} B(z_i) \\ \vdots \\ -z_i^{-n_a} B(z_i) \\ z_i^{-1} A(z_i) \\ \vdots \\ z_i^{-n_b} A(z_i) \end{pmatrix}^* + \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{|A|^2} \begin{pmatrix} -z^{-1} \\ \vdots \\ -z^{-n_a} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{pmatrix} -z^{-1} \\ \vdots \\ -z^{-n_a} \\ 0 \\ \vdots \\ 0 \end{pmatrix}^* d\omega.$$

Here $\alpha_i > 0$ are the weightings of the different frequencies. It is easily seen that the matrices under the sign of the sum are of (complex) rank one, while the integral is a matrix which has a rank of at most n_a. Thus the rank of $\bar{M}$ does not exceed $n_a + n_b - 1$ and $\bar{M}$ is singular. This concludes the proof. ∎

COROLLARY 3.8. *Any $\bar{M} \in \mathcal{M}$ is strictly positive definite.* ∎

This corollary ensures the existence of the inverse $\bar{M}^{-1}$ in the interior of the search space.

By inspecting (2.2), (2.4), and (2.8), the reader will have no difficulty in proving the following monotonicity property.

PROPOSITION 3.9. *Let $\bar{M}_1, \bar{M}_2$ be two positive semidefinite average information matrices, and suppose $\bar{M}_1 \leq \bar{M}_2$. Then the values of the cost functions $\mathcal{J}_1, \mathcal{J}_2$ at $\bar{M}_2$ do not exceed the respective values at $\bar{M}_1$.* ∎

Now we shall include the input energy constraint (2.7) into our framework. By Proposition 3.9, for any constant $\beta > 1$ the value of each of the considered cost functions at a particular input power spectrum Φ_u will be not less than its value at the input power spectrum $\beta \Phi_u$. Thus we can replace constraint (2.7) by

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \Phi_u(\omega) d\omega = c. \tag{3.4}$$

In [30] it was shown that relations like (3.4) determine affine hyperplanes in the space of feasible average information matrices. Indeed, we have

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\Phi_u(\omega)d\omega = \frac{\lambda_0}{2\pi}\int_{-\pi}^{\pi}\frac{\Phi_u}{\lambda_0|A|^2}\left(\sum_{i=0}^{n_a}a_i^2+\sum_{i=1}^{n_a}\left(2\sum_{k=0}^{n_a-i}a_{k+i}a_k\right)\cos i\omega\right)d\omega$$
$$= \lambda_0\left(\tilde{x}_0\sum_{i=0}^{n_a}a_i^2+\sum_{i=1}^{n_a}\tilde{x}_i\left(2\sum_{k=0}^{n_a-i}a_{k+i}a_k\right)\right)=c.$$

Thus we get

$$\tilde{x}_0 = \frac{1}{\sum_{i=0}^{n_a}a_i^2}\left(\frac{c}{\lambda_0}-\sum_{i=1}^{n_a}\tilde{x}_i\left(2\sum_{k=0}^{n_a-i}a_{k+i}a_k\right)\right). \tag{3.5}$$

Inserting (3.5) into (3.3), we obtain an LMI on the variables $\tilde{x}_1,\dots,\tilde{x}_n$, i.e., in an n-dimensional space instead of the initial $n+1$-dimensional one. The feasible set of LMI (3.5), (3.3) is a subset of $\mathbf{R}^n$, parametrized by new variables $x_1,\dots,x_n$, which we define by $x_1=\tilde{x}_1,\dots,x_n=\tilde{x}_n$. Denote by $\mathcal{X}_c$ the interior of this set and by $\mathcal{M}_c$ the set of average information matrices corresponding to points in $\mathcal{X}_c$. Thus the optimization takes place over the closure of $\mathcal{X}_c$. Let us stack the variables $x_1,\dots,x_n$ into a vector $x\in\mathbf{R}^n$, to be distinguished from $\tilde{x}\in\mathbf{R}^{n+1}$. While the latter parametrizes the set $\mathcal{M}$, the former parametrizes the set $\mathcal{M}_c$ or $\mathcal{X}_c$.

Using (3.5), we can represent average information matrices in the closure of $\mathcal{M}_c$ as affine functions of the variables $x_1,\dots,x_n$. We have

$$\bar{M} = \bar{M}_0+\sum_{i=1}^{n}x_i\bar{M}_i, \tag{3.6}$$

where

$$\bar{M}_0 = \frac{c}{2\lambda_0\sum_{i=0}^{n_a}a_i^2}\tilde{M}_0+\tilde{M},$$
$$\bar{M}_i = \frac{\tilde{M}_i+\tilde{M}_i^T}{2}-\frac{\sum_{k=0}^{n_a-i}a_{k+i}a_k}{\sum_{i=0}^{n_a}a_i^2}\tilde{M}_0, \qquad i=1,\dots,n_a,$$
$$\bar{M}_i = \frac{\tilde{M}_i+\tilde{M}_i^T}{2}, \qquad i=n_a+1,\dots,n.$$

Thus the closure of $\mathcal{M}_c$ is contained in an n-dimensional affine subspace of the space of symmetric $(n_a+n_b)\times(n_a+n_b)$-matrices.

PROPOSITION 3.10. *The search space of Problems* 1 *and* 2 *can be represented as a section of the trigonometric moment cone* $\mathcal{M}^{(n+1)}$ *and is thus a bounded closed* n*-dimensional convex set. It is parametrized by the variables* $x_1,\dots,x_n$.

Proof. What is left to prove is that relation (3.5) defines a section of the moment cone $\mathcal{M}^{(n+1)}$. Let $\tilde{x}$ be an arbitrary nonzero moment point and $\Phi_u(\omega)$ a measure generating this moment point. Then the ray $\beta\tilde{x}$, $\beta>0$, will be generated by the ray $\beta\Phi_u(\omega)$ of measures. On the latter, exactly one measure satisfies relation (3.4). Therefore exactly one point on the ray $\beta\tilde{x}$ satisfies relation (3.5). ∎

In this section we reduced the infinite-dimensional problem of searching the minimum of the cost functions over the set of all admissible input power spectra to the finite-dimensional problem of searching the minimum over a section of a moment cone. Moreover, we described the search space as an LMI, namely (3.3), (3.5), and showed that it is bounded.

4. Quasiconvexity. In the previous section we proved the search space to be a bounded convex set. In this section we prove quasiconvexity of cost functions $\mathcal{J}_1, \mathcal{J}_2$ and thus of Problems 1 and 2.

PROPOSITION 4.1. *On $\mathcal{M}$ cost function $\mathcal{J}_1$ is quasiconvex with respect to $\bar{M}$.*

Proof. The worst-case chordal distance can be expressed as a solution to a generalized eigenvalue problem (GEVP) [10, Theorem 5.1]. We have $\kappa_{WC}(G(e^{j\omega}, \hat{\theta}), \mathcal{D}) = \sqrt{\gamma_{opt}}$, where γ_{opt} is the solution of the GEVP

$$\text{minimize } \gamma \text{ subject to } \quad F_0 + \gamma F_1 + \tau R \geq 0, \quad \tau \geq 0. \tag{4.1}$$

Here F_0, F_1, R are symmetric matrices given by

$$F_0 = V \begin{pmatrix} -1 & 0 & -\mathrm{Im}G(e^{j\omega}, \hat{\theta}) & \mathrm{Re}G(e^{j\omega}, \hat{\theta}) \\ 0 & -1 & \mathrm{Re}G(e^{j\omega}, \hat{\theta}) & \mathrm{Im}G(e^{j\omega}, \hat{\theta}) \\ -\mathrm{Im}G(e^{j\omega}, \hat{\theta}) & \mathrm{Re}G(e^{j\omega}, \hat{\theta}) & -|G(e^{j\omega}, \hat{\theta})|^2 & 0 \\ \mathrm{Re}G(e^{j\omega}, \hat{\theta}) & \mathrm{Im}G(e^{j\omega}, \hat{\theta}) & 0 & -|G(e^{j\omega}, \hat{\theta})|^2 \end{pmatrix} V^T,$$

$$F_1 = (1 + |G(e^{j\omega}, \hat{\theta})|^2) V V^T,$$

$$R = \begin{pmatrix} I_{n_a+n_b} \\ -\hat{\theta}^T \end{pmatrix} \bar{M} \begin{pmatrix} I_{n_a+n_b} & -\hat{\theta} \end{pmatrix} - \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\chi^2_{n_a+n_b}(\alpha)}{N} \end{pmatrix}, \tag{4.2}$$

where V is a $(n_a + n_b + 1) \times 4$-matrix defined by

$$V = \begin{pmatrix} \mathrm{Re}Z_N^T & \mathrm{Im}Z_N^T & \mathrm{Im}Z_D^T & \mathrm{Re}Z_D^T \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

with Z_N, Z_D given by (2.3).

We will now show that γ_{opt} is quasiconvex with respect to R. Choose $\lambda \in (0,1)$ and let R_1, R_2 be symmetric matrices of appropriate dimension. Suppose γ, τ_1, τ_2 are nonnegative numbers such that $F_0 + \gamma F_1 + \tau_1 R_1 \geq 0, F_0 + \gamma F_1 + \tau_2 R_2 \geq 0$. We have to show that there exists $\tau \geq 0$ such that $F_0 + \gamma F_1 + \tau(\lambda R_1 + (1-\lambda) R_2) \geq 0$. If $\tau_1 = 0$ or $\tau_2 = 0$, then we can choose $\tau = 0$. Let $\tau_1 \tau_2 > 0$. Define

$$\lambda' = \frac{\lambda \tau_2}{\lambda \tau_2 + (1-\lambda)\tau_1}, \quad \tau = \frac{\tau_1 \tau_2}{\lambda \tau_2 + (1-\lambda)\tau_1}.$$

Obviously $\lambda' \in (0,1)$ and $\tau > 0$. It is easily verified that $\lambda\tau = \lambda'\tau_1, (1-\lambda)\tau = (1-\lambda')\tau_2$. Hence we have

$$F_0 + \gamma F_1 + \tau(\lambda R_1 + (1-\lambda) R_2) = \lambda'(F_0 + \gamma F_1 + \tau_1 R_1) + (1-\lambda')(F_0 + \gamma F_1 + \tau_2 R_2) \geq 0.$$

Thus if γ is feasible for $R = R_1$ and for $R = R_2$, then it is also feasible for any linear convex combination of R_1, R_2. It follows that $\gamma_{opt}|_{R=\lambda R_1 + (1-\lambda) R_2} \leq \max\{\gamma_{opt}|_{R=R_1}, \gamma_{opt}|_{R=R_2}\}$, i.e., quasiconvexity of γ_{opt} with respect to R.

Suppose $\omega \in [0, \pi]$ is fixed. Note that R affinely depends on $\bar{M}$, while F_0 and F_1 are constant for given ω. Therefore γ_{opt} is quasiconvex with respect to $\bar{M}$ for fixed ω. But quasiconvexity is preserved under the operation of taking the maximum over a family of functions and under rescaling by a strictly monotonic function (in this case the square root). This completes the proof. ∎

PROPOSITION 4.2. *On $\mathcal{M}$, cost function $\mathcal{J}_2$ is quasiconvex with respect to $\bar{M}$.*

Proof. Let us compute cost function $\mathcal{J}_2$.

$$\mathcal{J}_2 = \lim_{\varepsilon\to 0}\frac{\mathcal{J}_1(\varepsilon^{-2}\bar{M})}{\varepsilon} = \sup_{\omega\in[0,\pi]}\lim_{\varepsilon\to 0}\left(\varepsilon^{-1}\sup_{z\in U_\varepsilon(\omega)}\frac{|G(e^{j\omega},\hat{\theta})-z|}{\sqrt{1+|G(e^{j\omega},\hat{\theta})|^2}\sqrt{1+|z|^2}}\right)$$
$$= \sup_{\omega\in[0,\pi]}\frac{d}{d\varepsilon}|_{\varepsilon=0}\sup_{z\in U_\varepsilon(\omega)}\frac{|G(e^{j\omega},\hat{\theta})-z|}{\sqrt{1+|G(e^{j\omega},\hat{\theta})|^2}\sqrt{1+|z|^2}}.$$

Here $U_\varepsilon(\omega)$ denotes the set $\{z = G(e^{j\omega},\theta)|\frac{N}{\chi^2_{n_a+n_b}(\alpha)}(\theta-\hat{\theta})^T\bar{M}(\theta-\hat{\theta}) < \varepsilon^2\}$. The expression $\sqrt{1+|z|^2}$ tends to $\sqrt{1+|G(e^{j\omega},\hat{\theta})|^2}$ as $\varepsilon\to 0$; therefore

$$\mathcal{J}_2 = \sup_{\omega\in[0,\pi]}\frac{\frac{d}{d\varepsilon}|_{\varepsilon=0}\sup_{z\in U_\varepsilon(\omega)}|z-G(e^{j\omega},\hat{\theta})|}{1+|G(e^{j\omega},\hat{\theta})|^2} = \sup_{\omega\in[0,\pi]}\frac{\varepsilon^{-1}\sup_{z\in U_\varepsilon(\omega)}|T(\theta-\hat{\theta})|}{1+|G(e^{j\omega},\hat{\theta})|^2},$$

where the $2\times(n_a+n_b)$-matrix T is given by

$$T = \begin{pmatrix} \mathrm{Re}\frac{\partial G(e^{j\omega},\theta)}{\partial\theta}|_{\theta=\hat{\theta}} \\ \mathrm{Im}\frac{\partial G(e^{j\omega},\theta)}{\partial\theta}|_{\theta=\hat{\theta}} \end{pmatrix}.$$

If T has full rank, then, following [2], we can write the term $\varepsilon^{-1}\sup_{z\in U_\varepsilon(\omega)}|T(\theta-\hat{\theta})|$ as

$$\left(\lambda_{\min}\left(\left(T\left(\frac{N}{\chi^2_{n_a+n_b}(\alpha)}\bar{M}\right)^{-1}T^T\right)^{-1}\right)\right)^{-1/2} = \sqrt{\frac{\chi^2_{n_a+n_b}(\alpha)}{N}}(\lambda_{\max}(T\bar{M}^{-1}T^T))^{1/2}.$$

By $\lambda_{\min}$ and $\lambda_{\max}$ the minimal and maximal eigenvalues, respectively, are denoted.

If T is rank deficient, we can find vectors $w\in\mathbf{R}^{n_a+n_b}$ and $w_1\in\mathbf{R}^2$ such that $|w_1|=1$ and $T = w_1w^T$. We exclude the trivial case $T=0$ from consideration and assume $w\neq 0$. Then

$$\varepsilon^{-1}\sup_{z\in U_\varepsilon(\omega)}|T(\theta-\hat{\theta})| = \left(\left(w^T\left(\frac{N}{\chi^2_{n_a+n_b}(\alpha)}\bar{M}\right)^{-1}w\right)^{-1}\right)^{-1/2}$$
$$= \sqrt{\frac{\chi^2_{n_a+n_b}(\alpha)}{N}}(w^T\bar{M}^{-1}w)^{1/2}.$$

But we have anyway $w^T\bar{M}^{-1}w = \lambda_{\max}(T\bar{M}^{-1}T^T)$.

Hence in either case we obtain

$$\mathcal{J}_2 = \sqrt{\frac{\chi^2_{n_a+n_b}(\alpha)}{N}}\sup_{\omega\in[0,\pi]}\frac{(\lambda_{\max}(T(\omega)\bar{M}^{-1}T(\omega)^T))^{1/2}}{1+|G(e^{j\omega},\hat{\theta})|^2}. \tag{4.3}$$

It is well known that the inverse P^{-1} of a symmetric positive definite matrix P and the maximal eigenvalue $\lambda_{\max}(Q)$ of a symmetric positive semidefinite matrix Q are convex functions with respect to P or Q, respectively (see, e.g., [8]). Hence $\lambda_{\max}(T\bar{M}^{-1}T^T)$ is convex with respect to $\bar{M}$ for fixed ω. Since the operation of taking

the maximum over a family of functions preserves convexity, we have that $\mathcal{J}_2^2$ is a convex function with respect to $\bar{M}$. By strict monotonicity of the square root, this yields quasiconvexity of $\mathcal{J}_2$. ∎

In the preceding two sections we have shown that Problems 1 and 2 stated in section 2 are quasiconvex. In the next section we will provide the necessary tools that allow the user to apply standard convex algorithms to solve these problems numerically.

5. Cutting planes. Most methods in convex analysis are based on the notion of a cutting plane (see, e.g., [5]). Suppose $S \subset \mathbf{R}^m$ is a convex set and $f : S \to \mathbf{R}$ is a quasiconvex function defined on S.

DEFINITION 5.1. *A cutting plane to f at a point $x^{(0)} \in S$ is defined by a nonzero vector $g \in \mathbf{R}^m$ such that $f(x^{(0)}) \leq f(x)$ for any $x \in S$ satisfying the inequality $g^T(x - x^{(0)}) \geq 0$.*

Thus the global minimum of f on S lies in the half-space $\{x \,|\, g^T(x - x^{(0)}) \leq 0\}$. By definition of quasiconvexity, a cutting plane always exists.

In this section we will compute cutting planes for cost functions $\mathcal{J}_1, \mathcal{J}_2$ at an arbitrary point $x^{(0)} \in \mathcal{X}_c$.

Let $\bar{M}^{(0)}$ be the average information matrix corresponding to $x^{(0)}$. By Corollary 3.8 the matrix $\bar{M}^{(0)}$ is positive definite.

We shall now compute a cutting plane for $\mathcal{J}_1 = \max_{\omega \in [0,\pi]} \kappa_{WC}(G(e^{j\omega}, \hat{\theta}), \mathcal{D})$. Denote by $\omega^{(0)}$ the frequency where the worst-case chordal distance κ_{WC} attains its maximum. The value of $\omega^{(0)}$ can be found, e.g., by a grid search with subsequent refinement using a denser grid in the vicinity of the maximum. It is easily seen that a cutting plane to the function $\kappa_{WC}(G(e^{j\omega^{(0)}}, \hat{\theta}), \mathcal{D})$ or its square will also be a cutting plane to $\mathcal{J}_1$. In what follows we will assume that ω is equal to $\omega^{(0)}$ and omit it as an argument.

Thus our goal is to find a cutting plane for the optimum value γ_{opt} of GEVP (4.1), (4.2), considered as a function of x. Note that F_0, F_1 are independent of x, while R depends on x via $\bar{M}$. By (3.6), we can represent R as $R(x) = R_0 + \sum_{i=1}^n x_i R_i$ with

$$R_0 = \begin{pmatrix} I_{n_a+n_b} \\ -\hat{\theta}^T \end{pmatrix} \bar{M}_0 \begin{pmatrix} I_{n_a+n_b} & -\hat{\theta} \end{pmatrix} - \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\chi^2_{n_a+n_b}(\alpha)}{N} \end{pmatrix},$$

$$R_i = \begin{pmatrix} I_{n_a+n_b} \\ -\hat{\theta}^T \end{pmatrix} \bar{M}_i \begin{pmatrix} I_{n_a+n_b} & -\hat{\theta} \end{pmatrix}.$$

Let $\gamma_{opt}^{(0)}, \tau_{opt}^{(0)}$ be the optimal values for γ, τ in GEVP (4.1), (4.2) at $x = x^{(0)}$. Then the matrix $F_0 + \gamma_{opt}^{(0)} F_1 + \tau_{opt}^{(0)} R$ is both singular and positive semidefinite. Let V^0 be the nullspace of this matrix.

PROPOSITION 5.2. *If $\tau_{opt}^{(0)} > 0$, then there exists a unit length vector $v \in V^0$ such that $v^T R v = 0$. If $\tau_{opt}^{(0)} = 0$, then there exists a unit length vector $v \in V^0$ such that $v^T R v \leq 0$. In either case the vector $g \in \mathbf{R}^n$ given componentwise by $g_i = -v^T R_i v$, if it is nonzero, defines a cutting plane for the function $\mathcal{J}_1$. If g is zero, $\mathcal{J}_1$ achieves a minimum at $x^{(0)}$.*

The proof of this proposition can be found in the appendix.

Let us now compute a cutting plane for cost function $\mathcal{J}_2$, which is given by (4.3). Denote by $\omega^{(0)}$ the frequency at which the function $\frac{\lambda_{\max}(T(\omega)\bar{M}^{-1}T(\omega)^T)}{(1+|G(e^{j\omega},\hat{\theta})|^2)^2}$ attains its

maximum. Let $v \in \mathbf{R}^2$ be a unit length eigenvector to the maximal eigenvalue of the matrix $T(\omega^{(0)})\bar{M}^{-1}T(\omega^{(0)})^T$.

PROPOSITION 5.3. *Let $g \in \mathbf{R}^n$ be defined componentwise by $g_i = -v^T T(\omega^{(0)}) \cdot \bar{M}^{-1}\bar{M}_i\bar{M}^{-1}T(\omega^{(0)})^T v$. Then g defines a cutting plane for the cost function $\mathcal{J}_2$ at $x^{(0)}$, if $g \neq 0$, and $\mathcal{J}_2$ attains a minimum at $x^{(0)}$, if $g = 0$.*

Proof. Consider

$$f(x) = \sqrt{\frac{\chi^2_{n_a+n_b}(\alpha)}{N}} \frac{\left(v^T T(\omega^{(0)}) \left(\bar{M}_0 + \sum_{i=1}^n x_i \bar{M}_i\right)^{-1} T(\omega^{(0)})^T v\right)^{1/2}}{1 + |G(e^{j\omega^{(0)}}, \hat{\theta})|^2}.$$

By definition we have $f(x^{(0)}) = \mathcal{J}_2(x^{(0)})$, but $f(x) \leq \mathcal{J}_2(x)$ for any $x \in \mathcal{X}_c$.

Let $x \in \mathcal{X}_c$ be a point such that $g^T(x - x^{(0)}) \geq 0$. We shall show that $f(x^{(0)}) \leq f(x)$, which would imply $\mathcal{J}_2(x^{(0)}) \leq \mathcal{J}_2(x)$. This is equivalent to $\tilde{f}(x^{(0)}) \leq \tilde{f}(x)$, where $\tilde{f}$ is defined by

$$\begin{aligned}
\tilde{f}(x) &= \frac{N}{\chi^2_{n_a+n_b}(\alpha)}(1 + |G(e^{j\omega^{(0)}}, \hat{\theta})|^2)^2 f^2(x) \\
&= v^T T(\omega^{(0)}) \left(\bar{M}_0 + \sum_{i=1}^n x_i \bar{M}_i\right)^{-1} T(\omega^{(0)})^T v \\
&= \operatorname{tr}\left(T(\omega^{(0)})^T v v^T T(\omega^{(0)}) \left(\bar{M}_0 + \sum_{i=1}^n x_i \bar{M}_i\right)^{-1}\right).
\end{aligned}$$

In other words, we have to show that g defines a cutting plane for $\tilde{f}$. It is well known (see, e.g., [30, p. 39]) that $\tilde{f}$, being of the form $\operatorname{tr} W\bar{M}^{-1}$ with $W \geq 0$, is a smooth convex function on $\mathcal{X}_c$. Hence a cutting plane to $\tilde{f}$ is defined by its gradient, which is identical to g.

If $g = 0$, then $\tilde{f}$ attains a minimum at $x^{(0)}$. Hence f attains a minimum at $x^{(0)}$, which yields $\mathcal{J}_2(x^{(0)}) \leq \mathcal{J}_2(x)$ for any $x \in \mathcal{X}_c$. This concludes the proof. ∎

The results of sections 3 to 5, i.e., the LMI description of the feasible set and the knowledge of cutting planes, allow the user to apply a whole range of convex optimization methods for solving Problems 1 and 2. For a description of different methods, see, e.g., [5], [22].

6. Error assessment of the solution. Suppose we seek the minimum of a quasiconvex cost function $\mathcal{J}(x)$ on the closure of $\mathcal{X}_c$. Let us assume that with some method an approximation $x^{(0)} \in \mathcal{X}_c$ of the optimal value x^* was obtained together with an upper bound on the scalar product $g^T(x^{(0)} - x^*)$ (which is usually delivered by standard convex analysis methods), where g is a vector defining a cutting plane to $\mathcal{J}$ at $x^{(0)}$.

In this section we assess the quality of the approximation $x^{(0)}$, i.e., we derive a bound on the error $\mathcal{J}(x^{(0)}) - \mathcal{J}(x^*)$. The results presented can be used for designing termination criteria for iterative optimization algorithms, guaranteeing a prespecified level of accuracy.

PROPOSITION 6.1. *Let $x^{(0)} \in \mathcal{X}_c$ be a feasible point and $\omega^{(0)}$ a frequency at which the worst-case chordal distance $\kappa_{WC}(G(e^{j\omega}, \hat{\theta}), \mathcal{D}(x^{(0)}))$ attains its maximum. Suppose cost function $\mathcal{J}_1$ attains its minimum at x^*. Let vectors v and g be defined as in Proposition 5.2. If $v^T F_1 v > 0$, then the following bound on the error $\mathcal{J}_1(x^{(0)}) -$*

$\mathcal{J}_1(x^*)$ *holds:*

$$\mathcal{J}_1^2(x^{(0)}) - \mathcal{J}_1^2(x^*) \leq N \frac{\mathcal{J}_1^2(x^{(0)})}{\chi^2_{n_a+n_b}(\alpha) v^T F_1 v} (1 + |G(e^{j\omega^{(0)}}, \hat{\theta})|^2)^2 |1 + Z_D \hat{\theta}|^2 g^T (x^{(0)} - x^*),$$

where Z_D *is defined in* (2.3).

Note that the condition $v^T F_1 v > 0$ is satisfied whenever $\mathcal{J}_1(x^{(0)}) < 1$. This inequality holds at least in the vicinity of x^* if $\mathcal{J}_1$ is not identically 1 on $\mathcal{X}_c$.

Proof of Proposition 6.1. Denote by γ^*_{opt} the square of the worst-case chordal distance $\kappa_{WC}(G(e^{j\omega^{(0)}}, \hat{\theta}), \mathcal{D}(x^*))$ at frequency $\omega^{(0)}$ and at the point x^*. Let τ^*_{opt} be the corresponding optimal value of τ. Then we have $\mathcal{J}_1^2(x^*) \geq \gamma^*_{opt}$.

By definition we have at frequency $\omega^{(0)}$ the relations $v^T(F_0 + \gamma^{(0)}_{opt} F_1 + \tau^{(0)}_{opt} R)v = 0$, $(\tau - \tau^{(0)}_{opt}) v^T R(x^{(0)}) v \leq 0$ for any $\tau \geq 0$, and $v^T(R(x) - R(x^{(0)}))v \leq -g^T(x - x^{(0)})$ for any x. Hence

$$v^T(F_0 + \gamma^*_{opt} F_1 + \tau^*_{opt} R(x^*))v \leq (\gamma^*_{opt} - \gamma^{(0)}_{opt}) v^T F_1 v - \tau^*_{opt} g^T (x^* - x^{(0)}).$$

Since the left-hand side of this inequality is nonnegative, we obtain

$$\mathcal{J}_1^2(x^{(0)}) - \mathcal{J}_1^2(x^*) \leq \gamma^{(0)}_{opt} - \gamma^*_{opt} \leq \frac{\tau^*_{opt}}{v^T F_1 v} g^T (x^{(0)} - x^*).$$

Let us now derive a bound on τ^*_{opt}. We have $(v')^T(F_0 + \gamma^*_{opt} F_1 + \tau^*_{opt} R)v' \geq 0$ for any vector $v' \in \mathbf{R}^{n_a+n_b+1}$. Choose $v' = (\hat{\theta}^T \ 1)^T$. By direct calculation one can show that $(v')^T F_0 v' = 0, (v')^T F_1 v' = (1 + |G|^2)^2 |1 + Z_D \hat{\theta}|^2, (v')^T R v' = -\frac{\chi^2_{n_a+n_b}(\alpha)}{N}$. Thus we have

$$\tau^*_{opt} \leq \frac{N}{\chi^2_{n_a+n_b}(\alpha)} \gamma^*_{opt} (1 + |G|^2)^2 |1 + Z_D \hat{\theta}|^2 \leq \frac{N}{\chi^2_{n_a+n_b}(\alpha)} \gamma^{(0)}_{opt} (1 + |G|^2)^2 |1 + Z_D \hat{\theta}|^2.$$

Combining the obtained inequalities, we complete the proof. ∎

PROPOSITION 6.2. *Let* $x^{(0)} \in \mathcal{X}_c$ *be a feasible point. Let* $\omega^{(0)}$ *be a frequency at which the quantity* $\frac{\lambda_{\max}(T(\omega)\bar{M}^{-1}(x^{(0)})T(\omega)^T)}{(1+|G(e^{j\omega}, \hat{\theta})|^2)^2}$ *attains its maximum. Suppose cost function* $\mathcal{J}_2$ *attains its minimum at* x^*. *Let* g *be defined as in Proposition* 5.3. *Then the following bound on the error* $\mathcal{J}_2(x^{(0)}) - \mathcal{J}_2(x^*)$ *holds:*

$$\mathcal{J}_2^2(x^{(0)}) - \mathcal{J}_2^2(x^*) \leq \frac{\chi^2_{n_a+n_b}(\alpha)}{N(1 + |G(e^{j\omega^{(0)}}, \hat{\theta})|^2)^2} g^T (x^{(0)} - x^*).$$

Proof. Recall that we defined two functions $f(x), \tilde{f}(x)$ in the proof of Proposition 5.3 and identified g as the gradient of $\tilde{f}$. Since $\tilde{f}$ is convex, we can bound it by its first order Taylor polynomial, i.e., $\tilde{f}(x^{(0)}) - \tilde{f}(x^*) \leq g^T(x^{(0)} - x^*)$. Therefore we have

$$\begin{aligned}\mathcal{J}_2^2(x^{(0)}) - \mathcal{J}_2^2(x^*) \leq f^2(x^{(0)}) - f^2(x^*) &= \frac{\chi^2_{n_a+n_b}(\alpha)}{N(1 + |G(e^{j\omega^{(0)}}, \hat{\theta})|^2)^2} (\tilde{f}(x^{(0)}) - \tilde{f}(x^*)) \\ &\leq \frac{\chi^2_{n_a+n_b}(\alpha)}{N(1 + |G(e^{j\omega^{(0)}}, \hat{\theta})|^2)^2} g^T (x^{(0)} - x^*). \qquad ∎\end{aligned}$$

The propositions proven in this section enable the user to tell whether a given solution $x^{(0)}$, delivered, e.g., by the current iteration step, satisfies the prespecified accuracy requirements. This information can be used, e.g., to decide whether further iterations are necessary.

7. Design of input signals. Let us now turn to the question of how to design an input signal from $x^{(0)}$. By Theorem 3.6, there exist moment points which can be realized only by discrete spectra. On the other hand, any moment point can be realized by a discrete spectrum. Therefore we propose here the following two-step procedure. First a discrete input power spectrum generating the moment point $x^{(0)}$ is computed, and then a multisine input with the desired spectrum is generated. This procedure in no way restricts the optimality of the solution.

We weaken the condition $x^{(0)} \in \mathcal{X}_c$ and suppose that $x^{(0)}$ is in the closure of $\mathcal{X}_c$. The point $x^{(0)}$ corresponds to a point $\tilde{x} = (\tilde{x}_0, \tilde{x}_1, \ldots, \tilde{x}_n)$ in moment space $\mathcal{M}^{(n+1)}$. Here $\tilde{x}_i$ equals the ith component of $x^{(0)}$ for $i = 1, \ldots, n$, and $\tilde{x}_0$ is given by (3.5).

Our goal will be to construct a realization of $\tilde{x}$. By Theorem 3.6, there exists a discrete realization with index not greater than $\frac{n+1}{2}$.

Denote by $\tilde{x}^s(\omega) \in \mathcal{M}^{(n+1)}$ the moment point induced by the design measure that satisfies constraint (3.4) and concentrates all power at the single frequency ω. Then the ith entry of $\tilde{x}^s(\omega)$ is given by $\frac{c}{\lambda_0 |A(\omega)|^2} \cos(i\omega)$. Since (3.4) defines an affine section of the convex cone $\mathcal{M}^{(n+1)}$ and $\tilde{x}$ satisfies (3.4), $\tilde{x}$ is a convex combination of points on the curve $\{\tilde{x}^s(\omega) \,|\, \omega \in [0, \pi]\}$.

If $\tilde{x}$ equals $\tilde{x}^s(\pi)$, then we have already found a realization of index $\frac{1}{2}$. In this case this is the only possible realization.

Suppose now that $\tilde{x}$ is not equal to $\tilde{x}^s(\pi)$. To construct a realization of $\tilde{x}$, we will exploit an idea that is used to prove Theorem 3.6 (see, e.g., [14]).

Consider the line going through the points $\tilde{x}$ and $\tilde{x}^s(\pi)$. This line has an interval in common with the convex set $\mathcal{M}^{(n+1)}$. This interval is finite, because it lies on the section defined by (3.4), and nondegenerated, because it contains two different points $\tilde{x}$ and $\tilde{x}^s(\pi)$. By Theorem 3.6 part (i), $\tilde{x}^s(\pi)$ is one of the endpoints of this interval. Denote the other endpoint by $\tilde{x}^{bd}$. The computation of $\tilde{x}^{bd}$ from $\tilde{x}$ and $\tilde{x}^s(\pi)$ can be reduced to a standard GEVP using LMI description (3.3) of the set $\mathcal{M}^{(n+1)}$. For treatment of this type of problem, see, e.g., [5].

Thus $\tilde{x}$ is a linear convex combination of the points $\tilde{x}^s(\pi)$ and $\tilde{x}^{bd}$. Any realization of $\tilde{x}^{bd}$ will deliver us a realization of $\tilde{x}$. Note that $\tilde{x}^{bd}$ lies on the boundary of $\mathcal{M}^{(n+1)}$. By Theorem 3.6 part (i) it has only one realization, which is of index less than $\frac{n+1}{2}$. Hence the realization of $\tilde{x}$ that we obtain with the described procedure will have an index not exceeding $\frac{n+1}{2}$. If $\tilde{x}$ lies in the interior of $\mathcal{M}^{(n+1)}$, then this realization contains the frequency π and is therefore the upper principal realization of $\tilde{x}$.

We shall now construct the realization of $\tilde{x}^{bd}$. Denote the frequencies which are involved in this realization by ω_i, $i = 1, \ldots, k$. Then the point $\tilde{x}^{bd}$ is a nondegenerated linear convex combination of $\tilde{x}^s(\omega_1), \ldots, \tilde{x}^s(\omega_k)$. We can write $\tilde{x}^{bd} = \sum_{i=1}^{k} \lambda_i \tilde{x}^s(\omega_i)$, where $\lambda_i > 0$ and $\sum_{i=1}^{k} \lambda_i = 1$.

Since $\tilde{x}^{bd}$ lies on the boundary of $\mathcal{M}^{(n+1)}$, there exists a supporting hyperplane E at $\tilde{x}^{bd}$. Note that E is a linear subspace, because $\mathcal{M}^{(n+1)}$ is a convex cone. The construction of a supporting hyperplane proceeding from LMI description (3.3) of $\mathcal{M}^{(n+1)}$ is a standard procedure and is described, e.g., in [5].

LEMMA 7.1. *The points $\tilde{x}^s(\omega_1), \ldots, \tilde{x}^s(\omega_k)$ lie in E.*

Proof. Denote by n_E the normal vector to E that points toward $\mathcal{M}^{(n+1)}$ and by L_E the linear functional $x \mapsto \langle n_E, x \rangle$ defined by n_E. For any $\omega \in [0, \pi]$ we have $L_E(\tilde{x}^s(\omega)) \geq 0$. On the other hand, $L_E(\tilde{x}^{bd}) = \sum_{i=1}^{k} \lambda_i L_E(\tilde{x}^s(\omega_i)) = 0$, because $\tilde{x}^{bd}$ lies in E. Hence for all i we have $L_E(\tilde{x}^s(\omega_i)) = 0$, i.e., $\tilde{x}^s(\omega_i) \in E$. ∎

LEMMA 7.2. *There exist maximally $\frac{n}{2} + 1$ frequencies such that the corresponding points $\tilde{x}^s(\omega)$ lie in E.*

Proof. Consider $p_E : [0, \pi] \to \mathbf{R}$ defined by $p_E(\omega) = L_E(\tilde{x}^s(\omega))\frac{\lambda_0 |A(\omega)|^2}{c}$. By definition, p_E is a trigonometric polynomial. Since $L_E(\tilde{x}^s(\omega))$ is nonnegative for all ω, p_E is too. Now we can apply a classical result from Tchebycheff system theory which states that p_E can have at most $\frac{n}{2} + 1$ zeros. But the zeros of p_E lie exactly at those frequencies whose corresponding points $\tilde{x}^s(\omega)$ lie in E. □

Now we are able to obtain a finite set of frequencies that is guaranteed to contain $\omega_1, \ldots, \omega_k$. Namely, we have to find the zeros, which are at the same time local minima, of the trigonometric polynomial $p_E(\omega)$.

Once we have found a set of frequencies $\omega_1, \ldots, \omega_k, \omega_{k+1}, \ldots, \omega_{k+k'}$ such that the convex hull of the points $\tilde{x}^s(\omega_1), \ldots, \tilde{x}^s(\omega_{k+k'})$ contains $\tilde{x}^{bd}$, it is a standard linear quadratic (LQ) programming problem to find the weights associated with the different frequencies. Namely, the weights λ_j minimize the squared distance $|\tilde{x}^{bd} - \sum_{j=1}^{k+k'} \lambda_j \tilde{x}^s(\omega_j)|^2$, which is a quadratic polynomial in the λ_j. Note that the number of frequencies involved is not greater than $\frac{n}{2} + 1$ and hence not greater than n. Therefore the points $\tilde{x}^s(\omega_1), \ldots, \tilde{x}^s(\omega_{k+k'})$ are linearly independent and the minimized polynomial has a positive definite quadratic part. An efficient algorithm for solving this type of problems is, e.g., the Beale algorithm (see [1]).

Suppose that a discrete realization of $\tilde{x}$ with frequencies $\omega_1, \ldots, \omega_m$ and associated weights $\lambda_1, \ldots, \lambda_m$ is available. Then the multisine input $u(t) = \sum_{i=1}^{m} \alpha_i \sin(t\omega_i + \phi_i)$ with $\alpha_i = \sqrt{2c\lambda_i}$, ϕ_i arbitrary, if $\omega_i \neq 0, \pi$, and $\alpha_i = \sqrt{c\lambda_i}$, $\phi_i = \pm\frac{\pi}{2}$, if $\omega_i \in \{0, \pi\}$, has the desired input power spectrum (see, e.g., [30]).

Often it is also possible to obtain the desired power spectrum by using binary signals (see [30, p. 29] and references cited therein).

8. Simulation results. Consider the true system $y = G_0 u + H_0 e = \frac{B(z)}{A(z)} u + \frac{1}{A(z)} e$ with $G_0 = \frac{B(z)}{A(z)} = \frac{0.1047z^{-1} + 0.0872z^{-2}}{1 - 1.5578z^{-1} + 0.5769z^{-2}}$. Here u is the input, subject to the energy constraint $\bar{E}u^2(t) = 1$, and e is white Gaussian noise with variance 0.1.

The system is to be identified within an ARX model structure of order two. The number of data points to be collected is $N = 1000$. The aim is to minimize the worst-case ν-gap of the uncertainty region around the identified model corresponding to a confidence level of $\alpha = 0.95$.

In a Monte-Carlo simulation, 500 runs were performed. Each run consisted of five identification experiments: one preliminary experiment and four mutually independent secondary experiments based on this preliminary experiment, corresponding to the four different cost functions $\mathcal{J}_1$, $\mathcal{J}_2$, D-optimality, and E-optimality.

In the preliminary experiment, the input was chosen to be white Gaussian noise with variance 1. The parameter vector and noise variance identified in the preliminary experiment were used as a priori estimates of the true parameter vector and the true noise variance for designing the input power spectrum for the series of second experiments. In two of the second experiments, the input power spectrum minimized the cost functions $\mathcal{J}_1, \mathcal{J}_2$, respectively. The actual input sequence was a multisine having the evaluated optimal power spectrum. For comparison, two other second experiments with D-optimal and E-optimal input power spectra were performed. After each identification experiment the worst-case ν-gap of the identified uncertainty region was recorded.

The noise realizations for the five experiments within one run and for different runs were different, as well as the input realizations for the preliminary experiments of the different runs.

In Figure 8.1 the worst-case ν-gap obtained from the preliminary experiment with

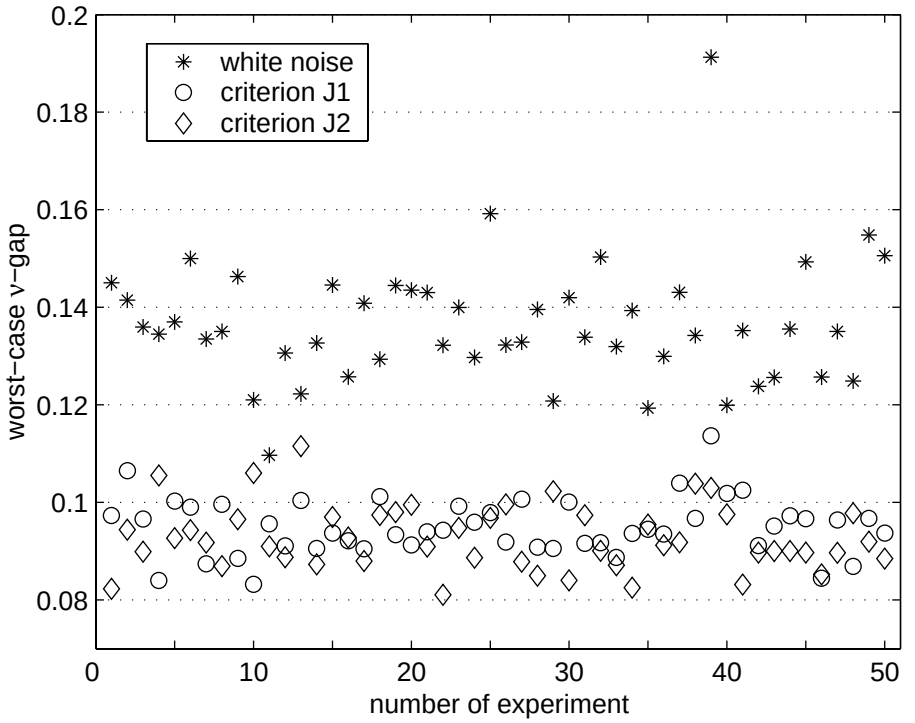


FIG. 8.1. *Identification with white and subsequently estimated optimal input.*

white noise input, as well as from the experiments with inputs optimized with respect to $\mathcal{J}_1$ and $\mathcal{J}_2$, respectively, is shown for the first 50 simulation runs. The mean over 500 runs of the worst-case ν-gap resulting from the preliminary experiments equals 0.1345. The means of the worst-case ν-gap resulting from the experiments with multisine input optimized with respect to criteria $\mathcal{J}_1, \mathcal{J}_2$ are 0.0937 and 0.0927, respectively. The difference between them is statistically significant (2×1.64 standard deviations). The means of the worst-case ν-gap resulting from the experiments with D- and E-optimal multisine input are equal to 0.1434 and 0.1055, respectively.

It is evident that using inputs optimized with respect to criteria $\mathcal{J}_1, \mathcal{J}_2$ gives better results than using white noise input or input optimized with respect to the classical D- and E-optimality criteria. Note also that the inputs optimized with respect to the cost function $\mathcal{J}_2$, which is a first order approximation of the exact cost function $\mathcal{J}_1$, give better results than $\mathcal{J}_1$, despite the fact that the plotted quantity is in fact $\mathcal{J}_1$. As mentioned already in section 2, this tendency was observed also in simulations with other systems. The reason is that the optimum of the input power spectrum with respect to $\mathcal{J}_2$ is less dependent on the preliminary estimate $\bar{\theta}$ of the true parameter vector than the optimum with respect to $\mathcal{J}_1$. Given the lower complexity of $\mathcal{J}_2$ and hence the lower computational effort in comparison with $\mathcal{J}_1$, it is preferable to use primarily the former.

9. Conclusions. Let us summarize the results obtained in the present paper. We have to design an input sequence for an identification experiment that makes the worst-case ν-gap between the identified model and the uncertainty region around it as small as possible. The design takes place via power spectrum optimization. Two

nonstandard cost criteria $\mathcal{J}_1$ and $\mathcal{J}_2$ are defined, which reflect the optimization task with different accuracy. $\mathcal{J}_1$ is the exact worst-case ν-gap one would want to minimize, while $\mathcal{J}_2$ is an approximation of $\mathcal{J}_1$. These functions fulfill the natural conditions of monotonicity and quasiconvexity with respect to the power spectrum.

It was shown that optimization of the input power spectrum with respect to these cost criteria can be reduced to a standard convex optimization problem involving LMI constraints. In Propositions 5.2 and 5.3 we demonstrate how to construct cutting planes to the cost functions $\mathcal{J}_1, \mathcal{J}_2$, which is essential for applying standard numerical methods such as the ellipsoid algorithm. In Propositions 6.1 and 6.2 we derive bounds on the difference between the actually achieved and the optimal value of the cost functions, which allows us to estimate the quality of the optimization result and to design stopping criteria for iterative search algorithms. We have also briefly touched on the problem of designing an input sequence with a prespecified power spectrum.

Simulations show clearly the superiority of the proposed cost functions over classical design criteria. They also suggest using cost function $\mathcal{J}_2$ rather than $\mathcal{J}_1$, due to both lower computational effort and higher performance.

Appendix A. Proof of Proposition 5.2.

LEMMA A.1. *Let γ_{opt}, τ_{opt} be the optimal values of γ, τ in GEVP* (4.1), (4.2). *Then the following conditions hold:*

(i) *The matrix F_0 is negative semidefinite.*
(ii) *The nullspace of F_1 is a subset of the nullspace of F_0.*
(iii) *The matrix $F_0 + F_1$ is positive semidefinite.*
(iv) *The nullspace of F_1 is a strict subset of the nullspace of $F_0 + F_1$.*
(v) *$\tau_{opt} > 0$ if and only if the restriction of R on the nullspace of $F_0 + F_1$ is strictly positive definite.*
(vi) *$\gamma_{opt} = 1$ if and only if $\tau_{opt} = 0$.*

Proof. (i) follows from the representation $F_0 = -VWW^TV^T$, where W is a 4×2-matrix given by

$$W = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \mathrm{Im}G(\hat{\theta}) & -\mathrm{Re}G(\hat{\theta}) \\ -\mathrm{Re}G(\hat{\theta}) & -\mathrm{Im}G(\hat{\theta}) \end{pmatrix}.$$

The nullspace of F_1 is given by the kernel of V^T. The latter is contained in the kernel of F_0, which yields (ii).

(iii) follows from the representation

$$F_0 + F_1 = V \begin{pmatrix} |G(\hat{\theta})|^2 & 0 & -\mathrm{Im}G(\hat{\theta}) & \mathrm{Re}G(\hat{\theta}) \\ 0 & |G(\hat{\theta})|^2 & \mathrm{Re}G(\hat{\theta}) & \mathrm{Im}G(\hat{\theta}) \\ -\mathrm{Im}G(\hat{\theta}) & \mathrm{Re}G(\hat{\theta}) & 1 & 0 \\ \mathrm{Re}G(\hat{\theta}) & \mathrm{Im}G(\hat{\theta}) & 0 & 1 \end{pmatrix} V^T = VW_\perp W_\perp^T V^T,$$

where $W_\perp$ is a 4×2-matrix given by

$$W_\perp = \begin{pmatrix} |G(\hat{\theta})| & 0 \\ 0 & |G(\hat{\theta})| \\ -\sin \arg G(\hat{\theta}) & \cos \arg G(\hat{\theta}) \\ \cos \arg G(\hat{\theta}) & \sin \arg G(\hat{\theta}) \end{pmatrix}.$$

Here $\arg G(\hat{\theta})$ is an arbitrary number if $G(\hat{\theta}) = 0$. Note that $W^T W = (1 + |G(\hat{\theta})|^2) I_2$, $W^T W_\perp = 0$, and $W_\perp^T W_\perp = (1 + |G(\hat{\theta})|^2) I_2$.

By (ii) the nullspace of F_1 is a subset of the nullspace of $F_0 + F_1$. We shall now show that the vector $v = (\zeta, 0, \dots, 0, 1, 0, \dots, 0, -\cos n_k\omega \operatorname{Re} G(\hat{\theta}) + \sin n_k\omega \operatorname{Re} G(\hat{\theta}) - \zeta \cos\omega)^T \in \mathbf{R}^{n_a+n_b+1}$, where $\zeta = -\cot n_k\omega \operatorname{Im} G(\hat{\theta}) - \operatorname{Re} G(\hat{\theta})$ if $\omega \in (0, \pi)$ and ζ arbitrary otherwise, is not contained in the kernel of V^T but is contained in the kernel of $W_\perp^T V^T$. The "1" in v is situated at position $n_a + 1$. Indeed, by (2.3) we obtain

$$V^T v = \begin{pmatrix} \cos n_k\omega \\ -\sin n_k\omega \\ -\zeta \sin\omega \\ -\cos n_k\omega \operatorname{Re} G(\hat{\theta}) + \sin n_k\omega \operatorname{Im} G(\hat{\theta}) \end{pmatrix}$$

$$= \cos n_k\omega \begin{pmatrix} 1 \\ 0 \\ \operatorname{Im} G(\hat{\theta}) \\ -\operatorname{Re} G(\hat{\theta}) \end{pmatrix} - \sin n_k\omega \begin{pmatrix} 0 \\ 1 \\ -\operatorname{Re} G(\hat{\theta}) \\ \operatorname{Im} G(\hat{\theta}) \end{pmatrix} \neq 0,$$

because $\cos n_k\omega$, $-\sin n_k\omega$ cannot both vanish. In case $\omega \in \{0, \pi\}$ the equality holds by $\operatorname{Im} G(\hat{\theta}) = 0$. On the other hand, we have $W_\perp^T V^T v = W_\perp^T W (\cos n_k\omega, -\sin n_k\omega)^T = 0$. This concludes the proof of (iv).

Let us prove (v) and (vi). Denote the nullspace of $F_0 + F_1$ by $\bar{V}^0$ and its orthogonal complement by $\bar{V}^\perp$. By definition there exists a positive number β_1 such that for any $v^\perp \in \bar{V}^\perp$ we have $(v^\perp)^T (F_0 + F_1) v^\perp \geq \beta_1 |v^\perp|^2$.

Suppose the restriction of R on $\bar{V}^0$ is strictly positive definite. Then there exists a positive number β_2 such that for any $v^0 \in \bar{V}^0$ we have $(v^0)^T R v^0 \geq \beta_2 |v^0|^2$. Let $v = v^0 + v^\perp$ be an arbitrary vector with v^0, $v^\perp$ being its orthogonal projections on $\bar{V}^0$, $\bar{V}^\perp$, respectively. Let $\tau > 0$ be a positive number. Then we have

$$v^T (F_0 + F_1 + \tau R) v = (v^\perp)^T (F_0 + F_1) v^\perp + \tau((v^\perp)^T R v^\perp + 2(v^0)^T R v^\perp + (v^0)^T R v^0)$$

$$\geq \beta_1 |v^\perp|^2 + \tau(\lambda_{\min}(R)|v^\perp|^2 - 2\min\{\lambda_{\min}(R), -\lambda_{\max}(R)\}|v^\perp||v^0| + \beta_2 |v^0|^2)$$

$$= \begin{pmatrix} |v^\perp| \\ |v^0| \end{pmatrix}^T \begin{pmatrix} \beta_1 + \tau\lambda_{\min}(R) & -\tau \min\{\lambda_{\min}(R), -\lambda_{\max}(R)\} \\ -\tau \min\{\lambda_{\min}(R), -\lambda_{\max}(R)\} & \tau\beta_2 \end{pmatrix} \begin{pmatrix} |v^\perp| \\ |v^0| \end{pmatrix}.$$

It is easily seen that the 2×2-matrix in the middle is positive definite if τ is small enough. Therefore there exists $\tau > 0$ such that the matrix $F_0 + F_1 + \tau R$ is strictly positive definite, while the matrix $F_0 + F_1$ is not. Thus in this case we have $\tau_{opt} \neq 0$ and $\gamma_{opt} < 1$.

Now suppose the restriction of R on $\bar{V}^0$ is not strictly positive definite. Since $\bar{M}$ is strictly positive definite, it follows from expression (4.2) that R has $n_a + n_b$ positive eigenvalues and one negative eigenvalue. Thus it can be represented as a difference $R = R_+ - R_-$, where R_+, R_- are positive semidefinite matrices of rank $n_a + n_b, 1$, respectively, and the linear hulls $\bar{V}_+, \bar{V}_-$ of their columns are orthogonal to each other. The whole space $\mathbf{R}^{n_a+n_b+1}$ splits into a direct sum $\bar{V}_+ \oplus \bar{V}_-$.

Let $v^0 \in \bar{V}^0$ be a nonzero vector such that $(v^0)^T R v^0 \leq 0$. The vector v^0 can be represented as a sum $v^0 = v_+ + v_-$, where $v_+ \in \bar{V}_+$, $v_- \in \bar{V}_-$. Since $(v^0)^T R v^0 =$

$(v_+)^T R_+ v_+ - (v_-)^T R_- v_- \leq 0$, the assumption $v_- = 0$ would imply $v_+ = 0$, which contradicts $v^0 \neq 0$. Hence $v_- \neq 0$. We can represent v_- as a sum $v_- = v_-^0 + v_-^\perp$, where $v_-^0 \in \bar{V}^0, v_-^\perp \in \bar{V}^\perp$. Let $\varepsilon > 0$ be a positive number and consider the vector $v = v^0 + \varepsilon v_-$. We have $v = v_+ + (1+\varepsilon)v_-$. Hence $v^T R v = (v_+)^T R_+ v_+ - (1+\varepsilon)^2 (v_-)^T R_- v_- = (v^0)^T R v^0 - (2\varepsilon + \varepsilon^2)(v_-)^T R_- v_-$. On the other hand, $v = (v^0 + \varepsilon v_-^0) + \varepsilon v_-^\perp$. Since $v^0 + \varepsilon v_-^0 \in \bar{V}^0$, this yields $v^T(F_0 + F_1)v = \varepsilon^2 (v_-^\perp)^T (F_0 + F_1) v_-^\perp$. We obtain

$$v^T(F_0 + F_1 + \tau R)v \leq \varepsilon^2 (v_-^\perp)^T (F_0 + F_1) v_-^\perp - \tau(2\varepsilon + \varepsilon^2)(v_-)^T R_- v_-.$$

Note that $(v_-)^T R_- v_-$ is strictly positive. Hence for any prespecified $\tau > 0$ we can choose a small $\varepsilon > 0$ such that $v^T(F_0 + F_1 + \tau R)v < 0$. Thus for any positive τ the matrix $F_0 + F_1 + \tau R$ is not positive semidefinite, while $F_0 + F_1$ is. This implies $\tau_{opt} = 0$. $\gamma_{opt} = 1$ now follows from (iv).

The proof of the lemma is complete. □

Proof of Proposition 5.2. Denote the orthogonal complement of V^0 by $V^\perp$. Then the restriction on $V^\perp$ of the quadratic form defined by the matrix $F_0 + \gamma_{opt}^{(0)} F_1 + \tau_{opt}^{(0)} R$ is strictly positive definite. Hence there exists a positive number β_1 such that for any vector $v^\perp \in V^\perp$ we have $(v^\perp)^T (F_0 + \gamma_{opt}^{(0)} F_1 + \tau_{opt}^{(0)} R) v^\perp \geq \beta_1 |v^\perp|^2$.

Suppose the restriction on V^0 of the quadratic form defined by the matrix R is strictly positive definite. Then there exists a positive number β_2 such that for any vector $v^0 \in V^0$ we have $(v^0)^T R v^0 \geq \beta_2 |v^0|^2$.

Let $v = v^0 + v^\perp$ be an arbitrary vector, where $v^0 \in V^0$ and $v^\perp \in V^\perp$ are its orthogonal projections on the subspaces V^0 and $V^\perp$, respectively. Let $\varepsilon > 0$ be a positive number. Then we have

$$\begin{aligned}
&v^T(F_0 + \gamma_{opt}^{(0)} F_1 + (\tau_{opt}^{(0)} + \varepsilon) R)v \\
&= (v^\perp)^T (F_0 + \gamma_{opt}^{(0)} F_1 + \tau_{opt}^{(0)} R) v^\perp + \varepsilon((v^\perp)^T R v^\perp + 2(v^0)^T R v^\perp + (v^0)^T R v^0) \\
&\geq \beta_1 |v^\perp|^2 + \varepsilon(\lambda_{\min}(R)|v^\perp|^2 + 2\min\{\lambda_{\min}(R), -\lambda_{\max}(R)\}|v^\perp||v^0| + \beta_2 |v^0|^2) \\
&= \begin{pmatrix} |v^\perp| \\ |v^0| \end{pmatrix}^T \begin{pmatrix} \beta_1 + \varepsilon\lambda_{\min}(R) & \varepsilon \min\{\lambda_{\min}(R), -\lambda_{\max}(R)\} \\ \varepsilon \min\{\lambda_{\min}(R), -\lambda_{\max}(R)\} & \varepsilon\beta_2 \end{pmatrix} \begin{pmatrix} |v^\perp| \\ |v^0| \end{pmatrix}.
\end{aligned}$$

It is easily seen that the 2×2-matrix in the middle is positive definite if ε is small enough. This implies that there exists a number $\tau > \tau_{opt}^{(0)}$ such that the matrix $F_0 + \gamma_{opt}^{(0)} F_1 + \tau R$ is strictly positive definite. This contradicts the optimality of $\gamma_{opt}^{(0)}$.

In a similar way it is shown that if the restriction on V^0 of the quadratic form R is strictly negative definite, then there exists a number $\varepsilon > 0$ such that for any $\tau \in [\tau_{opt}^{(0)} - \varepsilon, \tau_{opt}^{(0)})$ the matrix $F_0 + \gamma_{opt}^{(0)} F_1 + \tau R$ is strictly positive definite.

Thus the restriction on V^0 of the quadratic form R is neither strictly positive nor strictly negative definite if $\tau_{opt}^{(0)} > 0$ and it is negative semidefinite if $\tau_{opt}^{(0)} = 0$. This proves the first part of the proposition.

Now let $v \in V^0$ be a unit length vector satisfying the conditions of Proposition 5.2. Let $g \in \mathbf{R}^n$ be given componentwise by $g_i = -v^T R_i v$. Let $x \in \mathcal{X}_c$ be a vector satisfying the inequality $g^T(x - x^{(0)}) \geq 0$. Let τ be a nonnegative number and let γ be strictly less than $\gamma_{opt}^{(0)}$.

By assumption we have $v^T(F_0 + \gamma_{opt}^{(0)} F_1 + \tau_{opt}^{(0)} R(x^{(0)}))v = 0$. We obtain

$$v^T(F_0 + \gamma F_1 + \tau R(x))v = v^T((\gamma - \gamma_{opt}^{(0)})F_1 + \tau(R(x) - R(x^{(0)})) + (\tau - \tau_{opt}^{(0)})R(x^{(0)}))v$$

$$= (\gamma - \gamma_{opt}^{(0)})v^T F_1 v + (-\tau g^T(x - x^{(0)})) + (\tau - \tau_{opt}^{(0)})v^T R(x^{(0)})v \leq 0.$$

The last inequality follows from the fact that none of the three terms on the left-hand side exceeds zero. The first term is nonpositive because F_1 is positive semidefinite. The second term does not exceed zero by assumption on x. The third term is not greater than zero because by assumption on v we have $v^T R(x^{(0)})v \leq 0$ and the condition $\tau - \tau_{opt}^{(0)} < 0$ yields $\tau_{opt}^{(0)} > 0$ and hence $v^T R(x^{(0)})v = 0$. If the inequality is strict, then $F_0 + \gamma F_1 + \tau R(x)$ is not positive semidefinite.

Now assume that $v^T(F_0 + \gamma F_1 + \tau R(x))v = 0$. Then we have $v^T F_1 v = 0$ and v is an element of the nullspace of F_1. By Lemma A.1, part (iv), it is also an element of the nullspace of $F_0 + F_1$. Note that $v^T R(x^{(0)})v \leq 0$. By Lemma A.1, part (v), we then have $\tau_{opt}^{(0)} = 0$ and by part (vi) $\gamma_{opt}^{(0)} = 1$. Further we have either $\tau = \tau_{opt}^{(0)}$ or $v^T R(x^{(0)})v = 0$.

If $\tau = \tau_{opt}^{(0)} = 0$, then by Lemma A.1, parts (iii) and (iv), the matrix $F_0 + \gamma F_1 = F_0 + \gamma F_1 + \tau R(x)$ is not positive semidefinite.

If $\tau > 0$, then $v^T R(x^{(0)})v = 0$ and $v^T R(x)v = v^T(R(x) - R(x^{(0)}))v = -g^T(x - x^{(0)}) \leq 0$. Since v belongs to the nullspace of $F_0 + F_1$, by Lemma A.1, part (v) we have $\tau_{opt}(x) = 0$ and by part (vi) $\gamma_{opt}(x) = 1 > \gamma$. Hence the pair (γ, τ) is again not feasible for GEVP (4.1), (4.2) at x.

Thus in any case $\gamma_{opt}(x)$ is not less than $\gamma_{opt}^{(0)}$ and the vector g, if nonzero, defines a cutting plane for γ_{opt} and hence for $\mathcal{J}_1$.

If $g = 0$, however, then any x satisfies the relation $g^T(x - x^{(0)}) \geq 0$ and $\gamma_{opt}^{(0)}$ does not exceed γ_{opt} at any other point $x \in \mathcal{X}_c$. Hence we have $\mathcal{J}_1(x) \geq \kappa_{WC}(G(e^{j\omega^{(0)}}, \hat{\theta}), \mathcal{D}) = \sqrt{\gamma_{opt}(x)} \geq \sqrt{\gamma_{opt}^{(0)}} = \mathcal{J}_1(x^{(0)})$ and $\mathcal{J}_1$ attains a minimum at $x^{(0)}$.

This concludes the proof of the second part of Proposition 5.2. ∎

Acknowledgments. The authors would like to thank Yurii Nesterov for the very useful help he provided during the work on this paper.

REFERENCES

[1] J. Abadie, ed., *Nonlinear Programming*, North–Holland, Amsterdam, 1967, chap. 7.

[2] X. Bombois, B. Anderson, and M. Gevers, *Frequency domain image of a set of linearly parametrized transfer functions*, in Proceedings of the 6th European Control Conference, 2001, pp. 1416–1421.

[3] X. Bombois, M. Gevers, and G. Scorletti, *A measure of robust stability for a set of parametrized transfer functions*, IEEE Trans. Automat. Control, 45 (2000), pp. 2141–2145.

[4] X. Bombois, M. Gevers, G. Scorletti, and B. Anderson, *Robustness analysis tools for an uncertainty set obtained by prediction error identification*, Automatica, 37 (2001), pp. 1629–1636.

[5] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.

[6] B. L. Cooley, J. H. Lee, and S. P. Boyd, *Control-relevant experiment design: A plant-friendly, LMI-based approach*, in Proceedings of the 1998 American Control Conference, Vol. 2, IEEE Press, Piscataway, NJ, 1998, pp. 1240–1244.

[7] P. Date and G. Vinnicombe, *An algorithm for identification in the ν-gap metric*, in Proceedings of the 38th IEEE Conference of Decision and Control, IEEE Press, Piscataway, NJ, 1999, pp. 3230–3235.

[8] K. Fan, *On a theorem of Weyl concerning eigenvalues of linear transformations*, Proc. Natl. Acad. Sci. USA, 35 (1949), pp. 652–655.

[9] V. Fedorov, *Theory of Optimal Experiments*, Probab. Math. Statist. 12, Academic Press, New York-London, 1972.

[10] M. Gevers, X. Bombois, B. Codrons, G. Scorletti, and B. Anderson, *Model validation for control and controller validation in a prediction error framework*, in System Identification 2000 (SYSID 2000), Vol. 1, 2000, Pergamon Press, New York, pp. 319–324.

[11] G. Goodwin and R. Payne, *Design and characterization of optimal test signals for linear SISO parameter estimation*, in Preprints of 3rd IFAC Symposium, The Hague, North–Holland, Amsterdam, 1973, paper TT-1.

[12] G. Goodwin, M. Zarrop, and R. Payne, *Coupled design of test signal, sampling intervals and filters for system identification*, IEEE Trans. Automat. Control, 19 (1974), pp. 748–752.

[13] S. Karlin and L. Shapley, *Geometry of Moment Spaces*, Mem. Amer. Math. Soc. 12, AMS, Providence, RI, 1953.

[14] S. Karlin and W. Studden, *Tchebycheff Systems: With Applications in Analysis and Statistics*, Pure Appl. Math. 15, Interscience, New York, 1966.

[15] J. Kiefer, *General equivalence theory for optimum designs (approximate theory)*, Ann. Statist., 2 (1974), pp. 849–879.

[16] J. Kiefer and J. Wolfowitz, *Optimum designs in regression problems*, Ann. Math. Statist., 30 (1959), pp. 271–294.

[17] M. Krein, *The ideas of P.L. Čebyšev and A.A. Markov in the theory of limiting values of integrals and their further developments*, Amer. Math. Soc. Transl. Ser. 2, 12 (1951), pp. 1–122.

[18] K. Lindqvist and H. Hjalmarsson, *Identification for control: Adaptive input design using convex optimization*, in Proceedings of the 40th IEEE Conference on Decision and Control, Orlando, FL, 2001, pp. 4326–4331.

[19] L. Ljung, *System Identification: Theory for the User*, 2nd ed., Prentice Hall Information and System Sciences Series, Prentice–Hall, Upper Saddle River, NJ, 1999.

[20] R. Mehra, *Optimal input signals for parameter estimation in dynamic systems — survey and new results*, IEEE Trans. Automat. Control, 19 (1974), pp. 753–768.

[21] R. Mehra, *Optimal inputs for linear system identification*, IEEE Trans. Automat. Control, 19 (1974), pp. 192–200.

[22] Y. Nesterov and A. Nemirovskii, *Interior Point Polynomial Algorithms in Convex Programming*, SIAM Stud. Appl. Math. 13, SIAM, Philadelphia, 1994.

[23] R. Payne, *Optimal Experiment Design for Dynamic System Identification*, Ph.D. thesis, Imperial College, London, 1974.

[24] R. Payne and G. Goodwin, *Simplification of Frequency Domain Experiment Design for SISO Systems*, Publication 74/3, Dept. of Computing and Control, Imperial College, London, 1974.

[25] J. Schoukens, P. Guillaume, and R. Pintelon, *Perturbation signals for system identification*, in Prentice Hall International Series in Acoustics, Speech and Signal Processing, Prentice–Hall, Englewood Cliffs, NJ, 1993, ch. 3, pp. 126–160.

[26] A. van den Bos, *Selection of periodic test signals for estimation of linear system dynamics*, in Preprints of 3rd IFAC Symposium, The Hague, North–Holland, Amsterdam, 1973, paper TT-3.

[27] E. van den Eijnde and J. Schoukens, *On the design of optimal excitation signals*, in Proceedings of the 9th IFAC/IFORS Symposium on Identification and System Parameter Estimation, Budapest, Hungary, 1991, pp. 827–832.

[28] G. Vinnicombe, *Frequency domain uncertainty and the graph topology*, IEEE Trans. Automat. Control, 38 (1993), pp. 1371–1383.

[29] S. Wu, S. P. Boyd, and L. Vandenberghe, *FIR filter design via semidefinite programming and spectral factorization*, in Proceedings of the 35th IEEE Conference on Decision and Control, Kobe, Japan, 1996.

[30] M. Zarrop, *Optimal Experiment Design for Dynamic System Identification*, Lecture Notes in Control and Inform. Sci. 21, Springer-Verlag, Berlin, New York, 1979.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 5, pp. 1609–1625

STABILITY ANALYSIS OF SECOND-ORDER SWITCHED HOMOGENEOUS SYSTEMS*

DAVID HOLCMAN† AND MICHAEL MARGALIOT‡

Abstract. We study the stability of second-order switched homogeneous systems. Using the concept of *generalized first integrals* we explicitly characterize the "most destabilizing" switching-law and construct a Lyapunov function that yields an easily verifiable, necessary and sufficient condition for asymptotic stability. Using the duality between stability analysis and control synthesis, this also leads to a novel algorithm for designing a stabilizing switching controller.

Key words. absolute stability, switched linear systems, robust stability, hybrid systems, hybrid control

AMS subject classifications. 37N35, 93D15, 93D20, 93D30

PII. S0363012901389354

1. Introduction.

We consider the *switched homogeneous system*

$$\dot{\mathbf{x}}(t) \in \mathbf{\Omega}(\mathbf{x}(t)), \quad \mathbf{\Omega}(\mathbf{x}) := Co\{\mathbf{f}_1(\mathbf{x}), \mathbf{f}_2(\mathbf{x}), \ldots, \mathbf{f}_q(\mathbf{x})\}, \tag{1.1}$$

where $\mathbf{x}(t) = (x_1(t), \ldots, x_n(t))^T$, the $\mathbf{f}_i(\cdot)$'s are homogeneous functions (with equal degree of homogeneity), and *Co* denotes the convex hull. An important special case is $\mathbf{f}_i(\mathbf{x}) = A_i\mathbf{x}$, $i = 1, \ldots, q$, for which (1.1) reduces to a *switched linear system*.

Switched systems appear in many fields of science ranging from economics to electrical and mechanical engineering [15], [18]. In particular, switched linear systems were studied in the literature under various names, e.g., polytopic linear differential inclusions [4], linear polysystems [6], bilinear systems [5], and uncertain linear systems [20].

If $\mathbf{f}_i(\mathbf{0}) = \mathbf{0}$ for all i, then $\mathbf{0}$ is an equilibrium point of (1.1). Analyzing the stability of this equilibrium point is difficult because the system admits infinitely many solutions for every initial value.[1]

Stability analysis of switched *linear* systems can be traced back to the 1940s since it is closely related to the well-known *absolute stability problem* [4], [19]. Current approaches to stability analysis include (i) deriving sufficient but *not* necessary and sufficient stability conditions, and (ii) deriving necessary and sufficient stability conditions for the particular case of low-order systems. Popov's criterion, the circle criterion [19, Chapter 5], and the positive-real lemma [4, Chapter 2] can all be considered as examples of the first approach. Many other sufficient conditions exist in the literature.[2] Nevertheless, these conditions are sufficient but not necessary and sufficient and are known to be rather conservative conditions.

*Received by the editors May 10, 2001; accepted for publication (in revised form) July 5, 2002; published electronically January 14, 2003.
http://www.siam.org/journals/sicon/41-5/38935.html
†Department of Theoretical Mathematics, Weizmann Institute of Science, Rehovot, Israel 76100 (holcman@wisdom.weizmann.ac.il).
‡Department of Electrical Engineering-Systems, Tel Aviv University, Israel 69978 (michaelm@eng.tau.ac.il).
[1] An analysis of the computational complexity of some closely related problems can be found in [2].
[2] See, for example, the recent survey paper by Liberzon and Morse [11].

Far more general results were derived for the second approach, namely, the particular case of low-order linear switched systems. The basic idea is to single out the "most unstable" solution $\tilde{\mathbf{x}}(t)$ of (1.1), that is, a solution with the following property: If $\tilde{\mathbf{x}}(t)$ converges to the origin, then so do *all* the solutions of (1.1). Then, all that is left to analyze is the stability of this single solution (see, e.g., [3]).

Pyatnitskiy and Rapoport [16] and Rapoport [17] were the first to formulate the problem of finding the "most unstable" solution of (1.1) using a variational approach. Applying the maximum principle, they developed a characterization of this solution in terms of a two-point boundary value problem. Their characterization is not explicit but, nevertheless, using tools from convex analysis they proved the following result. Let Γ be the collection of all the q-sets of *linear* functions $\{A_1\mathbf{x},\ldots,A_q\mathbf{x}\}$ for which (1.1) is asymptotically stable, and denote the boundary[3] of Γ by $\partial\Gamma$. Pyatnitskiy and Rapoport proved that if $\{A_1\mathbf{x},\ldots,A_q\mathbf{x}\}\in\partial\Gamma$, then the "most unstable" solution of (1.1) is a *closed* trajectory. Intuitively, this can be explained as follows. If $\{A\mathbf{x},B\mathbf{x}\}\in\Gamma$, then, by the definition of Γ, $\tilde{\mathbf{x}}(t)$ converges to the origin; if $\{A\mathbf{x},B\mathbf{x}\}\notin(\Gamma\cup\overline{\Gamma})$, then $\tilde{\mathbf{x}}(t)$ is unbounded. Between these two extremes, that is, when $\{A\mathbf{x},B\mathbf{x}\}\in\partial\Gamma$, $\tilde{\mathbf{x}}(t)$ is a closed solution. This leads to a *necessary and sufficient* stability condition for second- and third-order switched linear systems [16], [17]; however, the condition is a nonlinear equation in several unknowns and, since solving this equation turns out to be difficult, it cannot be used in practice.

Margaliot and Langholz [14] introduced the novel concept of *generalized first integrals* and used it to provide a different characterization of the closed trajectory. Unlike Pyatnitskiy and Rapoport, the characterization is *constructive* and leads, for second-order switched linear systems, to an *easily verifiable*, necessary and sufficient stability condition. Furthermore, their approach yields an *explicit* Lyapunov function for switched linear systems.

In the general homogeneous case, the functions $\mathbf{f}_i(\cdot)$ are *nonlinear* functions, and therefore the approaches used for switched linear systems cannot be applied. Filippov [7] derived a necessary and sufficient stability condition for second-order switched homogeneous systems. However, his proof uses a Lyapunov function that is not constructed explicitly.

In this paper we combine Filippov's approach with the approach developed by Margaliot and Langholz to provide a necessary and sufficient condition for asymptotic stability of second-order switched homogeneous systems. We construct a suitable *explicit* Lyapunov function and derive a condition that is easy to check in practice.

A closely related problem is the stabilization of several unstable systems using switching. This problem has recently regained interest with the discovery that there are systems that can be stabilized by *hybrid controllers* whereas they cannot be stabilized by continuous state-feedback [18, Chapter 6]. To analyze the stability of (1.1), we synthesize the "most unstable" solution $\tilde{\mathbf{x}}(t)$ by switching between several asymptotically stable systems. Designing a switching controller is equivalent to synthesizing the "most stable" solution by switching between several unstable systems. These problems are dual and, therefore, a solution of the first is also a solution of the second. Consequently, we use our stability analysis to develop a novel procedure for designing a stabilizing switching controller for second-order homogeneous systems.

The rest of this paper is organized as follows. Section 2 includes some notations and assumptions. Section 3 develops the *generalized first integral* which will serve as our main analysis tool. Section 4 analyzes the sets Γ and $\partial\Gamma$. Section 5 provides an

[3]The set Γ is open [17].

explicit characterization of the "most destabilizing" switching-law. Section 6 presents an easily verifiable, *necessary and sufficient* stability condition. Section 7 describes a new algorithm for designing a switching controller. Section 8 summarizes.

2. Notations and assumptions. For $\beta > 1$, let

$$P_\beta := \left\{ f(\cdot,\cdot) : f(cx_1, cx_2) = c^\beta f(x_1, x_2) \text{ for all } c, x_1, x_2 \right\},$$

that is, the set of homogeneous functions of degree β. We denote by E_β the set of functions $\mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^2$ such that $\mathbf{f}(x_1, x_2) = (f_1(x_1, x_2), f_2(x_1, x_2))^T$ with $f_1, f_2 \in P_\beta$.

Consider the system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$, where $\mathbf{x} = (x_1, x_2)^T$ and $\mathbf{f} \in E_\beta$. Transforming to polar coordinates,

$$r(t) = \sqrt{x_1^2(t) + x_2^2(t)}, \qquad \theta(t) = \arctan\left(\frac{x_2(t)}{x_1(t)}\right),$$

we get

$$\dot{r} = r^\beta R(\theta), \qquad \dot{\theta} = r^{\beta-1} A(\theta), \tag{2.1}$$

where $R(\theta)$ and $A(\theta)$ are homogeneous functions of degree $\beta+1$ in the variables $\cos(\theta)$ and $\sin(\theta)$.

Following [9, Chapter III], we analyze the stability of (2.1) by considering two cases. If $A(\cdot)$ has no zeros, then the origin is a focus and (2.1) yields $r(\theta) = r_0 e^{\int_{\theta_0}^{\theta} \frac{R(u)}{A(u)} du} = r_0 p(\theta;\theta_0) e^{h\theta}$, where p is periodic in θ with period 2π, and $h := \frac{1}{2\pi}\int_0^{2\pi} \frac{R(u)}{A(u)} du$. Hence, $r(t) \to 0$ ($r(t) \to \infty$) if $sgn(h) \neq sgn(A)$ ($sgn(h) = sgn(A)$).

If A has zeros, say $A(\overline{\theta}) = 0$, then the line $\theta = \overline{\theta}$ is a solution of (2.1) (the origin is a node) and along this line $r(t) \to 0$ ($r(t) \to \infty$) if $R(\overline{\theta}) < 0$ ($R(\overline{\theta}) > 0$).

Hence, if $ES_\beta := \{\mathbf{f} \in E_\beta : \dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ is asymptotically stable$\}$, then $ES_\beta = ES_\beta^F \cup ES_\beta^N$,[4] where

$$ES_\beta^F := \left\{ \mathbf{f} \in E_\beta : A(\theta) \text{ has no zeros and } sgn(h) \neq sgn(A) \right\},$$
$$ES_\beta^N := \left\{ \mathbf{f} \in E_\beta : R(\overline{\theta}) < 0 \text{ for all } \overline{\theta} \text{ such that } A(\overline{\theta}) = 0 \right\}.$$

Given $\mathbf{f}(\mathbf{x}) \in E_\beta$, we denote its differential at $\mathbf{x}$ by

$$(D\mathbf{f})(\mathbf{x}) := \begin{pmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \frac{\partial f_1(\mathbf{x})}{\partial x_2} \\ \frac{\partial f_2(\mathbf{x})}{\partial x_1} & \frac{\partial f_2(\mathbf{x})}{\partial x_2} \end{pmatrix}.$$

The differential's norm is $||(Df)(\mathbf{x})|| := \sup_{\mathbf{h}\in\mathbb{R}^2, ||\mathbf{h}||=1} ||(Df)(\mathbf{x})\mathbf{h}||$, where $||\cdot|| : \mathbb{R}^2 \to \mathbb{R}_+$ is some vector norm on $\mathbb{R}^2$. The distance between two functions $\mathbf{f}, \mathbf{g} \in E_\beta$ is defined by [10]

$$d(\mathbf{f}, \mathbf{g}) := \sup_{\mathbf{x}\,:\,||\mathbf{x}||<1} (||\mathbf{f}(\mathbf{x}) - \mathbf{g}(\mathbf{x})|| + ||(D\mathbf{f})(\mathbf{x}) - (D\mathbf{g})(\mathbf{x})||). \tag{2.2}$$

Note that $(E_\beta, d(\cdot,\cdot))$ is a Banach space and that in the topology induced by $d(\cdot,\cdot)$ the set ES_β is *open*.

[4] Here F stands for focus and N for node.

For simplicity,[5] we consider the differential inclusion (1.1) with $q = 2$:

$$\dot{\mathbf{x}}(t) \in \mathbf{\Omega}(\mathbf{x}(t)), \quad \mathbf{\Omega}(\mathbf{x}) := Co\{\mathbf{f}(\mathbf{x}), \mathbf{g}(\mathbf{x})\} \tag{2.3}$$

with $\mathbf{f}, \mathbf{g} \in ES_\beta$.

Given an initial condition $\mathbf{x}_0$, a solution of (2.3) is an absolutely continuous function $\mathbf{x}(t)$, with $\mathbf{x}(0) = \mathbf{x}_0$, that satisfies (2.3) for almost all t. Clearly, there is an infinite number of solutions for any initial condition. To differentiate the possible solutions we use the concept of a switching-law.

DEFINITION 2.1. *A* switching-law *is a piecewise constant function* $\eta : [0, +\infty) \to [0, 1]$. *We refer to the solution of* $\dot{\mathbf{x}} = \eta(t)\mathbf{f}(\mathbf{x}) + (1 - \eta(t))\mathbf{g}(\mathbf{x})$ *as the solution corresponding to the switching-law* η.

The solution $\mathbf{x}(t) \equiv \mathbf{0}$ is said to be *uniformly*[6] *locally asymptotically stable* if

- given any $\epsilon > 0$, there exists $\delta(\epsilon) > 0$ such that *every* solution of (2.3) with $||\mathbf{x}(0)|| < \delta(\epsilon)$ satisfies $||\mathbf{x}(t)|| < \epsilon$ for all $t \geq 0$,
- there exists $c > 0$ such that *every* solution of (2.3) satisfies $\lim_{t\to\infty} \mathbf{x}(t) = \mathbf{0}$ if $||\mathbf{x}(0)|| < c$.

Since $\mathbf{f}$ and $\mathbf{g}$ are homogeneous, local asymptotic stability of (2.3) implies global asymptotic stability. Hence, when the above conditions hold, the system is uniformly globally asymptotically stable (UGAS).

DEFINITION 2.2. *A set* $P \subset \mathbb{R}^2$ *is an* invariant set *of* (2.3) *if* every *solution* $\mathbf{x}(t)$, *with* $\mathbf{x}(0) \in P$, *satisfies* $\mathbf{x}(t) \in P$ *for all* $t \geq 0$.

DEFINITION 2.3. *We will say that* $\mathbf{\Omega}(\mathbf{x}) = Co\{\mathbf{f}(\mathbf{x}), \mathbf{g}(\mathbf{x})\}$ *is* singular *if there exists an invariant set that does not contain an open neighborhood of the origin.*

We assume the following from here on.

Assumption 1. The set $\mathbf{\Omega}(\mathbf{x})$ is not singular.

The role of Assumption 1 will become clear in the proof of Lemma 5.4 below. Note that it is easy to check if the assumption holds by transforming the two systems $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ and $\dot{\mathbf{x}} = \mathbf{g}(\mathbf{x})$ to polar coordinates and examining the set of points where $\dot{\theta} = 0$ for each system. For example, if there exists a line l that is an invariant set for both $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ and $\dot{\mathbf{x}} = \mathbf{g}(\mathbf{x})$, then l is an invariant set of (2.3) and Assumption 1 does not hold.

To make the stability analysis nontrivial, we also assume the following.

Assumption 2. For any *fixed* $\overline{\eta} \in [0, 1]$, the origin is a globally asymptotically stable equilibrium point of $\dot{\mathbf{x}} = \overline{\eta}\mathbf{f}(\mathbf{x}) + (1 - \overline{\eta})\mathbf{g}(\mathbf{x})$.

3. The generalized first integral.

If the system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) \tag{3.1}$$

is Hamiltonian [8], then it admits a *classical* first integral, that is, a function $H(\mathbf{x})$ which satisfies $H(\mathbf{x}(t)) \equiv H(\mathbf{x}(0))$ along the trajectories of (3.1). In this case, the study of (3.1) is greatly simplified since its trajectories are nothing but the contours $H(\mathbf{x}) = \text{const}$. In particular, it turns out that the first integral provides a crucial analysis tool for switched linear systems [14]. The purpose of this section is to extend this idea to the case where $\mathbf{f} \in ES_\beta$ and, therefore, (3.1) is not Hamiltonian.

Let $v := x_2/x_1$; then

$$\frac{dv}{\frac{dx_2}{dx_1} - v} = \frac{dx_1}{x_1}.$$

[5] Our results can be easily generalized to the case $q > 2$.

[6] The term "uniform" is used here to describe uniformity with respect to switching signals.

If $\mathbf{f} \in ES_\beta$, then f_1 and f_2 are both homogeneous functions of degree β and, therefore, the ratio $\frac{f_2(x_1,x_2)}{f_1(x_1,x_2)}$ is a function of v only, which we denote by $\alpha(v)$. Hence, along the trajectories of (3.1), $\frac{dv}{\alpha(v)-v} = \frac{dx_1}{x_1}$, that is, $e^{\int \frac{dv}{v-\alpha(v)}}|x_1| = \text{const}$. Thus, we define the *generalized first integral* of (3.1) by

$$H(x_1, v) := \left(x_1 e^{L(v)}\right)^{2k}, \tag{3.2}$$

where $L(v) := \int \frac{dv}{v-\alpha(v)}$ and k is a positive integer. Note that we can write $H = H(x_1, x_2)$ by substituting $v = x_2/x_1$. Note also that $H(\lambda x_1, \lambda x_2) = \lambda^{2k} H(x_1, x_2)$.

Let S be the collection of points where $H(x_1, x_2)$ is not defined or not continuous; then, by construction, $H : \mathbb{R}^2 \setminus S \to \mathbb{R}_+$ is *piecewise constant* along the trajectories of (3.1). If $S = \emptyset$, then H is a classical first integral of the system. In general, however, $S \neq \emptyset$. Nevertheless, this does not imply that H cannot be used in the analysis of (3.1). Consider, for example, the case where S is a line and a trajectory $\mathbf{x}(t)$ of (3.1) can cross S but not stay on S. Then, $H(\mathbf{x}(t))$ will remain constant except perhaps at a crossing time where its value can "jump."[7] Thus, a trajectory of the system is a concatenation of several contours of H. This motivates the term *generalized first integral*.

To clarify the relationship between the trajectories of $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ and the contours $H(\mathbf{x}) = \text{const}$, we consider an example.

EXAMPLE 1. *Consider the system*

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} -x_2^3 - 2x_1^3 \\ x_1 x_2^2 \end{pmatrix}. \tag{3.3}$$

Here (3.2) *yields*

$$H(x_1, v) = \left(x_1 \frac{v(2 - v + v^2)^{\frac{1}{8}}}{(1+v)^{\frac{1}{4}}} e^{-\frac{3}{4\sqrt{7}} \arctan((-1+2v)/\sqrt{7})} \right)^{2k},$$

and using $k = 2$ *and* $v = x_2/x_1$ *we get*

$$H(x_1, x_2) = \frac{x_2^4 \sqrt{2x_1^2 - x_1 x_2 + x_2^2}}{x_1 + x_2} e^{-\frac{3}{\sqrt{7}} \arctan(\frac{2x_2 - x_1}{\sqrt{7} x_1})}.$$

In this case $S = l_1 \cup l_2$, *where* $l_1 := \{\mathbf{x} : x_1 + x_2 = 0\}$ *and* $l_2 := \{\mathbf{x} : x_1 = 0\}$. *It is easy to verify that* l_1 *is an invariant set of* (3.3), *that is,* $\mathbf{x}(t) \cap l_1 = \emptyset$ *(except for the trivial trajectory that starts and stays on* l_1*). Furthermore, it is easy to see that a trajectory of* (3.3) *cannot stay on the line* l_2.

Figure 1 *shows the trajectory* $\mathbf{x}(t)$ *of* (3.3) *for* $\mathbf{x}_0 = (3, 1)^T$. *Figure* 2 *displays* $H(\mathbf{x}(t))$ *as a function of time. It may be seen that* $H(\mathbf{x}(t))$ *is a piecewise constant function that attains two values. Note that the "jump" in* $H(\mathbf{x}(t))$ *occurs when* $x_1(t) = 0$, *that is, when* $\mathbf{x}(t) \in S$.

4. The boundary of stability. Let Γ be the set of all pairs $(\mathbf{f}, \mathbf{g})$ for which (2.3) is UGAS. In this section we study Γ and its boundary $\partial\Gamma$. Our first result, whose proof is given in the appendix, is an inverse Lyapunov theorem.

LEMMA 4.1. *If* $(\mathbf{f}, \mathbf{g}) \in \Gamma$, *then there exists a* C^1 *positive-definite function* $V(\mathbf{x}) : \mathbb{R}^2 \to [0, +\infty)$ *such that for all* $\mathbf{x} \in \mathbb{R}^2 \setminus \mathbf{0}$, $\nabla V(\mathbf{x})\mathbf{f}(\mathbf{x}) < 0$ *and* $\nabla V(\mathbf{x})\mathbf{g}(\mathbf{x}) < 0$. *Furthermore,* $V(\mathbf{x})$ *is positively homogeneous of degree one.*[8]

[7] That is, a time t_0 such that $\mathbf{x}(t_0) \in S$.

[8] That is, $V(c\mathbf{x}) = cV(\mathbf{x})$ for all $c > 0$ and all $\mathbf{x} \in \mathbb{R}^2$.

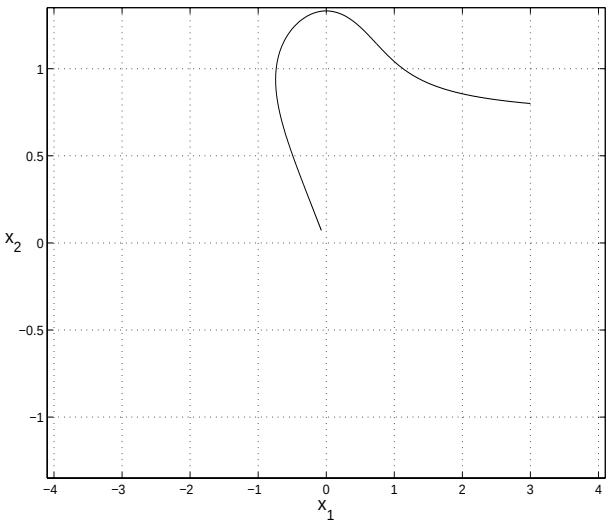


FIG. 1. *The trajectory of* (3.3) *for* $\mathbf{x}_0 = (3,1)^T$.

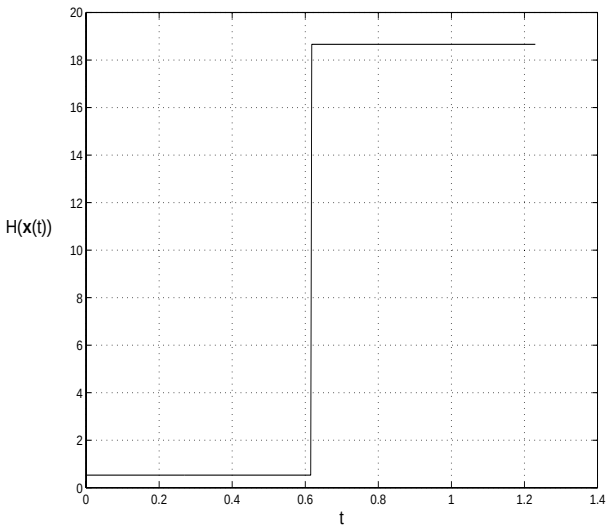


FIG. 2. $H(\mathbf{x}(t))$ *as a function of time.*

LEMMA 4.2. Γ *is an open cone.*

Proof. Let $(\mathbf{f},\mathbf{g}) \in \Gamma$. Clearly, $(c\mathbf{f}, c\mathbf{g}) \in \Gamma$ for all $c > 0$. Hence, Γ is a cone.

To prove that Γ is open, we use the common Lyapunov function V from Lemma 4.1. Denote $\gamma := \{\mathbf{x} : V(\mathbf{x}) = 1\}$, so γ is a closed curve encircling the origin. Hence, there exists $a < 0$ such that for all $\mathbf{x} \in \gamma$,

$$\nabla V(\mathbf{x})\mathbf{f}(\mathbf{x}) < a \quad \text{and} \quad \nabla V(\mathbf{x})\mathbf{g}(\mathbf{x}) < a. \tag{4.1}$$

If $\tilde{\mathbf{f}} \in ES_\beta$ and $\tilde{\mathbf{g}} \in ES_\beta$ are such that $d(\tilde{\mathbf{f}},\mathbf{f}) < \epsilon$ and $d(\tilde{\mathbf{g}},\mathbf{g}) < \epsilon$, with $\epsilon > 0$ sufficiently small, then for all $\mathbf{x} \in \gamma$, $\nabla V(\mathbf{x})\tilde{\mathbf{f}}(\mathbf{x}) < a/2 < 0$ and $\nabla V(\mathbf{x})\tilde{\mathbf{g}}(\mathbf{x}) < a/2 < 0$. It follows from the homogeneity of V, $\tilde{\mathbf{f}}$, and $\tilde{\mathbf{g}}$ that $(\tilde{\mathbf{f}},\tilde{\mathbf{g}}) \in \Gamma$. ∎

5. The worst-case switching-law. In this section we provide two explicit characterizations of the switching-law that yields the "most unstable" solution of (2.3).

Let $H^{\mathbf{f}}(\mathbf{x})$ ($H^{\mathbf{g}}(\mathbf{x})$) be the generalized first integral of $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ ($\dot{\mathbf{x}} = \mathbf{g}(\mathbf{x})$).

DEFINITION 5.1. *Define the* worst-case switching-law (WCSL) *by*

$$\lambda(\mathbf{x}) := \begin{cases} 0 & \text{if } \nabla H^{\mathbf{f}}(\mathbf{x})\mathbf{g}(\mathbf{x}) \geq 0, \\ 1 & \text{if } \nabla H^{\mathbf{f}}(\mathbf{x})\mathbf{g}(\mathbf{x}) < 0. \end{cases} \tag{5.1}$$

We denote

$$\mathbf{h}(\mathbf{x}) := \lambda(\mathbf{x})\mathbf{f}(\mathbf{x}) + (1 - \lambda(\mathbf{x}))\mathbf{g}(\mathbf{x})$$

so the solution corresponding to the WCSL satisfies $\dot{\mathbf{x}} = \mathbf{h}(\mathbf{x})$. Note that the WCSL is a *state-dependent* switching-law and that since $\lambda(\mathbf{x}) = 0$ or $\lambda(\mathbf{x}) = 1$, then $\mathbf{h}(\mathbf{x}) = \mathbf{g}(\mathbf{x})$ or $\mathbf{h}(\mathbf{x}) = \mathbf{f}(\mathbf{x})$, respectively, that is, the vertices of $\mathbf{\Omega}$. Furthermore, it is easy to see that $\mathbf{h}(\mathbf{x})$ is homogeneous of degree β.

Intuitively, the WCSL can be explained as follows. Consider a point $\mathbf{x}$ where $\mathbf{f}(\mathbf{x})$ and $\mathbf{g}(\mathbf{x})$ are as shown in Figure 3. A solution of $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ follows the contour $H^{\mathbf{f}}(\mathbf{x}) = \text{const}$, whereas a solution of $\dot{\mathbf{x}} = \mathbf{g}(\mathbf{x})$ crosses this contour going further away from the origin. In this case, $\nabla H^{\mathbf{f}}(\mathbf{x})\mathbf{g}(\mathbf{x}) > 0$, so the WCSL is $\lambda(\mathbf{x}) = 0$, which corresponds to setting $\dot{\mathbf{x}} = \mathbf{g}(\mathbf{x})$. Thus, the WCSL "pushes" the trajectory away from the origin as much as possible.

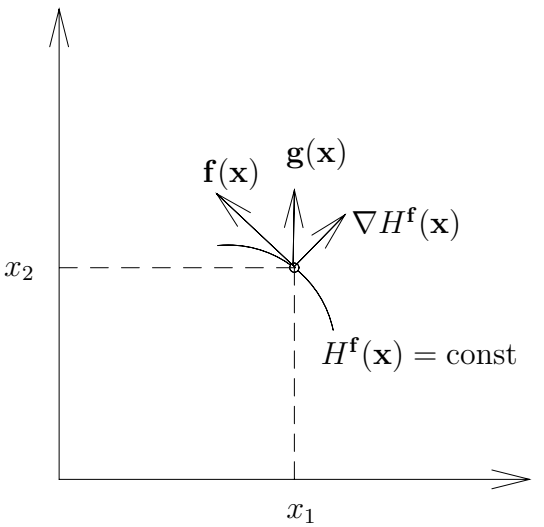


FIG. 3. *Geometrical explanation of the WCSL when* $\nabla H^{\mathbf{f}}(\mathbf{x})\mathbf{g}(\mathbf{x}) > 0$.

Note that the definition of WCSL using (5.1) is meaningful only for $\mathbf{x} \in \mathbb{R}^2 \setminus S$ since $\nabla H^{\mathbf{f}}(\mathbf{x})$ is not defined for $\mathbf{x} \in S$. However, extending the definition of WCSL to any $\mathbf{x} \in \mathbb{R}^2$ is immediate since $\mathbf{x} \in S$ implies one of two cases. In the first case, $\mathbf{x} \in l$, where l is a line in $\mathbb{R}^2$ which is an invariant set of $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$, that is, $\mathbf{f}(\mathbf{x}) = c\mathbf{x}$ (with $c < 0$ since $\mathbf{f}$ is asymptotically stable), so clearly the WCSL must use $\mathbf{g}$. In the second case, the trajectory of $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ crosses S so the value of the switching-law on the single point $\mathbf{x}$ can be chosen arbitrarily.

We expect the WCSL to remain unchanged if we swap the roles of $\mathbf{f}$ and $\mathbf{g}$. Indeed, this is guaranteed by the following lemma, whose proof is given in the appendix.

LEMMA 5.2. *For all* $\mathbf{x} \in D := \{\mathbf{x} : \mathbf{f}^T(\mathbf{x})\mathbf{g}(\mathbf{x}) > 0\}$

$$sgn(\nabla H^{\mathbf{f}}(\mathbf{x})\mathbf{g}(\mathbf{x})) = -sgn(\nabla H^{\mathbf{g}}(\mathbf{x})\mathbf{f}(\mathbf{x})), \tag{5.2}$$

where $sgn(\cdot)$ *is the sign function.*

We can now state the main result of this section.

THEOREM 5.3. $(\mathbf{f}, \mathbf{g}) \in \partial\Gamma$ *if and only if the solution corresponding to the WCSL is closed.*[9]

Proof. Denote the solution corresponding to the WCSL by $\mathbf{x}(t)$ and suppose that $\mathbf{x}(t)$ is closed. Let γ be the closed curve $\{\mathbf{x}(t) : t \in [0, T]\}$, where $T > 0$ is the smallest time such that $\mathbf{x}(T) = \mathbf{x}(0)$. Note that using the explicit construction of $\lambda(\mathbf{x})$ (see (5.1)) we can easily define γ *explicitly* as a concatenation of several contours of $H^{\mathbf{f}}(\mathbf{x})$ and $H^{\mathbf{g}}(\mathbf{x})$. Note also that the switching between $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ and $\dot{\mathbf{x}} = \mathbf{g}(\mathbf{x})$ takes place at points $\mathbf{x}$ where $\nabla H^f(\mathbf{x})\mathbf{g}(\mathbf{x}) = 0$ (see (5.1)), that is, when $\mathbf{g}(\mathbf{x})$ and $\mathbf{f}(\mathbf{x})$ are collinear. Hence, the curve γ has no corners.

We define the function $V(\mathbf{x})$ by $V(\mathbf{0}) = 0$, and for all $\mathbf{x} \neq \mathbf{0}$

$$V(\mathbf{x}) = k \quad \text{such that} \quad \mathbf{x} \in k\gamma, \tag{5.3}$$

that is, the contours of V are obtained by scaling γ (see [1]). The function $V(\mathbf{x})$ is positively homogeneous (that is, for any $c \geq 0$, $V(c\mathbf{x}) = cV(\mathbf{x})$), radially unbounded, and differentiable on $\mathbb{R}^2 \setminus \{\mathbf{0}\}$.

Let $\mathbf{p}(\mathbf{x}) = ||\mathbf{x}||^{\beta-1}\mathbf{x}$ and denote $\mathbf{f}^\epsilon(\mathbf{x}) := \mathbf{f}(\mathbf{x}) + \epsilon\mathbf{p}(\mathbf{x})$ and $\mathbf{g}^\epsilon(\mathbf{x}) := \mathbf{g}(\mathbf{x}) + \epsilon\mathbf{p}(\mathbf{x})$. Note that both $\mathbf{f}^\epsilon(\mathbf{x})$ and $\mathbf{g}^\epsilon(\mathbf{x})$ belong to E_β. We use $V(\mathbf{x})$ to analyze the stability of the perturbed system $\dot{\mathbf{x}} \in \mathbf{\Omega}^\epsilon(\mathbf{x}) := Co\{\mathbf{f}^\epsilon(\mathbf{x}), \mathbf{g}^\epsilon(\mathbf{x})\}$. Consider the derivative of V along the trajectories of $\dot{\mathbf{x}} \in \mathbf{\Omega}^\epsilon(\mathbf{x})$:

$$\begin{aligned}\dot{V}(\mathbf{x}) &= \nabla V(\mathbf{x})\left(\eta(t)(\mathbf{f}^\epsilon(\mathbf{x}) + (1-\eta(t))\mathbf{g}^\epsilon(\mathbf{x})\right) \\ &= \epsilon\nabla V(\mathbf{x})\mathbf{p}(\mathbf{x}) + \eta(t)\nabla V(\mathbf{x})\mathbf{f}(\mathbf{x}) + (1-\eta(t))\nabla V(\mathbf{x})\mathbf{g}(\mathbf{x}),\end{aligned}$$

where $\eta(t) \in [0, 1]$ for all t. If at some $\mathbf{x}$, $V(\mathbf{x})$ corresponds to a contour $H^{\mathbf{f}}(\mathbf{x}) = \text{const}$, then $\nabla V(\mathbf{x})\mathbf{f}(\mathbf{x}) = 0$ and, by the definition of WCSL (see (5.1)), $\nabla V(\mathbf{x})\mathbf{g}(\mathbf{x}) \leq 0$ so $\dot{V}(\mathbf{x}) \leq \epsilon\nabla V(\mathbf{x})\mathbf{p}(\mathbf{x})$. Otherwise, $V(\mathbf{x})$ corresponds to a contour $H^{\mathbf{g}}(\mathbf{x}) = \text{const}$, so $\nabla V(\mathbf{x})\mathbf{g}(\mathbf{x}) = 0$, $\nabla V(\mathbf{x})\mathbf{f}(\mathbf{x}) \leq 0$, and again $\dot{V}(\mathbf{x}) \leq \epsilon\nabla V(\mathbf{x})\mathbf{p}(\mathbf{x})$. Hence, for any $\epsilon < 0$ we have

$$\dot{V}(\mathbf{x}) \leq \epsilon\nabla V(\mathbf{x})\mathbf{p}(\mathbf{x}) = \epsilon||\mathbf{x}||^{\beta-1}\nabla V(\mathbf{x})\mathbf{x} < 0;$$

since this holds for all $\mathbf{x}$ and all $\eta(t) \in [0, 1]$, we get that for $\epsilon < 0$, $\mathbf{\Omega}^\epsilon \in \Gamma$.

On the other hand, for $\epsilon > 0$ and $\eta(t) = \lambda(\mathbf{x}(t))$ we have

$$\dot{V}(\mathbf{x}) = \epsilon\nabla V(\mathbf{x})\mathbf{p}(\mathbf{x}) = \epsilon||\mathbf{x}||^{\beta-1}\nabla V(\mathbf{x})\mathbf{x} > 0;$$

since this holds for all $\mathbf{x}$, $\dot{\mathbf{x}} \in \mathbf{\Omega}^\epsilon(\mathbf{x})$ admits an unbounded solution for $\epsilon > 0$. The derivations above hold for arbitrarily small ϵ and, therefore, $\mathbf{\Omega} \in \partial\Gamma$.

For the opposite direction, assume that $(\mathbf{f}, \mathbf{g}) \in \partial\Gamma$, and let $\mathbf{x}(t)$ be the solution corresponding to the WCSL, that is, $\mathbf{x}(t)$ satisfies $\dot{\mathbf{x}} = \mathbf{h}(\mathbf{x}) := \lambda(\mathbf{x})\mathbf{f}(\mathbf{x}) +$

[9] We omit specifying the initial condition because the fact that $\mathbf{h}(\mathbf{x})$ is homogeneous implies that, if the solution starting at some $\mathbf{x}_0$ is closed, then all solutions are closed.

$(1-\lambda(\mathbf{x}))\mathbf{g}(\mathbf{x})$. To prove that $\mathbf{x}(t)$ is a closed trajectory, we use the following lemma, whose proof appears in the appendix.

LEMMA 5.4. *If* $(\mathbf{f},\mathbf{g}) \in \partial\Gamma$*, then the solution corresponding to the WCSL rotates around the origin.*

Thus, for a given $\mathbf{x}_0 \neq \mathbf{0}$, there exists $t_1 > 0$ such that $\mathbf{x}(t)$, with $\mathbf{x}(0) = \mathbf{x}_0$, satisfies $\mathbf{x}(t_1) = c\mathbf{x}_0$, and since $\mathbf{h}(\mathbf{x})$ is homogeneous, we get $\mathbf{x}(nt_1) = c^n\mathbf{x}(0)$, $n = 1, 2, 3, \ldots$. We consider two cases. If $c > 1$, then $\mathbf{x}(t)$ is unbounded, and using the homogeneity of $\mathbf{h}(\mathbf{x})$ we conclude that $\mathbf{0}$ is a (spiral) source. It follows from the theory of structural stability (see, e.g., [10]) that there exists an $\epsilon > 0$ such that for *all* $(\tilde{\mathbf{f}},\tilde{\mathbf{g}})$ with $d(\tilde{\mathbf{f}},\mathbf{f}) < \epsilon$ and $d(\tilde{\mathbf{g}},\mathbf{g}) < \epsilon$, the origin is a source of the perturbed dynamical system $\dot{\mathbf{x}} = \lambda(x)\tilde{\mathbf{f}}(\mathbf{x}) + (1-\lambda(x))\tilde{\mathbf{g}}(\mathbf{x})$. This implies that $(\mathbf{f},\mathbf{g}) \notin \partial\Gamma$, which is a contradiction.

If $c < 1$, then $\mathbf{x}(t)$ converges to the origin and, by the construction of the WCSL, so does any other solution, so $(\mathbf{f},\mathbf{g}) \in \Gamma$, which is again a contradiction. Hence, $c = 1$, that is, $\mathbf{x}(t)$ is closed. □

The characterization of the WCSL using the generalized first integrals leads to a simple and *constructive* proof of Theorem 5.3. However, to actually check whether the solution corresponding to the WCSL is closed, a characterization of the WCSL in polar coordinates is more suitable.

Representing (2.3) in polar coordinates, we get

$$\begin{pmatrix} \dot{r} \\ \dot{\theta} \end{pmatrix} \in \begin{pmatrix} \cos\theta & \sin\theta \\ -\frac{\sin\theta}{r} & \frac{\cos\theta}{r} \end{pmatrix} Co\{\mathbf{f}(r,\theta),\mathbf{g}(r,\theta)\}. \tag{5.4}$$

If $(\mathbf{f},\mathbf{g}) \in \partial\Gamma$, then the WCSL yields a closed solution. By using the transformation $\overline{r} = r$, $\overline{\theta} = -\theta$ (if necessary), we may always assume that this solution rotates around the origin in a counterclockwise direction, that is, $\dot{\theta}(r,\theta) > 0$ for all $\theta \in [0, 2\pi)$. Note that this implies that if at some point $\mathbf{x}$ the trajectories of one of the systems are in the clockwise direction, then the WCSL will use the second system. Hence, determining the WCSL is nontrivial only at points where the trajectories of both systems rotate in the same direction, and we assume from here on that both rotate in a clockwise direction. (Note that this explains why in Lemma 5.2 it is enough to consider $\mathbf{x} \in D$.)

Let $\mathbf{j}_\eta(r,\theta) := \eta\mathbf{f}(r,\theta) + (1-\eta)\mathbf{g}(r,\theta)$ and

$$F(r,\theta) := \{\eta \in [0,1] : (-\sin\theta \quad \cos\theta)\mathbf{j}_\eta(r,\theta) > 0\} \tag{5.5}$$

so F is a parameterization of the set of directions in $\boldsymbol{\Omega}$ for which $\dot{\theta} > 0$.

For any (r,θ) we define the switching-law

$$\zeta(r,\theta) := \arg\max_{\eta\in F} \frac{1}{r}\frac{\dot{r}}{\dot{\theta}};$$

that is, ζ is the switching-law that selects, among all the directions which yield $\dot{\theta} > 0$, the direction that maximizes $\frac{d\ln r}{d\theta}$. Using (5.4), we get

$$\zeta(r,\theta) = \arg\max_{\eta\in F} \frac{(\cos\theta \quad \sin\theta)\mathbf{j}_\eta(r,\theta)}{(-\sin\theta \quad \cos\theta)\mathbf{j}_\eta(r,\theta)}. \tag{5.6}$$

Let

$$m(r,\theta) := \frac{(\cos\theta \quad \sin\theta)\mathbf{j}_\zeta(r,\theta)}{(-\sin\theta \quad \cos\theta)\mathbf{j}_\zeta(r,\theta)} \tag{5.7}$$

so that along the trajectory corresponding to ζ, $\frac{1}{r}\dot{r}/\dot{\theta} = m$. Note that since $\mathbf{f}$ and $\mathbf{g}$ are homogeneous, $\zeta = \zeta(\theta)$ and $m = m(\theta)$.

It is easy to verify that the function $q(y) := \frac{ay+b(1-y)}{cy+d(1-y)}$, $y \in [0,1]$ (where c and d are such that the denominator is never zero), is monotonic and, therefore, $\zeta(r,\theta)$ in (5.6) is always 0 or 1 and $m(r,\theta)$ in (5.7) is always one of the two values,

$$m_0(\theta) := \frac{(\cos\theta \quad \sin\theta)\mathbf{g}(r,\theta)}{(-\sin\theta \quad \cos\theta)\mathbf{g}(r,\theta)}, \qquad m_1(\theta) := \frac{(\cos\theta \quad \sin\theta)\mathbf{f}(r,\theta)}{(-\sin\theta \quad \cos\theta)\mathbf{f}(r,\theta)},$$

respectively.

The next lemma, whose proof is given in the appendix, shows that the switching-law ζ is just the WCSL λ.

LEMMA 5.5. *The switching-law ζ yields a closed solution if and only if λ yields a closed solution.*

Let

$$\begin{aligned} I &:= \int_0^{2\pi} m(\theta)d\theta \\ &= \int_0^{2\pi} \frac{d\ln r}{d\theta}d\theta \\ &= \ln(r(T)) - \ln(r(0)), \end{aligned} \tag{5.8}$$

where $(r(t),\theta(t))$ is the solution corresponding to the switching-law ζ, and T is the time needed to complete a rotation around the origin. This solution is closed if and only if $\ln(r(T)) - \ln(r(0)) = 0$. Combining this with Lemma 5.5 and Theorem 5.3, we immediately obtain the following.

THEOREM 5.6. $(\mathbf{f},\mathbf{g}) \in \partial\Gamma$ *if and only if* $I = 0$.

It is easy to calculate I numerically and, therefore, Theorem 5.6 provides us with a simple recipe for determining whether $(\mathbf{f},\mathbf{g}) \in \partial\Gamma$. However, note that we assumed throughout that the closed solution of the system rotates in a counterclockwise direction. Thus, to use Theorem 5.6 correctly, I has to be computed twice: first for the original system and then for the transformed system $r' = r$, $\theta' = -\theta$ (denote this value by I'). $(\mathbf{f},\mathbf{g}) \in \partial\Gamma$ if and only if $\max(I, I') = 0$. In this way, we find whether the system has a closed trajectory, rotating around the origin in a clockwise or counterclockwise direction.

The following example demonstrates the use of Theorem 5.6.

EXAMPLE 2 (detecting the boundary of stability). *Consider the system*

$$\dot{\mathbf{x}} \in \Omega_k(\mathbf{x}) := Co\{\mathbf{f}(\mathbf{x}), \mathbf{g}_k(\mathbf{x})\}, \tag{5.9}$$

where

$$\mathbf{f}(\mathbf{x}) = \begin{pmatrix} -x_2^3 - 2x_1^3 \\ x_1x_2^2 \end{pmatrix}, \qquad \mathbf{g}_k(\mathbf{x}) = \begin{pmatrix} (kx_1 - x_2)^3 - 2x_1^3 \\ x_1(x_2 - kx_1)^2 \end{pmatrix}. \tag{5.10}$$

It is easy to verify that $\mathbf{f} \in ES_3^N$, and since $\mathbf{g}_0(\mathbf{x}) = \mathbf{f}(\mathbf{x})$, we have $\Omega_0 \in \Gamma$. The problem is to determine the smallest $k^* > 0$ such that $(\mathbf{f}(\mathbf{x}), \mathbf{g}_{k^*}(\mathbf{x})) \in \partial\Gamma$.

Transforming to polar coordinates we get

$$\mathbf{f}(r,\theta) = r^3 \begin{pmatrix} -\sin^3\theta - 2\cos^3\theta \\ \cos\theta\sin^2\theta \end{pmatrix}, \quad \mathbf{g}_k(r,\theta) = r^3 \begin{pmatrix} (k\cos\theta - \sin\theta)^3 - 2\cos^3\theta \\ \cos\theta(\sin\theta - k\cos\theta)^2 \end{pmatrix},$$

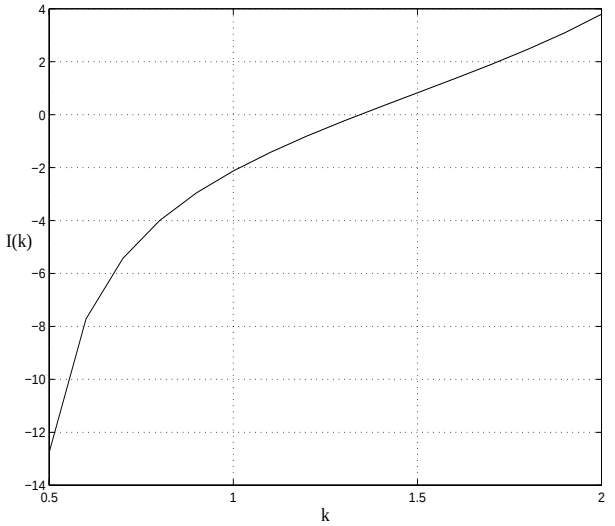


FIG. 4. *$I(k)$ as a function of k.*

so

$$j_\eta(r,\theta) = r^3\eta \begin{pmatrix} -\sin^3\theta - 2\cos^3\theta \\ \cos\theta\sin^2\theta \end{pmatrix} + r^3(1-\eta)\begin{pmatrix} (k\cos\theta - \sin\theta)^3 - 2\cos^3\theta \\ \cos\theta(\sin\theta - k\cos\theta)^2 \end{pmatrix}$$

and

$$I = \int_0^{2\pi} m(\theta)d\theta = \int_0^{2\pi} \max_{\eta\in F(\theta)} \frac{(\cos\theta \quad \sin\theta)\mathbf{j}_\eta(r,\theta)}{(-\sin\theta \quad \cos\theta)\mathbf{j}_\eta(r,\theta)}d\theta, \tag{5.11}$$

where $F(\theta)$ includes 0 if $(-\sin\theta \quad \cos\theta)\mathbf{j}_0(r,\theta) > 0$ and 1 if $(-\sin\theta \quad \cos\theta)\mathbf{j}_1(r,\theta) > 0$. Note that although j_η is a function of both r and θ, the integrand in (5.11) *is a function of θ (and k) but not of r.*

We calculated $I(k)$ numerically for different values of k. The results are shown in Figure 4. *The value k^* for which $I(k^*) = 0$ is*

$$k^* = 1.3439$$

(to four-digit accuracy), and it may be seen that for $k < k^*$ ($k > k^*$), $I(k) < 0$ ($I(k) > 0$). We repeated the computation for the transformed system $\overline{r} = r$, $\overline{\theta} = -\theta$ and found that there exists no closed solution rotating around the origin in a clockwise direction. Hence, the system (5.9) *and* (5.10) *is UGAS for all $k \in [0, k^*)$ and unstable for all $k > k^*$.*

The WCSL (see (5.6)*) for $k = k^*$ is*

$$\zeta(\theta) = \begin{cases} 0 & \text{if } \theta \in [0, 0.6256) \cup [1.1811, 3.7672) \cup [4.3227, 2\pi), \\ 1 & \text{otherwise.} \end{cases} \tag{5.12}$$

Figure 5 *depicts the solution of the system given by* (5.9) *and* (5.10) *with $k =$ 1.3439, WCSL* (5.12)*, and $\mathbf{x}_0 = (1,0)^T$. It may be seen that the solution is a closed trajectory, as expected. Note that this trajectory is* not *convex, which implies that the Lyapunov function used in the proof of Theorem* 5.3 *(see* (5.3)*) is not convex. This is*

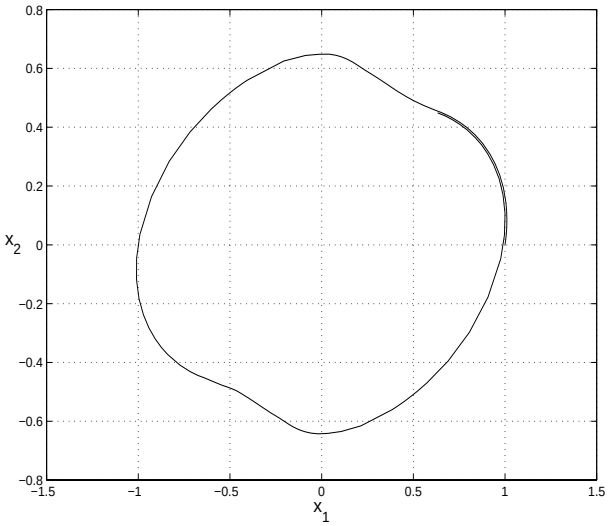


FIG. 5. *The solution of* (5.9) *and* (5.10) *for* $k = k^*$ *and the WCSL, with* $\mathbf{x}_0 = (1,0)^T$.

a phenomenon that is unique to nonlinear systems. For switched linear *systems the closed trajectory is convex and, therefore, so is the Lyapunov function* V *that yields a sufficient and necessary stability condition* [14].

6. Stability analysis. In this section we transform the original problem of analyzing the stability of (2.3) to one of detecting the boundary of stability $\partial\Gamma$. The latter problem was solved in section 5.

Given $\mathbf{\Omega} = Co\{\mathbf{f}, \mathbf{g}\}$, we define a new homogeneous function $\mathbf{h}_k(\mathbf{x})$ with the following properties: (1) $\mathbf{h}_0(\mathbf{x}) = \mathbf{f}(\mathbf{x})$; (2) $\mathbf{h}_1(\mathbf{x}) = \mathbf{g}(\mathbf{x})$; and (3) for all $k_1 < k_2$, $\{\mathbf{h}_k(\mathbf{x}) : 0 \le k \le k_1\} \subset \{\mathbf{h}_k(\mathbf{x}) : 0 \le k \le k_2\}$. One possible example that satisfies the above is

$$\mathbf{h}_k(\mathbf{x}) := \mathbf{f}(\mathbf{x}) + k(\mathbf{g}(\mathbf{x}) - \mathbf{f}(\mathbf{x})).$$

Consider the switched homogeneous system

$$\dot{\mathbf{x}}(t) \in \mathbf{\Omega}_k(\mathbf{x}(t)), \quad \mathbf{\Omega}_k := Co\{\mathbf{f}(\mathbf{x}), \mathbf{h}_k(\mathbf{x})\}. \tag{6.1}$$

The *absolute stability problem* is to find the smallest $k^* > 0$, when it exists, such that $\mathbf{\Omega}_{k^*} \in \partial\Gamma$. Noting that $\mathbf{\Omega}_0 = Co\{\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x})\} \in \Gamma$, $\mathbf{\Omega}_1 = Co\{\mathbf{f}(\mathbf{x}), \mathbf{g}(\mathbf{x})\} = \mathbf{\Omega}$, and $\mathbf{\Omega}_{k_1} \subset \mathbf{\Omega}_{k_2}$ for all $k_1 < k_2$, we immediately obtain the following result.

LEMMA 6.1. *The system* (2.3) *is UGAS if and only if* $k^* > 1$.

Thus, we can always transform the problem of analyzing the stability of a switched dynamical system into an absolute stability problem. We already know how to solve the latter problem for second-order homogeneous systems. To illustrate this consider the following example.

EXAMPLE 3. *Consider the system* (2.3) *with*

$$\mathbf{f}(\mathbf{x}) = \begin{pmatrix} -x_2^3 - 2x_1^3 \\ x_1 x_2^2 \end{pmatrix}, \qquad \mathbf{g}(\mathbf{x}) = \begin{pmatrix} (x_1 - x_2)^3 - 2x_1^3 \\ x_1(x_2 - x_1)^2 \end{pmatrix}. \tag{6.2}$$

It is easy to verify that $\mathbf{f}(\mathbf{x})$ *and* $\mathbf{g}(\mathbf{x})$ *belong to* ES_3 *and that both Assumptions* 1 *and* 2 *are satisfied.*

To analyze the stability of the system we use Lemma 6.1. *Defining*

$$\mathbf{h}_k(\mathbf{x}) = \begin{pmatrix} (kx_1 - x_2)^3 - 2x_1^3 \\ x_1(x_2 - kx_1)^2 \end{pmatrix}, \tag{6.3}$$

we must find the smallest k^* *such that* $(\mathbf{f}, \mathbf{h}_{k^*}) \in \partial\Gamma$. *We already calculated* k^* *in Example* 2 *and found that* $k^* = 1.3439 > 1$. *Hence, the system* (2.3) *with* $\mathbf{f}$ *and* $\mathbf{g}$ *given in* (6.2) *is UGAS.*

7. Designing a switching controller. In this section we employ our results to derive an algorithm for designing a switching controller for stabilizing homogeneous systems. To be concrete, we focus on linear systems rather than on the general homogeneous case. Hence, consider the system

$$\dot{\mathbf{x}} = A\mathbf{x} + B\mathbf{u}, \quad \mathbf{u} \in \mathcal{U} := Co\{K_1\mathbf{x}, K_2\mathbf{x}\}, \tag{7.1}$$

where K_1 and K_2 are given matrices that represent constraints on the possible controls.[10] We would like to design a stabilizing state-feedback controller $\mathbf{u}(t) = \mathbf{u}(\mathbf{x}(t))$ that satisfies the constraint $\mathbf{u}(t) \in \mathcal{U}$ for all t.

We assume that for any fixed matrix $K \in Co\{K_1, K_2\}$ the matrix $A + BK$ is strictly unstable and, therefore, a linear controller $\mathbf{u} = K\mathbf{x}$ will not stabilize the system. However, it is still possible that a switching controller will stabilize the system, and designing such a controller (if one exists) is the purpose of this section.

Roughly speaking, we are trying to find a switching-law that yields an asymptotically stable solution of $\dot{\mathbf{x}} \in \mathbf{\Omega} := Co\{A + BK_1, A + BK_2\}\mathbf{x}$, where each matrix in $\mathbf{\Omega}$ is strictly unstable. Using the transformation $\bar{t} = -t$, we see that such a solution exists if and only if this switching-law yields an unstable solution of $\dot{\mathbf{x}} \in \mathbf{\Omega}^- := Co\{-(A + BK_1), -(A + BK_2)\}\mathbf{x}$. Clearly, every matrix in $\mathbf{\Omega}^-$ is asymptotically stable. Hence, we obtain the main result of this section.

THEOREM 7.1. *Let* $\lambda = \lambda(\mathbf{x})$ *be the WCSL for the system*

$$\dot{\mathbf{x}} \in Co\{-(A + BK_1), -(A + BK_2)\}$$

and let $\tilde{\mathbf{x}}$ *be the corresponding solution. There exists a switching controller that asymptotically stabilizes* (7.1) *if and only if* $\tilde{\mathbf{x}}$ *is unbounded and, in this case,* $\mathbf{u}(\mathbf{x}) = \lambda(\mathbf{x})K_1\mathbf{x} + (1 - \lambda(\mathbf{x}))K_2\mathbf{x}$ *is a stabilizing controller.*

Note that Theorem 7.1 provides an algorithm for designing a stabilizing switching controller whenever such a controller exists. We already solved the problem of analyzing $\tilde{\mathbf{x}}$ for second-order systems.

EXAMPLE 4 (designing a stabilizing switching controller). *Consider the system* (7.1) *with*

$$A = \begin{pmatrix} 0 & 1 \\ -2 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix}, \quad K_1 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad K_2 = \begin{pmatrix} k & 0 \\ 0 & 0 \end{pmatrix}, \tag{7.2}$$

where $k > 0$ *is a constant. It is easy to verify that for any fixed* $K \in Co\{K_1, K_2\}$, *the matrix* $A + BK$ *is unstable and, therefore, no linear controller* $\mathbf{u} = K\mathbf{x}$ *can stabilize the system. Therefore, we design a switching controller. By Theorem* 7.1 *we must analyze the stability of the switched system* (6.1) *with*

$$\mathbf{f}(\mathbf{x}) = -\begin{pmatrix} 0 & 1 \\ -2 & 1 \end{pmatrix}\mathbf{x}, \qquad \mathbf{h}_k(\mathbf{x}) = -\begin{pmatrix} 0 & 1 \\ -(2+k) & 1 \end{pmatrix}\mathbf{x}.$$

[10] Determined, for example, by the physical limitations of the actuators.

Transforming $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ *to polar coordinates, we get*

$$\begin{pmatrix} \dot{r} \\ \dot{\theta} \end{pmatrix} = \begin{pmatrix} (\cos\theta - \sin\theta) r \sin\theta \\ (\sin\theta - \frac{1}{2}\cos\theta)^2 + \frac{7}{4}\cos^2\theta \end{pmatrix},$$

whereas $\dot{\mathbf{x}} = \mathbf{h}_k(\mathbf{x})$ *becomes*

$$\begin{pmatrix} \dot{r} \\ \dot{\theta} \end{pmatrix} = \begin{pmatrix} ((1+k)\cos\theta - \sin\theta) r \sin\theta \\ (\sin\theta - \frac{1}{2}\cos\theta)^2 + (\frac{7}{4}+k)\cos^2\theta \end{pmatrix}.$$

Clearly, the solutions of both these systems always rotate in a counterclockwise direction ($\dot{\theta} > 0$ *for all* θ) *and, therefore, for all* θ, *we have* $m(\theta) = \max(m_0(\theta), m_1(\theta))$, *where*

$$m_0(\theta) = \frac{((1+k)\cos\theta - \sin\theta)\sin\theta}{(\sin\theta - \frac{1}{2}\cos\theta)^2 + (\frac{7}{4}+k)\cos^2\theta}, \qquad m_1(\theta) = \frac{(\cos\theta - \sin\theta)\sin\theta}{(\sin\theta - \frac{1}{2}\cos\theta)^2 + \frac{7}{4}\cos^2\theta}.$$

It is easily verified that $m_1(\theta) \le m_0(\theta)$ *if and only if* $\tan\theta \ge 0$. *Hence, the WCSL is*

$$\zeta(\theta) = \begin{cases} 0 & \text{if } \tan\theta \ge 0, \\ 1 & \text{otherwise} \end{cases} = \begin{cases} 0 & \text{if } \theta \in [0, \pi/2) \cup [\pi, 3\pi/2), \\ 1 & \text{otherwise} \end{cases}$$

and

$$I(k) = \int_0^{\pi/2} m_0(\theta)d\theta + \int_{\pi/2}^{\pi} m_1(\theta)d\theta + \int_{\pi}^{3\pi/2} m_0(\theta)d\theta + \int_{3\pi/2}^{2\pi} m_1(\theta)d\theta.$$

Computing numerically, we find that the value of k *for which* $I = 0$ *is* $k^* = 6.98513$. *Hence, there exists a switching controller that asymptotically stabilizes* (7.1) *and* (7.2) *if and only if* $k > 6.98513$ *and*

$$\mathbf{u}(\mathbf{x}) = \begin{cases} K_2\mathbf{x} & \text{if } \arctan(x_2/x_1) \in [0, \pi/2) \cup [\pi, 3\pi/2), \\ K_1\mathbf{x} & \text{otherwise} \end{cases} \tag{7.3}$$

is a stabilizing controller.

Figure 6 *depicts the trajectory of the closed-loop system given by* (7.1) *and* (7.2) *with* $k = 10$, *the switching controller* (7.3), *and* $\mathbf{x}_0 = (1, 0)^T$. *As we can see, the system is indeed asymptotically stable.*

8. Summary. We presented a new approach to stability analysis of second-order switched homogeneous systems based on the idea of generalized first integrals. Our approach leads to an explicit Lyapunov function that provides an easily verifiable, *necessary and sufficient* stability condition.

Using our stability analysis, we designed a novel algorithm for constructing a *switching controller* for stabilizing second-order homogeneous systems. The algorithm determines whether the system can be stabilized using switching, and if the answer is affirmative, outputs a suitable controller.

Interesting directions for further research include the complete characterization of the boundary of stability $\partial\Gamma$ and the study of higher-order switched homogeneous systems.

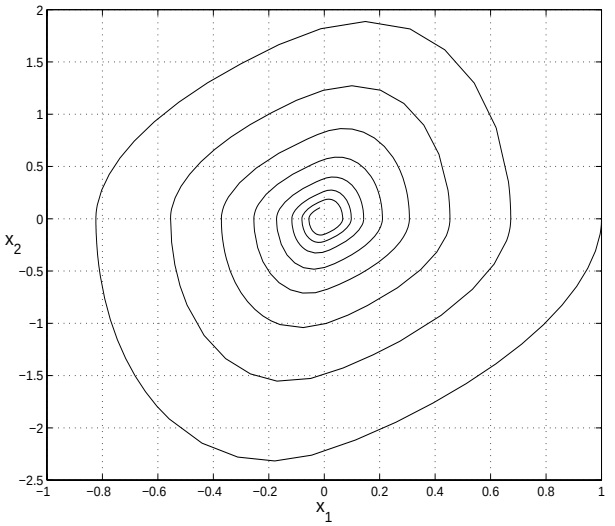


FIG. 6. *Trajectory of the closed-loop system with the switching controller with* $\mathbf{x}_0 = (1,0)^T$.

Appendix.

Proof of Lemma 4.1. The existence of a common Lyapunov function $V'(\mathbf{x})$ follows from Theorem 3.1 in [13] (see also [12]). However, V' is not necessarily homogeneous. Denote $\gamma := \{\mathbf{x} : V'(\mathbf{x}) = 1\}$ so γ is a closed curve encircling the origin. We define a new function $V(\mathbf{x})$ by $V(\mathbf{0}) = 0$ and, for all $\mathbf{x} \neq \mathbf{0}$,

$$V(\mathbf{x}) = k \quad \text{such that} \quad \mathbf{x} \in k\gamma;$$

that is, the contours of V are obtained by scaling γ (see [1]). $V(\mathbf{x})$ is differentiable on $\mathbb{R}^2 \setminus \{\mathbf{0}\}$, positively homogeneous of order one, and radially unbounded.

For any $\mathbf{x} \in \gamma$ we have $\nabla V(\mathbf{x})\mathbf{f}(\mathbf{x}) = \nabla V'(\mathbf{x})\mathbf{f}(\mathbf{x}) < 0$, and using the homogeneity of $V(\mathbf{x})$ and $\mathbf{f}(\mathbf{x})$ this holds for any $\mathbf{x} \in \mathbb{R}^2 \setminus \{\mathbf{0}\}$. Similarly, $\nabla V(\mathbf{x})\mathbf{g}(\mathbf{x}) < 0$ for all $\mathbf{x} \in \mathbb{R}^2 \setminus \{\mathbf{0}\}$. □

Proof of Lemma 5.2. Let $\mathbf{v}(\mathbf{x}) = \frac{\mathbf{f}(\mathbf{x})}{||\mathbf{f}(\mathbf{x})||}$ and $\mathbf{w}(\mathbf{x}) = \frac{(\nabla H^{\mathbf{f}}(\mathbf{x}))^T}{||\nabla H^{\mathbf{f}}(\mathbf{x})||}$. These two vectors form an orthonormal basis of $\mathbb{R}^2$ and, therefore, $\mathbf{g}(\mathbf{x}) = a_1\mathbf{v}(\mathbf{x}) + a_2\mathbf{w}(\mathbf{x})$ and $(\nabla H^{\mathbf{g}}(\mathbf{x}))^T = b_1\mathbf{v}(\mathbf{x}) + b_2\mathbf{w}(\mathbf{x})$, where $a_1 = \mathbf{g}^T(\mathbf{x})\mathbf{v}(\mathbf{x})$, $a_2 = \mathbf{g}^T(\mathbf{x})\mathbf{w}(\mathbf{x})$, $b_1 = \nabla H^{\mathbf{g}}(\mathbf{x})\mathbf{v}(\mathbf{x})$, and $b_2 = \nabla H^{\mathbf{g}}(\mathbf{x})\mathbf{w}(\mathbf{x})$. Now $\nabla H^{\mathbf{g}}(\mathbf{x})\mathbf{g}\mathbf{x} = 0$ yields

$$a_1 b_1 + a_2 b_2 = 0. \tag{8.1}$$

For any $\mathbf{x} \in D$ we have $a_1 > 0$ and since $\nabla H^{\mathbf{f}}(\mathbf{x})$ $(\nabla H^{\mathbf{g}}(\mathbf{x}))$ is orthogonal to $\mathbf{f}(\mathbf{x})$ $(\mathbf{g}(\mathbf{x}))$, we also have $b_2 > 0$. Substituting in (8.1) yields $sgn(a_2) = -sgn(b_1)$, which is just (5.2). □

Proof of Lemma 5.4. The system $\dot{\mathbf{x}} = \mathbf{h}(\mathbf{x})$ is homogeneous and we can represent it in polar coordinates as in (2.1). If $A(\overline{\theta}) = 0$ for some $\overline{\theta} \in [0, 2\pi]$, then the solution corresponding to the WCSL follows the line $l := \theta = \overline{\theta}$. If $R(\overline{\theta}) < 0$, then the solution follows the line l to the origin. However, by the definition of WCSL this is possible only if both the solutions of $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ and $\dot{\mathbf{x}} = \mathbf{g}(\mathbf{x})$ coincide with the line l. Thus, the line l is an invariant set of the system which is a contradiction to

Assumption 1. If $R(\bar{\theta}) \geq 0$, then we get a contradiction of Assumption 2. Hence, $A(\theta) \neq 0$ for all $\theta \in [0, 2\pi]$ and, therefore, there exists $c > 0$ such that $A(\theta) > c$ or $A(\theta) < -c$ for all $\theta \in [0, 2\pi]$. Thus, the solution rotates around the origin. ∎

Proof of Lemma 5.5. Suppose that the WCSL yields a closed trajectory $\tilde{\mathbf{x}}(t)$ that rotates around the origin in a counterclockwise direction ($\dot{\theta} > 0$). Assume that at some point $\mathbf{x}$ along this trajectory, $\lambda(\mathbf{x}) = 1$, that is,

$$\nabla H^{\mathbf{f}}(\mathbf{x})\mathbf{g}(\mathbf{x}) < 0. \tag{8.2}$$

Note that by the definition of the generalized first integral, $\nabla H^{\mathbf{f}}(\mathbf{x})\mathbf{f}(\mathbf{x}) = 0$ for any $\mathbf{x} \in \mathbb{R}^2 \setminus S$. This implies that $\nabla H^{\mathbf{f}}(\mathbf{x}) = k(f_2(\mathbf{x}), -f_1(\mathbf{x}))$ for some $k > 0$, so (8.2) yields

$$f_2(\mathbf{x})g_1(\mathbf{x}) - f_1(\mathbf{x})g_2(\mathbf{x}) < 0. \tag{8.3}$$

Let r, θ be the polar coordinates of $\mathbf{x}$. Since $\tilde{\mathbf{x}}(t)$ rotates around the origin in a counterclockwise direction and satisfies $\dot{\tilde{\mathbf{x}}} = \mathbf{f}(\tilde{\mathbf{x}})$ at $\mathbf{x}$, we have $(-\sin\theta \quad \cos\theta)\mathbf{f}(r, \theta) > 0$. If $(-\sin\theta \quad \cos\theta)\mathbf{g}(r, \theta) < 0$, then $0 \notin F(r, \theta)$ and, therefore, $\zeta(\theta) = 1$. If, on the other hand, $(-\sin\theta \quad \cos\theta)\mathbf{g}(r, \theta) > 0$, then by the definition of ζ (see (5.6)), $\zeta(\theta) = 1$ if and only if

$$\frac{(\cos\theta \quad \sin\theta)\mathbf{f}(r, \theta)}{(-\sin\theta \quad \cos\theta)\mathbf{f}(r, \theta)} > \frac{(\cos\theta \quad \sin\theta)\mathbf{g}(r, \theta)}{(-\sin\theta \quad \cos\theta)\mathbf{g}(r, \theta)}. \tag{8.4}$$

Simplifying, we see that (8.4) is equivalent to $f_1(r, \theta)g_2(r, \theta) - f_2(r, \theta)g_1(r, \theta) > 0$, which is just (8.3), hence, $\zeta(r, \theta) = 1$. Summarizing, we proved that $\lambda(\mathbf{x}) = 1$ if and only if $\zeta(\theta) = 1$. ∎

Acknowledgments. We thank the anonymous reviewers for many helpful comments.

REFERENCES

[1] F. Blanchini, *Set invariance in control*, Automatica, 35 (1999), pp. 1747–1767.
[2] V. D. Blondel and J. N. Tsitsiklis, *A survey of computational complexity results in systems and control*, Automatica, 36 (2000), pp. 1249–1274.
[3] U. Boscain, *Stability of planar switched systems: The linear single input case*, SIAM J. Control Optim., 41 (2002), pp. 89–112.
[4] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.
[5] R. W. Brockett, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millman, and H. J. Sussmann, eds., Birkhäuser, Boston, 1983, pp. 181–191.
[6] W. P. Dayawansa and C. F. Martin, *A converse Lyapunov theorem for a class of dynamical systems which undergo switching*, IEEE Trans. Automat. Control, 44 (1999), pp. 751–760.
[7] A. F. Filippov, *Stability conditions in homogeneous systems with arbitrary regime switching*, Automat. Remote Control, 41 (1980), pp. 1078–1085.
[8] H. Goldstein, *Classical Mechanics*, 2nd ed., Addison–Wesley, Reading, MA, 1980.
[9] W. Hahn, *Stability of Motion*, Springer–Verlag, New York, 1967.
[10] J. H. Hubbard and B. H. West, *Differential Equations: A Dynamical Systems Approach. Higher-Dimensional Systems*, Springer–Verlag, New York, 1995.
[11] D. Liberzon and A. S. Morse, *Basic problems in stability and design of switched systems*, IEEE Control Systems Magazine, 19 (1999), pp. 59–70.
[12] Y. Lin, E. D. Sontag, and Y. Wang, *A smooth converse Lyapunov theorem for robust stability*, SIAM J. Control Optim., 34 (1996), pp. 124–160.

[13] J. L. Mancilla-Aguilar and R. A. Garcia, *A converse Lyapunov theorem for nonlinear switched systems*, Systems Control Lett., 41 (2000), pp. 67–71.

[14] M. Margaliot and G. Langholz, *Necessary and sufficient conditions for absolute stability: The case of second-order systems*, IEEE Trans. Circuits Systems I Fund. Theory Appl., to appear.

[15] A. S. Morse, ed., *Control Using Logic-Based Switching*, Lecture Notes in Control and Inform. Sci. 222, Springer–Verlag, London, 1997.

[16] E. S. Pyatnitskiy and L. B. Rapoport, *Criteria of asymptotic stability of differential inclusions and periodic motions of time-varying nonlinear control systams*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 43 (1996), pp. 219–229.

[17] L. B. Rapoport, *Asymptotic stability and periodic motions of selector-linear differential inclusions*, in Robust Control via Variable Structure and Lyapunov Techniques, Lecture Notes in Control and Inform. Sci. 217, F. Garofalo and L. Glielmo, eds., Springer–Verlag, London, 1996, pp. 269–285.

[18] A. J. van der Schaft and H. Schumacher, *An Introduction to Hybrid Dynamical Systems*, Lecture Notes in Control and Inform. Sci. 251, Springer–Verlag, London, 2000.

[19] M. Vidyasagar, *Nonlinear Systems Analysis*, Prentice–Hall, Upper Saddle River, NJ, 1993.

[20] A. L. Zelentsovsky, *Nonquadratic Lyapunov functions for robust stability analysis of linear uncertain systems*, IEEE Trans. Automat. Control, 39 (1994), pp. 135–138.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 5, pp. 1626–1659

MINIMAX CONTROL OF DISCRETE-TIME STOCHASTIC SYSTEMS*

J. I. GONZÁLEZ-TREJO†, O. HERNÁNDEZ-LERMA‡, AND L. F. HOYOS-REYES†

Abstract. This paper gives a unified, self-contained presentation of minimax control problems for discrete-time stochastic systems on Borel spaces, with possibly unbounded costs. The main results include conditions for the existence of minimax strategies for finite-horizon problems and infinite-horizon discounted and undiscounted (average) cost criteria. The results are specialized to control systems with *unknown* disturbance distributions—also known as *games against nature*. Two examples illustrate the theory, one of them on the *mold level control problem*, which is a key problem in the steelmaking industry.

Key words. minimax control problems, Markov games with complete information, discrete-time stochastic control systems, games against nature

AMS subject classifications. 90C47, 91A25, 93E20

PII. S0363012901383837

1. Introduction.

This paper has several aims. First, it is a survey of results on minimax control problems for discrete-time, stochastic, Markov-like systems in Borel spaces, with possibly unbounded costs. Second, for such systems, it presents a unified, self-contained study that extends virtually all of the results known to date on the minimax control theory, including finite- and infinite-horizon (discounted and undiscounted) problems. It also includes stochastic control problems with unknown disturbance distribution—also known as *games against nature*. And, third, we use the developed theory to analyze a mold level control problem, which is a key problem in the steelmaking industry, and which provided the initial motivation for this paper.

In contrast to the standard optimal control problem, in which there is a *single* decision-maker, in a minimax control problem there are two decision-makers, namely, the controller himself and an "opponent." Thus, in the discrete-time stochastic case that we are concerned with, the system's state process $\{x_t\}$ typically evolves according to a model of the form, say,

$$x_{t+1} = F(x_t, a_t, b_t, \xi_t), \qquad t = 0, 1, \dots, \tag{1.1}$$

where $\pi = \{a_t\}$ and $\gamma = \{b_t\}$ are strategies for the controller and the opponent, and $\{\xi_t\}$ is a sequence of random disturbances. Then if $K(x, \pi, \gamma)$ denotes the system's performance criterion for each initial state $x_0 = x$, the controller's problem is to find a *minimax strategy* π^*, which means that π^* guarantees the best performance in the worst possible situation in the sense that it minimizes

$$K^{\#}(x, \pi) := \sup_{\gamma} K(x, \pi, \gamma) \qquad \forall x \tag{1.2}$$

*Received by the editors January 22, 2001; accepted for publication (in revised form) May 21, 2002; published electronically January 14, 2003.
http://www.siam.org/journals/sicon/41-5/38383.html

†Departamento de Sistemas, Universidad Autónoma Metropolitana–Azcapotzalco, Av. San Pablo No. 180, México D.F. 02200, México (gtji@correo.azc.uam.mx, hrlf@correo.azc.uam.mx).

‡Departamento de Matemáticas, CINVESTAV-IPN, A. Postal 14-740, México D.F. 07000, México. (ohernand@math.cinvestav.mx). The research of this author was performed partially during a visit to the LAAS-CNRS, Toulouse, France, under the auspices of the CONACyT (México)–CNRS (France) Scientific Cooperation Program and was also partially supported by CONACyT grants 32299-E and 37355-E.

over the set of all admissible strategies π. Hence π^* is a "minimax" strategy because it minimizes the maximum expected cost $K^{\#}(x,\cdot)$, where the maximum is taken over all possible strategies γ of the opponent. On the other hand, it follows from the above description that a minimax control problem is a special class of a two-person zero-sum stochastic game. In fact, defining this game's *upper value* as usual, i.e.,

$$U(x) := \inf_{\pi} \sup_{\gamma} K(x,\pi,\gamma) = \inf_{\pi} K^{\#}(x,\pi),$$

we can see that a minimax strategy π^* is precisely one that attains the upper value because

$$K^{\#}(x,\pi^*) = \inf_{\pi} K^{\#}(x,\pi) = U(x) \qquad \forall x. \tag{1.3}$$

A standard application of the minimax approach is to control systems that depend on *unknown* parameters. In this case, the opponent is the "nature," which somehow chooses the unknown parameters at each time t. For example, instead of (1.1) consider the usual, single-controller system

$$x_{t+1} = F(x_t, a_t, \xi_t), \qquad t = 0, 1, \dots, \tag{1.4}$$

and suppose that the disturbances ξ_t are independent random variables with *unknown distributions*. Then a nature's strategy, say $\gamma = \{b_t\}$, would choose a distribution b_t for ξ_t at each time $t = 0, 1, \dots$. Problems of this kind are naturally called *games against nature*.

As far as we can tell, except for some results by Küenle [32, 33, 34] and Kurano [35], there is no *general* theory for minimax control problems in Borel spaces. In fact, all of the previous literature consists either of results specialized from the theory of zero-sum games (which we can argue are not "true" minimax problems—see the next-to-last paragraph in section 2.1 and Remark 2.2) or of minimax problems for particular classes of controlled systems, for instance, queueing and inventory models; see [1, 2, 6, 22, 24, 36] and their references.

Here we propose a general theory for minimax control problems based on the weighted supremum norm approach, which can be traced back (at least) to Wessels [64], and further developed by many authors for different classes of Markov games and control processes, e.g., [2, 12, 13, 14, 17, 18, 19, 23, 25, 31, 32, 33, 34, 35, 42, 46, 48, 64]. This theory is presented in sections 2 to 5. After introducing the basic setup and performance criteria in section 2, in section 3 we first state our main hypothesis (Assumption 3.1), and then we present a full solution (Theorem 3.1) of the finite-horizon, *n-stage problem* ($n = 1, 2, \dots$). In section 4 we consider the infinite-horizon *α-discounted cost* (α-DC), with a "discount factor" α in (0,1). Our main results include the existence of minimax strategies and the approximation of the optimal α-DC by finite-horizon costs (Theorem 4.2), as well as the convergence, in a suitable sense (Definition 4.5), of finite-horizon minimax strategies to an infinite-horizon one (Corollary 4.6). Similar results are presented in section 5 for the *average cost* (AC) problem (Theorem 5.2 and Corollary 5.4).

In section 6 we turn our attention to the games against nature concerning a system of the form (1.4) with unknown disturbance distribution. Here we give conditions that ensure (most of) the requirements in Assumption 3.1, which in turn guarantees that (most of) the results in sections 3 to 5 hold for the games against nature.

Finally, in section 7, we thoroughly analyze two examples, obtaining the corresponding optimal cost functions and minimax strategies. The first one, Example 7.1,

gives a solution to the *mold level control problem*, which essentially consists of controlling a valve to regulate the mold level in a continuous casting machine. From a mathematical viewpoint this problem looks deceptively simple, but it has many critical implications for steelmaking [26, 47]. For instance, from an operational point of view the mold level must be kept constant to avoid molten steel overflows, mold emptying, or strand breakouts, which may cause significant economic losses [9, 28, 30]. From a metallurgical perspective, regulation of the mold is important to avoid slag trapping and steel oxidizing. Moreover, from the quality viewpoint the mold level must be kept constant to obtain a final product that is free of internal and surface cracks [5]. The problem is greatly complicated by many uncertainties and unknown parameters coming into play. For example, measuring the level of molten steel in the mold is not an easy task because the environment is very harsh (over 1600 degrees Celsius), and the molten surface is covered with an insulating powder [62]. In addition, the caster dynamic performance is affected by the material's clogging and unclogging in the valve's nozzle [29, 54]. These complications, among others, naturally induced us to pose the mold level control problem as a minimax problem. To our surprise, the minimax strategy turned out to be an easy-to-compute *myopic* strategy. For other approaches to the mold level control problem, see, for instance, [3, 9, 20, 27] and their references. In the second example, Example 7.2, we study a general, scalar LQ (linear systems, quadratic cost) minimax problem. Although the latter problem is scalar (i.e., the state space is $X = \mathbb{R}$), it clearly shows how one can deal with the vector case.

Before getting into details, a word needs to be said concerning our approach vs. other approaches in the related literature.

Existence and computation of values and minimax strategies. As happens with many mathematical problems, it is one thing to give conditions for a zero-sum game to have a value and/or minimax or maximin strategies (see Remark 2.2 below), and another—sometimes quite different—story to give conditions for the existence and *computation* of the game's value and optimal strategies. For instance, Rieder's [50] Example 4.1 shows a zero-sum game (with a Borel measurable payoff function) for which the value function exists, but it is *not* universally measurable. On the other hand, Küenle [31], Maitra and Sudderth [38], Nowak [43, 44], and other authors give conditions for a zero-sum game to have an either upper or lower semianalytic—hence universally measurable—value function, which is *not* necessarily Borel measurable.

As a consequence, in all of these cases it is, of course, unclear that one can in fact *compute* the value function. Similarly, in, for instance, the latter references one can find conditions for the existence of ε-optimal (for each $\varepsilon > 0$) *universally*—or even *limit*—measurable strategies; and, again, the question of how to *compute* these strategies remains open.

Actually, the existence of the value function and minimax/maximin strategies has been studied in an almost incredible generality. For instance, extending a result by Martin [41] for games on finite sets, Maitra and Sudderth [40] recently proved that a two-person, zero-sum stochastic game with *arbitrary* state and action spaces, a *finitely additive* law of motion, and a bounded Borel measurable payoff has a value. Moreover, the payoff function may depend on the *whole history* of the game (i.e., as in (2.4) with $n = \infty$), and so the main result in [40] is a vast improvement of Theorem 2.1 in [37] on games with a so-called lim sup payoff, which *includes, e.g., discounted and average cost games* (see (2.6) and (2.7) below). A countably additive (as opposed to finitely

additive) version of lim sup games is provided by Maitra and Sudderth in [38]—see also [39]—in which the state and action spaces are Borel spaces, with some additional requirements of compactness and continuity. In the latter case, Maitra and Sudderth give a *transfinite* algorithm for calculating the value; this algorithm, however, may not "terminate" if the state space is not finite.

Another work of extreme generality is Rieder's [51]. He considers zero- and nonzero-sum, *nonstationary* stochastic games, with a *non-Markov* transition law and utility functions that may depend on the *whole story* of the game. Under several different sets of assumptions, for each stage of the nonstationary game he shows the existence of values, the existence of optimal strategies, and the convergence of the value iteration algorithm.

Our paper is at a more mundane level: we consider a standard, time-homogeneous minimax control model with Borel state and action spaces (see (2.1)), and we are interested in the existence of *Borel measurable*—as opposed to, say, semianalytic or universally measurable—value functions and optimal strategies. This "loss" of generality, however, is compensated by the fact that our assumptions are reasonably mild and not hard to verify, and, at the same time, they are sufficiently general to include most of the minimax control problems that appear in applications; and above all, they allow us to obtain computable results, at least in principle. Moreover, replacing our Assumption 3.1 with the standard continuity/compactness conditions (see, e.g., [18, 25, 45, 46, 61]) the results in sections 3, 4, 5 can be extended in the obvious manner to get both minimax and maximin strategies for zero-sum games.

2. Minimax control problems. In this section we introduce the basic components of a minimax control problem, namely, the minimax control model, the sets of admissible strategies, and the performance criteria. We will use the following notation.

REMARK 2.1. *If X is a Borel space (that is, a Borel subset of a complete and separable metric space), its Borel σ-algebra is denoted by $\mathfrak{B}(X)$. Let X and Y be Borel spaces. Then a transition probability (or stochastic kernel) from Y to X is a function $\varphi(D \mid y)$ such that $\varphi(\cdot \mid y)$ is a probability measure on $\mathfrak{B}(X)$ for each $y \in Y$, and $\varphi(D \mid \cdot)$ is a measurable function on Y for each $D \in \mathfrak{B}(X)$. If $Y = X$, then φ is called a* Markov transition probability.

2.1. The minimax control model. As we have already noted, a minimax control problem is in fact a special class of a two-person zero-sum dynamic game. Thus the corresponding minimax control model is (as in [32, 33, 34, 35]) of the form

$$MCM := (X, A, B, \mathbb{K}_A, \mathbb{K}, Q, c), \tag{2.1}$$

where X is the *state space*, A is the *controller's* (player 1) *action space*, and B is the *opponent's* (player 2) *action space*. These spaces are all assumed to be Borel spaces. In addition we have the following:

(a) $\mathbb{K}_A \in \mathfrak{B}(X \times A)$ is the *constraint set for the controller*. That is, for each state $x \in X$, the x-section

$$A(x) := \{a \in A \mid (x, a) \in \mathbb{K}_A\} \tag{2.2}$$

represents the set of admissible actions for the controller in the state x. We assume that $\mathbb{K}_A$ contains the graph of a measurable function from X to A. (This will indeed be the case under Assumption 3.1 below.)

(b) $\mathbb{K} \in \mathfrak{B}(X \times A \times B)$ is the *constraint set for the opponent*, so that for each pair (x, a) in $\mathbb{K}_A$ the (x, a)-section

$$B(x, a) := \{b \in B \mid (x, a, b) \in \mathbb{K}\} \tag{2.3}$$

is the set of admissible actions for the opponent when the state is $x \in X$ and the controller uses the action $a \in A(x)$. We suppose that $\mathbb{K}$ contains the graph of a measurable map from $\mathbb{K}_A$ to B. (This, again, will be the case under Assumption 3.1.)

(c) Q or, more explicitly, $Q(D \mid x, a, b)$, denotes the *transition law*, a stochastic kernel from $\mathbb{K}$ to X.

(d) $c : \mathbb{K} \to \mathbb{R}$ stands for the *cost-per-stage* function.

The above minimax control model is also known as a *Markov* (or *stochastic*) *game with complete information* [32, 33, 34]. If the opponent's action set $B(x, a)$ in (2.3) does not depend on $a \in A(x)$, we then obtain the usual zero-sum Markov game in which the players choose their actions $a \in A(x)$ and $b \in B(x)$ independently.

The minimax control model represents a dynamic game that evolves in discrete time ($n = 0, 1, \dots$) as follows: if the state at time n is $x_n = x \in X$, then the controller chooses an action $a_n = a$ in $A(x)$, and the opponent chooses an action $b_n = b$ in $B(x, a)$. As a consequence of this, two things happen: (1) the controller pays $c(x, a, b)$ to the opponent, and (2) the system moves to a new state $x_{n+1} \in X$ with probability $Q(\cdot \mid x, a, b)$.

2.2. Strategies. Let $H_0 := X$, $H_0^\# := \mathbb{K}_A$, and for $n \geq 1$ let $H_n := \mathbb{K}^n \times X$ and $H^\# := \mathbb{K}^n \times \mathbb{K}_A$. Generic elements of H_n and $H_n^\#$ are "histories" of the form

$$h_n = (x_0, a_0, b_0, \dots, x_{n-1}, a_{n-1}, b_{n-1}, x_n) \qquad \text{and} \qquad h_n^\# = (h_n, a_n), \tag{2.4}$$

respectively.

A *strategy for the controller* is a sequence $\pi = \{\pi_n\}$ of stochastic kernels from H_n to A that satisfy the constraint

$$\pi_n(A(x_n) \mid h_n) = 1 \qquad \forall h_n \in H_n \qquad \text{and} \qquad n = 0, 1, \dots .$$

We shall denote by Π the set of all the strategies (or control policies) for the controller.

A *strategy for the opponent* is a sequence $\gamma = \{\gamma_n\}$ of stochastic kernels γ_n from $H_n^\#$ to B such that $\gamma_n(B(x_n, a_n) \mid h_n^\#) = 1$ for all $h_n^\# \in H_n^\#$ and $n = 0, 1, \dots$. The set of all these strategies is denoted by Γ.

DEFINITION 2.1. *$\mathbb{F}_A$ denotes the set of all measurable functions $f : X \to A$ such that $f(x)$ is in $A(x)$ for all $x \in X$, and $\mathbb{F}_B$ stands for the set of measurable functions g from $X \times A$ to B such that $g(x, a)$ is in $B(x, a)$ for all $(x, a) \in \mathbb{K}_A$.*

A strategy $\pi = \{\pi_n\}$ for the controller is said to be a *Markov strategy* if there is a sequence of functions $f_n \in \mathbb{F}_A$ such that $\pi_n(\cdot \mid h_n)$ is concentrated at $f_n(x_n)$ for all $n = 0, 1, \dots$. In this case we shall identify π_n with f_n. Moreover, if $\pi = \{f_n\}$ is a Markov strategy and $f_n = f_0$ for all $n \geq 0$, then π is called a *stationary strategy*, and we shall identify π with $f_0 \in \mathbb{F}_A$. In other words, $\mathbb{F}_A$ will be identified with the family of stationary strategies for the controller.

The Markov and stationary strategies for the opponent are defined similarly, replacing $f_n \in \mathbb{F}_A$ with $g_n \in \mathbb{F}_B$.

Let $(\Omega, \mathfrak{F})$ be the (canonical) measurable space consisting of the sample space $\Omega := (X \times A \times B)^\infty$ and its product σ-algebra $\mathfrak{F}$. Then, by a theorem of Ionescu-Tulcea [4, 32] for each pair of strategies $\pi \in \Pi$ and $\gamma \in \Gamma$, and each initial state $x \in X$,

there is a probability measure $\mathbb{P}_x^{\pi,\gamma}$ and a stochastic process $\{(x_t, a_t, b_t), t = 0, 1, \dots\}$ defined on $(\Omega, \mathfrak{F})$ in a canonical way, where x_t, a_t, and b_t represent the state and the actions of the controller and the opponent, respectively, at each time $t = 0, 1, \dots$. The expectation operator with respect to $\mathbb{P}_x^{\pi,\gamma}$ is denoted by $\mathbb{E}_x^{\pi,\gamma}$.

2.3. Performance criteria. Let α be a given positive number. We shall consider three performance criteria, each depending on the initial state $x_0 = x \in X$, and strategies $\pi \in \Pi$ for the controller and $\gamma \in \Gamma$ for the opponent.

- The finite-horizon *n-stage cost* ($n = 1, 2, \dots$):

$$V_{n,\alpha}(x, \pi, \gamma) := \mathbb{E}_x^{\pi,\gamma}\left[\sum_{t=0}^{n-1} \alpha^t c(x_t, a_t, b_t)\right]. \tag{2.5}$$

  For $\alpha = 1$ we shall write $V_{n,1}(x, \pi, \gamma) \equiv J_n(x, \pi, \gamma)$.
- The infinite-horizon *α-discounted cost* (α-DC) with $0 < \alpha < 1$:

$$V_\alpha(x, \pi, \gamma) := \mathbb{E}_x^{\pi,\gamma}\left[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t, b_t)\right]. \tag{2.6}$$

- The infinite-horizon expected *average cost* (AC):

$$J(x, \pi, \gamma) := \limsup_{n\to\infty} \frac{1}{n} J_n(x, \pi, \gamma). \tag{2.7}$$

DEFINITION 2.2. *Let $K(x, \pi, \gamma)$ be any of the cost functions in* (2.5)–(2.7)*, and let*

$$K^{\#}(x, \pi) := \sup_{\gamma\in\Gamma} K(x, \pi, \gamma). \tag{2.8}$$

A strategy $\pi^* \in \Pi$ (for the controller) is said to be a minimax strategy with respect to the cost function K if π^* minimizes $K^{\#}(x, \cdot)$ over Π for all $x \in X$, that is,

$$K^{\#}(x, \pi^*) = \inf_{\pi\in\Pi} K^{\#}(x, \pi) = \inf_{\pi\in\Pi} \sup_{\gamma\in\Gamma} K(x, \pi, \gamma) \qquad \forall x \in X. \tag{2.9}$$

REMARK 2.2. *In game-theoretic terminology the function on the right-hand side of* (2.9)*, i.e.,*

$$U(x) := \inf_{\pi\in\Pi} \sup_{\gamma\in\Gamma} K(x, \pi, \gamma), \tag{2.10}$$

is called the zero-sum game's upper value *(with respect to K). Thus a minimax strategy is one that attains the game's upper value. On the other hand, one could consider the "dual,"* maximin problem *in which* (2.8) *is replaced with*

$$K_{\#}(x, \gamma) := \inf_{\pi\in\Pi} K(x, \pi, \gamma).$$

Then the game's lower value *(with respect to K) is*

$$L(x) := \sup_{\gamma\in\Gamma} K_{\#}(x, \gamma) = \sup_{\gamma\in\Gamma} \inf_{\pi\in\Pi} K(x, \pi, \gamma),$$

and $\gamma^* \in \Gamma$ is called a maximin strategy *if*

$$K_{\#}(x, \gamma^*) = L(x) \qquad \forall x \in X. \tag{2.11}$$

In general $L(\cdot) \le U(\cdot)$, and if the equality holds, then $L(\cdot) = U(\cdot)$ is called the game's value function. If in addition π^* and γ^* satisfy (2.9) *and* (2.11)*, then the pair (π^*, γ^*) is said to be a* saddle point *or a* noncooperative equilibrium.

3. Finite-horizon minimax problems. The following assumption is supposed to hold throughout the remainder of the paper. Except for parts (f) and (g) in this assumption, all of the conditions concern the usual continuity/compactness hypotheses for Markov control processes and Markov games [1, 2, 10, 11, 12, 17, 18, 23, 31, 32, 33, 34, 35, 45, 46, 48, 51, 52, 55, 56, 61, 64].

Assumption 3.1. Let MCM be as in (2.1).

(a) $c(x, a, b)$ is lower semicontinuous (l.s.c.) on $\mathbb{K}$.

(b) There is a constant $\bar{c} \geq 0$ and a measurable function $w(\cdot) \geq 1$ on X such that

$$|c(x, a, b)| \leq \bar{c} w(x) \qquad \forall (x, a, b) \in \mathbb{K}. \tag{3.1}$$

(c) The transition law Q is weakly continuous, that is, for each continuous bounded function $u : X \to \mathbb{R}$, the function

$$\hat{u}(x, a, b) := \int_X u(y) Q(dy|x, a, b) \tag{3.2}$$

is continuous on $\mathbb{K}$.

(d) The function $w(\cdot) \geq 1$ in (b) as well as $\hat{w}(x, a, b) := \int w(y) Q(dy|x, a, b)$ are continuous on X and $\mathbb{K}$, respectively. (See Remark 3.1(a).)

(e) There is a constant $\beta > 0$ such that $\hat{w}(x, a, b) \leq \beta w(x)$ for all (x, a, b) in $\mathbb{K}$.

(f) The set $A(x) \subset A$ in (2.2) is compact for each $x \in X$, and in addition the set-valued mapping $x \mapsto A(x)$ is upper semicontinuous (u.s.c.), that is, if $x^k \to x$ and $a^k \in A(x^k)$ is such that $a^k \to a$, then a is in $A(x)$. (See Remark 3.1(c).)

(g) The set $B(x, a)$ in (2.3) is σ-compact for each $(x, a) \in \mathbb{K}_A$, and, moreover, the set-valued mapping $(x, a) \mapsto B(x, a)$ is l.s.c., that is, if $(x^k, a^k) \in \mathbb{K}_A$ converges to $(x, a) \in \mathbb{K}_A$ and b is in $B(x, a)$, then there exists b^k in $B(x^k, a^k)$ such that $b^k \to b$.

Before stating our optimality result for the finite-horizon criterion (2.5), we shall make some comments on Assumption 3.1 and introduce some useful concepts.

REMARK 3.1. (a) *If the cost-per-stage* $c(x, a, b)$ *is* bounded below, *then part* (d) *in Assumption* 3.1 *is* not *required. (Indeed, if* $c(x, a, b) \geq -k$ *for all* (x, a, b) *in* $\mathbb{K}$ *and some constant* k, *then* $\hat{c}(x, a, b) := c(x, a, b) + k$ *is nonnegative, and so the fact that Assumption* 3.1(d) *is not required follows from the proof of Lemma* 3.2(b), *below. Observe that replacing* c *with* $\hat{c}$ *in* (2.5)–(2.7) *does not essentially alter the corresponding minimax control problem.) Furthermore, if* c *is* bounded *(above and below), then the "weight" or "bounding" function* $w(\cdot)$ *can be taken as a* constant, *for instance,* $w(\cdot) \equiv 1$, *and so Assumption* 3.1(e) *can be omitted.*

(b) *Consider the dynamic model* (1.1) *and suppose that* $\{\xi_t\}$ *is a sequence of independently and identically distributed (i.i.d.) random variables in a Borel space* S. *Let* μ *be the common probability distribution of the* ξ_t. *Then the function* $\hat{u}$ *in* (3.2) *becomes*

$$\hat{u}(x, a, b) = \mathbb{E}\left[u(x_{t+1}) \mid (x_t, a_t, b_t) = (x, a, b)\right] = \int_S u[F(x, a, b, s)]\mu(ds). \tag{3.3}$$

Therefore, by the dominated convergence theorem, Assumption 3.1(c) *holds if the measurable function* $F : \mathbb{K} \times S \to X$ *is continuous in* $(x, a, b) \in \mathbb{K}$ *for each* $s \in S$.

(c) *A set-valued mapping is said to be* continuous *if it is l.s.c. and u.s.c. Thus the semicontinuity requirements in Assumption* 3.1(f) *and* (g) *will be trivially verified in many cases because the mappings* $x \mapsto A(x)$ *and* $(x, a) \mapsto B(x, a)$ *are continuous*

in most applications (see (7.6) *and* (7.18)*, for instance). In fact, it is not uncommon to find situations in which they are* constant*—hence continuous—mappings, that is, $A(x) = A$ for all $x \in X$ or $B(x,a) = B$ for all $(x,a) \in \mathbb{K}_A$.*

Let $w(\cdot) \geq 1$ be as in Assumption 3.1*. We shall denote by $\mathbb{B}_w(X)$ the Banach space of measurable functions u on X that have a finite w-norm, $\|u\|_w < \infty$, which is defined as*

$$\|u\|_w := \sup_{x \in X} [|u(x)|/w(x)]. \tag{3.4}$$

For each $u \in \mathbb{B}_w(X)$, $0 \leq \alpha \leq 1$, and $(x,a,b) \in \mathbb{K}$ let

$$H_\alpha(u;x,a,b) := c(x,a,b) + \alpha \int_X u(y) Q(dy|x,a,b), \tag{3.5}$$

$$H_\alpha^\#(u;x,a) := \sup_{b \in B(x,a)} H_\alpha(u;x,a,b), \tag{3.6}$$

$$T_\alpha u(x) := \inf_{a \in A(x)} H_\alpha^\#(u;x,a). \tag{3.7}$$

More explicitly, we can write (3.7) as

$$T_\alpha u(x) = \inf_{a \in A(x)} \sup_{b \in B(x,a)} \left[c(x,a,b) + \alpha \int_X u(y) Q(dy|x,a,b) \right]. \tag{3.8}$$

The operator T_α is usually referred to as the *dynamic programming* (DP) *operator.*

We shall denote by $\mathbb{B}_{lsc}$ the family of l.s.c. functions in $\mathbb{B}_w(X)$.

THEOREM 3.1. *Fix $\alpha > 0$, and define on X the functions*

$$v_{n,\alpha} := T_\alpha v_{n-1,\alpha} = T_\alpha^n v_{0,\alpha} \qquad \forall n = 1,2,\dots, \tag{3.9}$$

with $v_{0,\alpha}(\cdot) \equiv 0$. Then for each $n = 1,2,\dots,$

(a) *$v_{n,\alpha}$ is in $\mathbb{B}_{lsc}$;*

(b) *there exists $f_n \in \mathbb{F}_A$ such that*

$$v_{n,\alpha}(x) = H_\alpha^\#(v_{n-1,\alpha};x,f_n(x)) \qquad \forall x \in X,$$

i.e.,

$$v_{n,\alpha}(x) = \sup_{b \in B(x,f_n(x))} \left[c(x,f_n(x),b) + \alpha \int_X v_{n-1,\alpha}(y) Q(dy|x,f_n(x),b) \right], \tag{3.10}$$

and, moreover,

$$v_{n,\alpha}(x) = \min_{a \in A(x)} \sup_{b \in B(x,a)} \left[c(x,a,b) + \alpha \int_X v_{n-1,\alpha}(y) Q(dy|x,a,b) \right]; \tag{3.11}$$

(c) *$v_{n,\alpha}$ is the optimal n-stage cost, that is, from* (2.5) *and* (2.9)*,*

$$v_{n,\alpha}(x) = \inf_{\pi \in \Pi} \sup_{\gamma \in \Gamma} V_{n,\alpha}(x,\pi,\gamma) \qquad \forall x \in X;$$

(d) *the Markov policy* $\pi^n := \{f_n, f_{n-1}, \dots, f_1\}$ *is a minimax strategy for the* n*-stage problem; that is, by* (c) *and Definition* 2.2,

$$v_{n,\alpha}(x) = \sup_{\gamma\in\Gamma} V_{n,\alpha}(x, \pi^n, \gamma) \qquad \forall x \in X.$$

To prove Theorem 3.1 we shall state first some useful preliminary results. Observe in particular that the following lemma gives the existence of minimax strategies for a *single-stage minimax problem* with cost function $v(x, a, b)$, say.

LEMMA 3.2. *Let* $v : \mathbb{K} \to \mathbb{R}$ *be an l.s.c. function such that*

$$|v(x, a, b)| \le \bar{v}w(x) \qquad \forall (x, a, b) \in \mathbb{K} \tag{3.12}$$

for some constant $\bar{v} \ge 0$. *Let*

$$v^{\#}(x, a) := \sup_{b\in B(x,a)} v(x, a, b) \qquad \textit{and} \qquad v^*(x) := \inf_{a\in A(x)} v^{\#}(x, a).$$

Then

(a) $v^{\#}$ *is l.s.c. on* $\mathbb{K}_A$ *and satisfies that*

$$|v^{\#}(x, a)| \le \bar{v}w(x) \qquad \forall (x, a) \in \mathbb{K}_A; \tag{3.13}$$

(b) v^* *is in* $\mathbb{B}_{lsc}$ *and there exists* $f \in \mathbb{F}_A$ *such that*

$$v^*(x) = \min_{a\in A(x)} v^{\#}(x, a) = v^{\#}(x, f(x)) \qquad \forall x \in X, \tag{3.14}$$

that is, by definition of $v^{\#}$,

$$v^*(x) = \sup_{b\in B(x,f(x))} v(x, f(x), b) \qquad \forall x \in X.$$

Proof. (a) The inequality (3.13) obviously follows from (3.12). Now, to prove that $v^{\#}$ is l.s.c., let $(x^k, a^k) \in \mathbb{K}_A$ be a sequence converging to $(x, a) \in \mathbb{K}_A$. Choose an arbitrary $b \in B(x, a)$. Then, by Assumption 3.1(g), there exists $b^k \in B(x^k, a^k)$ such that $b^k \to b$. Therefore, as $v^{\#}(x^k, a^k) \ge v(x^k, a^k, b^k)$ and v is l.s.c., we have

$$\liminf_{k\to\infty} v^{\#}(x^k, a^k) \ge v(x, a, b).$$

Thus, as $b \in B(x, a)$ was arbitrary, we obtain $\liminf v^{\#}(x^k, a^k) \ge v^{\#}(x, a)$; that is, $v^{\#}$ is l.s.c.

(b) By (a) and the continuity of $w(\cdot)$ (Assumption 3.1(d)), the function $v^{\#}(x, a) + \bar{v}w(x)$ is *nonnegative* and l.s.c. on $\mathbb{K}_A$. Combining this fact with Assumption 3.1(f), a well-known result of Schäl [55] (reproduced in [16, Proposition D.5]) yields that the function

$$\inf_{a\in A(x)} [v^{\#}(x, a) + \bar{v}w(x)] \tag{3.15}$$

is l.s.c. and that there exists $f \in \mathbb{F}_A$ that realizes the minimum in (3.15), i.e.,

$$\min_{a\in A(x)} [v^{\#}(x, a) + \bar{v}w(x)] = v^{\#}(x, f(x)) + \bar{v}w(x) \qquad \forall x \in X.$$

This gives (3.14). Finally, (3.13) gives that $|v^*(\cdot)| \le \bar{v}w(\cdot)$, and so we conclude that v^* is in $\mathbb{B}_{lsc}$. □

In part (b) of the following lemma we replace the function v in (3.12) with the function $H_\alpha(u;\cdot)$ in (3.5), with u in $\mathbb{B}_{lsc}$. This yields in particular that *the DP operator T_α maps $\mathbb{B}_{lsc}$ into itself.*

LEMMA 3.3. *Let u be an arbitrary function in $\mathbb{B}_{lsc}$, and let $\hat{u}$ and $T_\alpha u$ be as in* (3.2) *and* (3.7), (3.8)*, respectively. Then*

(a) *$\hat{u}$ is l.s.c. on $\mathbb{K}$, and, therefore,*

(b) *$H_\alpha(u;x,a,b)$ is l.s.c. on $\mathbb{K}$, and*

$$|H_\alpha(u;x,a,b)| \le [\bar{c} + \alpha\beta\|u\|_w]w(x) \qquad \forall (x,a,b) \in \mathbb{K}; \tag{3.16}$$

(c) *$T_\alpha u$ is in $\mathbb{B}_{lsc}$, and, furthermore, there exists $f \in \mathbb{F}_A$ such that, for all $x \in X$,*

$$T_\alpha u(x) = \sup_{b \in B(x,f(x))} \left[c(x,f(x),b) + \alpha \int_X u(y) Q(dy|x,f(x),b) \right], \tag{3.17}$$

and (3.8) *holds with "min" in lieu of "inf," i.e.,*

$$T_\alpha u(x) = \min_{a \in A(x)} \sup_{b \in B(x,f(x))} \left[c(x,a,b) + \alpha \int_X u(y) Q(dy|x,a,b) \right]. \tag{3.18}$$

Proof. (a) Let u be in $\mathbb{B}_{lsc}$, and let us assume for a moment that u is *nonnegative*. Then there exists a sequence $\{u^n\}$ of continuous bounded functions such that $u^n \uparrow u$ pointwise. Hence, as $\int u(y)Q(dy|x,a,b) \ge \int u^n(y)Q(dy|x,a,b)$ for all n, if $(x^k,a^k,b^k) \to (x,a,b)$ we get from Assumption 3.1(c) that

$$\liminf_{k\to\infty} \int_X u(y)Q(dy|x^k,a^k,b^k) \ge \int_X u^n(y)Q(dy|x,a,b) \qquad \forall n.$$

Thus, letting $n \to \infty$ we conclude that $\hat{u}$ is l.s.c. on $\mathbb{K}$, when u is nonnegative. Consider now an *arbitrary* function u in $\mathbb{B}_{lsc}$. Then, by (3.4) and the continuity of $w(\cdot)$, the function $u(\cdot) + \|u\|_w w(\cdot)$ is l.s.c. and nonnegative. Therefore, by the result in the nonnegative case and the continuity of $\hat{w}$ (Assumption 3.1(d)), it follows that $\hat{u}$ is l.s.c.

(b) By (a) and Assumption 3.1(a), the function $H_\alpha(u;\cdot)$ is l.s.c. on $\mathbb{K}$. On the other hand, from (3.4) and Assumption 3.1(e),

$$\int |u(y)|Q(dy|x,a,b) \le \|u\|_w \int w(y)Q(dy|x,a,b) \le \|u\|_w \beta w(x) \tag{3.19}$$

for all (x,a,b) in $\mathbb{K}$. From the latter inequality together with (3.1) and (3.5) we get (3.16).

(c) This follows from (b) and Lemma 3.2(b) with $v(\cdot) = H_\alpha(u;\cdot)$. □

Using Lemma 3.3 we can now easily prove Theorem 3.1.

Proof of Theorem 3.1. Part (a) can be obtained by induction. As $v_{0,\alpha}(\cdot) \equiv 0$, (a) trivially holds for $n = 0$. Suppose now that $v_{n-1,\alpha}$ is in $\mathbb{B}_{lsc}$ for some $n \ge 1$. Then (3.9) and Lemma 3.3(c) yield that $v_{n,\alpha}$ is in $\mathbb{B}_{lsc}$, and (a) follows. In turn, (a) and Lemma 3.3(c) again yield (b). Finally, (c) and (d) follow from (b) and standard dynamic programming arguments (see, for instance, Theorem 3.2.1 in [16]). □

4. Infinite-horizon discount cost. In this section we consider the α-DC in (2.6), with $0 < \alpha < 1$. The clue to our optimality result is provided by the following fact due to Küenle [32, 33].

THEOREM 4.1. *Suppose that there exists a measurable function* $v^* : X \to \mathbb{R}$ *and a stationary policy* $f^* \in \mathbb{F}_A$ *such that, for all* $x \in X$,

$$v^*(x) = \sup_{b \in B(x, f^*(x))} \left[c(x, f^*(x), b) + \alpha \int_X v^*(y) Q(dy|x, f^*(x), b) \right] \tag{4.1}$$

and

$$\lim_{n \to \infty} \alpha^n \mathbb{E}_\alpha^{\pi,\gamma}[v^*(x_n)] = 0 \qquad \forall \pi \in \Pi, \gamma \in \Gamma, x \in X. \tag{4.2}$$

Then

$$v^*(x) = \sup_{\gamma \in \Gamma} V_\alpha(x, f^*, \gamma) \qquad \forall x \in X. \tag{4.3}$$

If in addition v^* *satisfies that (with* T_α *as in* (3.7), (3.8)*)*

$$v^* = T_\alpha v^*, \tag{4.4}$$

then v^* *is the optimal* α*-DC function, and* f^* *is an* α*-DC minimax strategy, that is,* (4.3) *holds and also*

$$v^*(x) = \inf_{\pi \in \Pi} \sup_{\gamma \in \Gamma} V_\alpha(x, \pi, \gamma) \qquad \forall x \in X. \tag{4.5}$$

To obtain (4.1), (4.2), and (4.4) it suffices to impose an additional condition on Assumption 3.1. The precise result is as follows. (Recall that $\mathbb{B}_{lsc}$ denotes the family of l.s.c. functions in $\mathbb{B}_w(X)$.)

THEOREM 4.2. *Suppose that Assumption* 3.1 *holds but, in addition, the constant* β *in part* (e) *satisfies that*

$$1 \le \beta < 1/\alpha. \tag{4.6}$$

Then there exist a function $v^* : X \to \mathbb{R}$ *and a stationary policy* $f^* \to \mathbb{F}_A$ *such that the following hold:*

(a) v^**is the unique function in* $\mathbb{B}_{lsc}$ *that satisfies* (4.4). *Moreover, it satisfies* (4.2), *and*

$$\|v_{n,\alpha} - v^*\|_w \le \bar{c}(\alpha\beta)^n/(1 - \alpha\beta) \qquad \forall n \ge 0, \tag{4.7}$$

with $\bar{c}$ *and* $v_{n,\alpha}$ *as in* (3.1) *and* (3.9), *respectively, with* $v_{0,\alpha} \equiv 0$.

(b) v^* *and* f^* *satisfy* (4.1).

Hence, v^* *and* f^* *satisfy the conclusions of Theorem* 4.1.

To prove Theorem 4.2 we will first state a few general results, some of which are in fact well known, but we include them here for completeness and ease of reference. After the proof of the theorem we state an interesting consequence (Corollary 4.6) of (4.7).

LEMMA 4.3. (a) *If* $\{v_n\}$ *is a sequence in* $\mathbb{B}_{lsc}$ *and* $\|v_n - v\|_w \to 0$, *then* v *is in* $\mathbb{B}_{lsc}$. *Hence,* (b) $\mathbb{B}_{lsc}$ *is a* complete *subset of* $\mathbb{B}_w(X)$, *that is, if* $\{v_n\} \subset \mathbb{B}_{lsc}$ *is a Cauchy sequence in* w*-norm, then it converges in* w*-norm to some function in* $\mathbb{B}_{lsc}$.

Proof. (a) The proof follows from the inequality $v(\cdot) \ge -\|v_n - v\|_w w(\cdot) + v_n(\cdot)$ and the continuity of w (Assumption 3.1(d)). Part (b) follows from (a) and the fact that $\mathbb{B}_w(X)$ is a Banach space. □

LEMMA 4.4. *Let $\mathbb{B}_0$ be a complete subset of $\mathbb{B}_w(X)$, and let T be a mapping from $\mathbb{B}_0$ into itself. Suppose that* (i) *T is monotone (i.e., $u \le u'$ implies $Tu \le Tu'$), and* (ii) *there is a constant $0 < \tau < 1$ such that*

$$T(u + rw) \le Tu + \tau rw \qquad \forall u \in \mathbb{B}_0 \qquad \text{and} \qquad r \ge 0. \tag{4.8}$$

Then T is a contraction mapping with modulus τ, i.e.,

$$\|Tu - Tu'\|_w \le \tau \|u - u'\|_w \qquad \forall u, u' \in \mathbb{B}_0, \tag{4.9}$$

and, therefore, there is a unique function u^* in $\mathbb{B}_0$ that satisfies

$$\text{(a)} \quad u^* = Tu^* \quad \text{and} \quad \text{(b)} \quad \|T^n u - u^*\|_w \le \tau^n \|u - u^*\|_w \quad \forall u \in \mathbb{B}_0, n \ge 0. \tag{4.10}$$

Further, if there is a constant $k \ge 0$ such that

$$\|Tu\|_w \le k + \tau \|u\|_w \qquad \forall u \in \mathbb{B}_0, \tag{4.11}$$

then

$$\|u^*\|_w \le k/(1 - \tau). \tag{4.12}$$

Proof. By (3.4), for any two functions u and u' in $\mathbb{B}_w(X)$ we have $u \le u' + \|u - u'\|_w w$. Hence, by (i) and (4.8) with $r := \|u - u'\|_w$, we obtain

$$Tu \le Tu' + \tau \|u - u'\|_w w,$$

so that $Tu - Tu' \le \tau \|u - u'\|_w w$. Similarly, $Tu - Tu' \ge -\tau \|u - u'\|_w w$, and (4.9) follows. Finally, (4.10) follows from Banach's fixed point theorem, whereas (4.12) follows from (4.11) and (4.10)(a). □

Lemma 4.4 is essentially the same as Proposition 7.2.9 in [17, p. 6], but the latter is *incorrectly stated* (!): it requires $r \in \mathbb{R}$, rather than $r \ge 0$ as in (4.8).

Finally, before passing to the proof of Theorem 4.2 we should mention that (4.2) holds for all u in $\mathbb{B}_w$, $\pi \in \Pi$, $\gamma \in \Gamma$, *and* $x \in X$, i.e.,

$$\lim_{n \to \infty} \alpha^n \mathbb{E}_x^{\pi,\gamma}[u(x_n)] = 0. \tag{4.13}$$

Indeed, using Assumption 3.1(e), a straightforward calculation gives

$$\mathbb{E}_x^{\pi,\gamma}[w(x_{n+1})] \le \beta \mathbb{E}_x^{\pi,\gamma}[w(x_n)] \qquad \forall n = 0, 1, \dots,$$

and so

$$\mathbb{E}_x^{\pi,\gamma}[w(x_n)] \le \beta^n w(x). \tag{4.14}$$

Hence, by (4.6),

$$\lim_{n \to \infty} \alpha^n \mathbb{E}_x^{\pi,\gamma}[w(x_n)] = 0.$$

This implies (4.13) because, by (3.4),

$$\mathbb{E}_x^{\pi,\gamma}[u(x_n)] \le \|u\|_w \mathbb{E}_x^{\pi,\gamma}[w(x_n)] \qquad \forall u \in \mathbb{B}_w(X). \tag{4.15}$$

Proof of Theorem 4.2. (a) We shall first use Lemma 4.4 to show that T_α is a contraction mapping on $\mathbb{B}_0 := \mathbb{B}_{lsc}$ with modulus $\tau := \alpha\beta < 1$, that is,

$$\|T_\alpha u - T_\alpha u'\|_w \le \tau \|u - u'\|_w \qquad \forall u, u' \in \mathbb{B}_{lsc}. \tag{4.16}$$

(Observe that T_α may not map all of $\mathbb{B}_w(X)$ into itself because for an arbitrary function u in $\mathbb{B}_w(X)$, the function $T_\alpha u$ is not necessarily measurable.) Now, by Lemma 4.3, $\mathbb{B}_{lsc}$ is a complete subset of $\mathbb{B}_w(X)$, whereas by Lemma 3.3, T_α maps $\mathbb{B}_{lsc}$ into itself. On the other hand, it is evident that T_α is monotone. Thus, to obtain (4.16) it suffices to verify that T_α satisfies (4.8). To do this, note that Assumption 3.1(e) yields

$$\sup_{b \in B(x,a)} \int_X w(y) Q(dy|x,a,b) \le \beta w(x) \qquad \forall (x,a) \in \mathbb{K}_A. \tag{4.17}$$

This inequality and (3.8) give (4.8) for $T = T_\alpha$ and $\tau := \alpha\beta$, and so (4.16) follows. Hence, there is a unique function v^* in $\mathbb{B}_{lsc}$ that satisfies (4.4), and, moreover (taking $u = v_{0,\alpha}$ in (4.10)(b)),

$$\|v_{n,\alpha} - v^*\|_w \le \tau^n \|v^*\|_w \qquad \forall n \ge 0, \quad \text{with } \tau = \alpha\beta. \tag{4.18}$$

Finally, observe that (3.16) gives

$$\|T_\alpha u\|_w \le \bar{c} + \tau \|u\|_w \qquad \forall u \in \mathbb{B}_{lsc},$$

which in turn, by (4.12), gives

$$\|v^*\|_w \le \bar{c}/(1-\tau). \tag{4.19}$$

From the latter inequality and (4.18) we obtain (4.7). Therefore, as v^* satisfies (4.2) (by (4.13)), the proof of part (a) is complete.

(b) The proof follows from (a) and Lemma 3.3(c). □

It is worth noting that (4.7) yields the *geometric convergence*, in the w-norm, of the so-called *value iteration* (or *successive approximations*) procedure $T_\alpha^n v_{0,\alpha} \to v^*$, where $v_{0,\alpha}(\cdot) \equiv 0$. Another noteworthy fact is Corollary 4.6 below, in which we use the following concept.

DEFINITION 4.5. *A sequence of stationary strategies $\{f_n\} \subset \mathbb{F}_A$ is said to converge in the sense of Schäl* [55] *if there exist $f_* \in \mathbb{F}_A$ such that $f_*(x) \in A(x)$ is an accumulation point of $\{f_n(x)\} \subset A(x)$ for each $x \in X$; that is, for each $x \in X$ there is a subsequence $\{n_i(x)\}$ of $\{n\}$ such that*

$$f_{n_i(x)}(x) \to f_*(x) \qquad as \qquad i \to \infty. \tag{4.20}$$

For example, *under Assumption* 3.1(f)*, any sequence in $\mathbb{F}_A$ converges in the sense of Schäl.* This is a special case of a result in [55], reproduced in [16, Proposition D.7]. This result trivially holds, and in fact it takes a *stronger* form, if the state space X is a *countable set* (with the discrete topology) and $A(x)$ is compact for each $x \in X$. Indeed, in the latter, countable case, a standard "diagonalization" argument shows that for any sequence $\{f_n\}$ in $\mathbb{F}_A$ there exists $f_* \in \mathbb{F}_A$ and a subsequence $\{n_i\} \subset \{n\}$ *independent of* $x \in X$ such that $f_{n_i}(x) \to f_*(x)$ for all $x \in X$.

COROLLARY 4.6. *Suppose that the hypotheses of Theorem* 4.2 *are satisfied. For each $n = 1, 2, \ldots$, let $f_n \in \mathbb{F}_A$ be as in* (3.10)*. Then*

(a) $\{f_n\}$ *converges in the sense of Schäl to some* $f_* \in \mathbb{F}_A$, *and*
(b) f_* *is a minimax strategy for the infinite-horizon* α*-DC problem.*

Proof. As mentioned in the previous paragraph, part (a) comes from [55]. To prove (b), first observe that (4.14), (4.6), and (3.1) yield

$$|V_{n,\alpha}(x,\pi,\gamma)| \le w(x)\bar{c}\sum_{t=0}^{n-1}(\alpha\beta)^t \le w(x)\bar{c}/(1-\tau), \qquad \text{with } \tau := \alpha\beta,$$

for all $n \ge 1$, π, γ, and x. Therefore the sequence $\{v_{n,\alpha}\}$ of optimal n-stage costs is *bounded in the* w*-norm*, i.e.,

$$\|v_{n,\alpha}\|_w \le \bar{c}/(1-\tau) \qquad \forall n. \tag{4.21}$$

Moreover, from (3.10),

$$v_{n,\alpha}(x) \ge c(x, f_n(x), b) + \alpha \int_X v_{n-1,\alpha}(y)Q(dy|x, f_n(x), b) \tag{4.22}$$

for all b in $B(x, f_n(x))$. Now choose an arbitrary $x \in X$, and let $n_i \equiv n_i(x)$ be as in (4.20), so that $(x, f_{n_i}(x)) \to (x, f_*(x))$. Next, choose an arbitrary b_* in $B(x, f_*(x))$. By Assumption 3.1(g), there exists b_{n_i} in $B(x, f_{n_i}(x))$ such that $b_{n_i} \to b_*$, and so $(x, f_{n_i}(x), b_{n_i}) \to (x, f_*(x), b_*)$. Finally, in (4.22) replace n, $f_n(x)$, and b with n_i, $f_{n_i}(x)$, and b_{n_i}, respectively, and let $i \to \infty$. Then, by (4.7) and the extension of Fatou's lemma in [17, Lemma 8.3.7, p. 48] (which is applicable in the present case because of (4.21)), we obtain

$$v^*(x) \ge c(x, f_*(x), b_*) + \alpha \int_X v^*(y)Q(dy|x, f_*(x), b_*).$$

Actually, as $b_* \in B(x, f_*(x))$ was arbitrary,

$$v^*(x) \ge \sup_{b \in B(x, f_*(x))} \left[c(x, f_*(x), b) + \alpha \int_X v^*(y)Q(dy|x, f_*(x), b) \right]. \tag{4.23}$$

Therefore, as x was also arbitrary, (4.23) holds for all $x \in X$. On the other hand, by the definition (3.8) of T_α, together with (4.4), the right-hand side of (4.23) is minorized by $T_\alpha v^* = v^*$, and so the equality (4.1) follows. Hence, by Theorem 4.2, f_* is a minimax strategy. □

5. Infinite-horizon AC. For the AC criterion in (2.7), Küenle [32, 33] and Kurano [35] give the following analogue of Theorem 4.1, in which $T \equiv T_1$ is the DP operator in (3.7), (3.8) when $\alpha = 1$. (In fact, Küenle assumes that the function h^* in Theorem 5.1 is *bounded*, but, as in [18], it is easy to verify that the boundedness of h^* can be replaced with (5.2). On the other hand, Kurano assumes that $c(x,a,b)$ is *nonnegative* and that $B(x,a) \equiv B$ does not depend on (x,a). Combining their techniques, we get Theorem 5.1.)

THEOREM 5.1 (see [32, 33, 34, 35]). *Suppose that there exists a constant* ρ^*, *a measurable function* h^* *on* X, *and a stationary strategy* $f^* \in \mathbb{F}_A$ *such that, for all* $x \in X$,

$$\rho^* + h^*(x) = \sup_{b \in B(x, f^*(x))} \left[c(x, f^*(x), b) + \int_X h^*(y)Q(dy|x, f^*(x), b) \right] \tag{5.1}$$

and

$$\lim_{n\to\infty}\frac{1}{n}\mathbb{E}_x^{\pi,\gamma}[h^*(x_n)]=0 \qquad \forall \pi\in\Pi,\ \gamma\in\Gamma. \tag{5.2}$$

Then

$$\rho^*=\sup_{\gamma\in\Gamma}J(x,f^*,\gamma) \qquad \forall x\in X. \tag{5.3}$$

If in addition

$$\rho^*+h^*(x)=Th^*(x) \qquad \forall x\in X, \tag{5.4}$$

then ρ^* is the optimal AC, and f^* is a minimax strategy for the AC criterion, that is, (5.3) *holds and also*

$$\rho^*=\inf_{\pi\in\Pi}\sup_{\gamma\in\Gamma}J(x,\pi,\gamma) \qquad \forall x\in X. \tag{5.5}$$

We will next give conditions for the existence of a "triplet" (ρ^*,h^*,f^*) that satisfies (5.1), (5.2), and (5.4). We shall use the notation $\mu(v):=\int v\,d\mu$ whenever the integral is well defined.

First of all, throughout this section we suppose that Assumption 3.1 holds except for part (e). (Actually, (e) can be obtained from the following assumption taking $\beta:=\theta+\nu(w)/\nu(X)$ because $\delta(x,a,b)\le 1/\nu(X)$. Assumption 5.1(c) also implies (4.6) for suitable values of α; see Remark 8.3.5(a) in [17], for instance.) In addition we suppose the following.

Assumption 5.1. There exists a measure ν on X, with $\nu_*:=\nu(X)>0$, a nonnegative u.s.c. function δ on $\mathbb{K}$, and a constant $0<\theta<1$ such that

(a) $Q(D\mid x,a,b)\ge\delta(x,a,b)\,\nu(D)$ for all $(x,a,b)\in\mathbb{K}$ and $D\in\mathfrak{B}(X)$;
(b) $\nu(w):=\int_X w(x)\nu(dx)<\infty$;
(c) $\int_X w(y)Q(dy\mid x,a,b)\le\theta w(x)+\delta(x,a,b)\,\nu(w)$ for all $(x,a,b)\in\mathbb{K}$; and
(d) $\int_X\delta(x,f(x),g(x,f(x)))\nu(dx)>0$ for each $f\in\mathbb{F}_A$ and $g\in\mathbb{F}_B$.

See Remark 5.3 for comments on Assumption 5.1 and other similar hypotheses used in the literature on stochastic games and Markov control problems under the AC criterion.

THEOREM 5.2. *If Assumptions* 3.1 *and* 5.1 *hold, then there is a constant ρ^*, a function h^* in $\mathbb{B}_{lsc}$, and a stationary policy $f^*\in\mathbb{F}_A$ that satisfy* (5.1), (5.2), *and* (5.4). *Hence the conclusions of Theorem* 5.1 *hold.*

To prove Theorem 5.2 we shall first introduce some notation and a preliminary result.

For any real-valued measurable function u on $\mathbb{K}$ and any two stationary policies $f\in\mathbb{F}_A$ and $g\in\mathbb{F}_B$ we write

$$u(x,f,g):=u(x,f(x),g(x,f(x))) \qquad \forall x\in X.$$

In particular, $c(x,f,g):=c(x,f(x),g(x,f(x)))$ and similarly for $Q(\cdot\mid x,f,g)$ and $\delta(x,f,g)$. The following result can be proved as Vega-Amaya's [61] Theorem 3.3 and Lemma 4.3.

LEMMA 5.3. *Suppose that Assumption* 5.1 *holds. Then for each pair of stationary strategies $f\in\mathbb{F}_A$ and $g\in\mathbb{F}_B$ we have the following:*

(a) *$Q(\cdot\mid x,f,g)$ is positive Harris recurrent—hence, in particular, $Q(\cdot\mid x,f,g)$ admits a unique invariant probability measure (i.p.m.), say $\mu_{f,g}$. Moreover,*

(b) $\mu_{f,g}(w) := \int w\, d\mu_{f,g} < \infty$, *and*
(c) *the function* $\delta(x,a,b)$ *and the constant* $\mu_{f,g}(\delta) := \int \delta(x,f,g)\mu_{f,g}(dx)$ *satisfy*

$$\mu_{f,g}(\delta) = \lim_{n\to\infty} \frac{1}{n} E_x^{f,g}\left[\sum_{t=0}^{n-1} \delta(x_t,f,g)\right] \qquad \forall x \in X.$$

(d) *If in addition* (3.1) *holds, then (by* (b)*)*

$$J(f,g) := \int_X c(x,f,g)\mu_{f,g}(dx) \qquad \textit{and} \qquad \rho^* := \inf_{f\in\mathbb{F}_A} \sup_{g\in\mathbb{F}_B} J(f,g) \tag{5.6}$$

are both finite.

Observe that the constant $J(f,g)$ in (5.6) coincides with the average cost (2.7) if $\pi = f$ and $\gamma = g$, i.e.,

$$J(x,f,g) = J(f,g) \qquad \forall x \in X.$$

On the other hand, integrating with respect to the i.p.m. $\mu_{f,g}$ both sides of the inequality in Assumption 5.1(c), we obtain that

$$\mu_{f,g}(\delta) \geq \mu_{f,g}(w)(1-\theta)/\nu(w).$$

Hence, as $w(\cdot) \geq 1$, letting $\delta_* := (1-\theta)/\nu(w)$ we get

$$\mu_{f,g}(\delta) \geq \delta_* > 0 \qquad \forall f \in \mathbb{F}_A,\ g \in \mathbb{F}_B. \tag{5.7}$$

Furthermore, if $\hat{Q}$ denotes the substochastic kernel

$$\hat{Q}(\cdot \mid x,a,b) := Q(\cdot \mid x,a,b) - \delta(x,a,b)\nu(\cdot) \tag{5.8}$$

(see Assumption 5.1(a)), then we may express the Assumption 5.1(c) as

$$\int_X w(y)\hat{Q}(dy \mid x,a,b) \leq \theta w(x) \qquad \forall (x,a,b) \in \mathbb{K}. \tag{5.9}$$

From these facts we may obtain the proof of Theorem 5.2 as follows.

Proof of Theorem 5.2. Let ρ^* and $\hat{Q}$ be as in (5.6) and (5.8), respectively, and for each function u in $\mathbb{B}_{lsc}$ and $x \in X$ let

$$Mu(x) := \inf_{a\in A(x)} \sup_{b\in B(x,a)} \left[c(x,a,b) + \int_X u(y)\hat{Q}(dy \mid x,a,b) - \rho^*\right] \tag{5.10}$$

or, more explicitly,

$$Mu(x) = \inf_{a\in A(x)} \sup_{b\in B(x,a)} \left[c(x,a,b) + \int_X u(y)Q(dy \mid x,a,b) - \delta(x,a,b)\nu(u) - \rho^*\right].$$

It is easily seen that M is a *contraction mapping* from $\mathbb{B}_{lsc}$ into itself. Indeed, by Assumption 5.1, the function δ is u.s.c., and so $-\delta$ is l.s.c. Thus, as in Lemma 3.3, it follows that if u is in $\mathbb{B}_{lsc}$, then so is Mu. Moreover, it is obvious that M is monotone, and, on the other hand, by (5.9),

$$M(u + rw) \leq Mu + \theta rw \qquad \forall u \in \mathbb{B}_{lsc}, r \geq 0.$$

Thus Lemma 4.4 yields that M is a contraction mapping on $\mathbb{B}_{lsc}$ with modulus θ, and it follows that M has a unique fixed point h^* in $\mathbb{B}_{lsc}$, i.e., $h^* = Mh^*$, or, equivalently,

(5.11)
$$\rho^* + h^*(x) = \inf_{a\in A(x)} \sup_{b\in B(x,a)} \left[c(x,a,b) + \int_X h^*(y) Q(dy \mid x,a,b) - \delta(x,a,b)\nu(h^*) \right].$$

We will next show that

$$\nu(h^*) = 0 \tag{5.12}$$

so that (5.11) reduces to the AC "optimality equation" (5.4).

To this end, first note that as in Lemma 3.3(c) (or Lemma 3.2(b)) there exists a stationary strategy f^* in $\mathbb{F}_A$ such that

(5.13)
$$\rho^* + h^*(x) = \sup_{b\in B(x,f^*(x))} \left[c(x,f^*(x),b) + \int_X h^*(y) Q(dy \mid x, f^*(x), b) - \delta(x,f^*(x),b)\nu(h^*) \right]$$

for all $x \in X$. Therefore, for any strategy $g \in \mathbb{F}_B$ we have

$$\rho^* + h^*(x) \geq c(x,f^*,g) + \int_X h^*(y) Q(dy \mid x, f^*, g) - \delta(x,f^*,g)\nu(h^*),$$

and integrating both sides of this inequality with respect to $\mu_{f^*,g}$ we get

$$\rho^* \geq J(f^*,g) - \mu_{f^*,g}(\delta)\nu(h^*) \qquad \forall g \in \mathbb{F}_B.$$

Now suppose that $\nu(h^*) < 0$. Hence, by (5.7),

$$-\mu_{f^*,g}(\delta)\nu(h^*) > -\delta_*\nu(h^*) > 0,$$

and so

$$\rho^* \geq J(f^*,g) - \delta_*\nu(h^*) > J(f^*,g) \qquad \forall g \in \mathbb{F}_B.$$

This inequality contradicts the definition of ρ^* in (5.6), according to which

$$\rho^* \leq \sup_{g\in\mathbb{F}_B} J(f,g) \qquad \forall f \in \mathbb{F}_A.$$

As a consequence, $\nu(h^*)$ cannot be negative; that is, necessarily $\nu(h^*) \geq 0$. Hence suppose that $\nu(h^*) > 0$, and let us go back to (5.11). By well-known measurable selection theorems—see, e.g., Corollary 4.3 in [49]—for each $\varepsilon > 0$ there exists an ε-maximizer $g_\varepsilon \in \mathbb{F}_B$ of the right-hand side of (5.11), i.e.,

$$\sup_{b\in B(x,a)} \left[c(x,a,b) + \int_X h^*(y) Q(dy \mid x,a,b) - \delta(x,a,b)\nu(h^*) \right]$$
$$\leq c(x,a,g_\varepsilon(x,a)) + \int_X h^*(y) Q(dy \mid x,a,g_\varepsilon(x,a)) - \delta(x,a,g_\varepsilon(x,a))\nu(h^*) + \varepsilon$$

for all $(x,a) \in \mathbb{K}_A$. The latter inequality and (5.11) yield that for any stationary strategy $f \in \mathbb{F}_A$ and $x \in X$,

$$\rho^* + h^*(x) \leq c(x,f,g_\varepsilon) + \int_X h^*(y) Q(dy \mid x, f, g_\varepsilon) - \delta(x,f,g_\varepsilon)\nu(h^*) + \varepsilon,$$

and so integration with respect to the i.p.m. μ_{f,g_ε} gives

$$\rho^* \le J(f, g_\varepsilon) - \mu_{f,g_\varepsilon}(\delta)\nu(h^*) + \varepsilon \qquad \forall f \in \mathbb{F}_A.$$

Hence, as we have assumed that $\nu(h^*) > 0$, from (5.7) we get

$$\rho^* \le \inf_{f\in\mathbb{F}_A} J(f, g_\varepsilon) - \delta_*\nu(h^*) + \varepsilon.$$

In particular, if we take $0 < \varepsilon < \delta_*\nu(h^*)$, there exists $g_\varepsilon \in \mathbb{F}_B$ such that

$$\rho^* < \inf_{f\in\mathbb{F}_A} J(f, g_\varepsilon),$$

which again contradicts the definition of ρ^* in (5.6), because the latter gives us

$$\rho^* \ge \inf_{f\in\mathbb{F}_A} J(f, g) \qquad \forall g \in \mathbb{F}_B.$$

Thus, necessarily (5.12) holds, and so (5.11) and (5.13) yield (5.4) and (5.1), respectively. Finally, using (4.15) and Assumption 5.1(c) a straightforward calculation shows that (5.2) *holds for any function in* $\mathbb{B}_w(X)$. This completes the proof of Theorem 5.2. □

A special case. The contraction mapping in (5.10) gives the desired theoretical results in Theorem 5.2, but for practical purposes it is not very useful because it is defined in terms of ρ^*, which is precisely what we would like to compute! Hence, to show a "computable" case, *throughout the rest of this section we consider the special case in which Assumption* 5.1 *holds with*

$$\delta(x, a, b) \equiv 1 \qquad \forall (x, a, b) \in \mathbb{K}. \tag{5.14}$$

Now let $v_n \equiv v_{n,1}$ be the optimal n-stage cost when $\alpha = 1$ (see Theorem 3.1(c)), with $v_0 \equiv 0$. We denote by $\{u_n\}$ the sequence in $\mathbb{B}_{lsc}$ defined as $u_0 := v_0$, $u_1 := v_1$, and for $n \ge 2$

$$u_n := v_n - m_n, \qquad \text{with} \qquad m_n := \sum_{j=1}^{n-1} (1 - \nu_*)^{n-1-j}\nu(v_j), \tag{5.15}$$

where $\nu_* := \nu(X)$. Furthermore, let

$$\rho_n := \nu(u_n) = \nu(v_n) - \nu_* m_n \qquad \forall n \ge 0. \tag{5.16}$$

Observe that all of these quantities are indeed "computable." Moreover, they yield a solution h^*, ρ^* of (5.4) as follows.

COROLLARY 5.4. *Suppose that the hypotheses of Theorem* 5.2 *are satisfied, with* δ *as in* (5.14). *Then there exist* $\rho^* \in \mathbb{R}$, $h^* \in \mathbb{B}_{lsc}$, *and* $f^* \in \mathbb{F}_A$ *as in Theorem* 5.2 *and, in addition, for all* $n = 0, 1, \ldots$, *we have*

$$\|u_n - h^*\|_w \le \theta^n \bar{c}/(1 - \theta), \tag{5.17}$$

$$|\rho_n - \rho^*| \le \theta^n \nu(w)\bar{c}/(1 - \theta). \tag{5.18}$$

Proof. Instead of the mapping Mu in (5.10), consider

$$Nu(x) := \inf_{a\in A(x)} \sup_{b\in B(x,a)} \left[c(x,a,b) + \int_X u(y)\hat{Q}(dy \mid x,a,b) \right] \tag{5.19}$$

for each $u \in \mathbb{B}_{lsc}$, with $\hat{Q}(\cdot \mid x,a,b) := Q(\cdot \mid x,a,b) - \nu(\cdot)$; see (5.8) and (5.14). Then, arguing exactly as in the proof of Theorem 5.2, it follows that N is a contraction mapping from $\mathbb{B}_{lsc}$ into itself, with modulus θ. Therefore, there is a unique function h^* in $\mathbb{B}_{lsc}$ such that $h^* = Nh^*$; equivalently, letting $\rho^* := \nu(h^*)$, the pair (h^*, ρ^*) is the unique solution of (5.4). Hence, as the existence of $f^* \in \mathbb{F}_A$ satisfying (5.1) follows from Theorem 5.2, to complete the proof it remains only to consider (5.17) and (5.18).

To this end, let $T := T_1$ be the DP operator in (3.8) when $\alpha = 1$, and note that (5.19) can be written as

$$Nu = Tu - \nu(u). \tag{5.20}$$

Now let $u_0 := 0$, and for $n \geq 1$ define u_n in $\mathbb{B}_{lsc}$ as

$$u_n := Nu_{n-1} = N^n u_0, \tag{5.21}$$

which, by (5.20), we can also write as $u_n = Tu_{n-1} - \nu(u_{n-1})$. Then, as (3.9) gives (with $\alpha = 1$) $v_n = Tv_{n-1} = T^n v_0$, a direct induction argument shows that the sequence in (5.21) is precisely the sequence $\{u_n\}$ in (5.15). Consequently, the contraction property of N and (4.10)(b) with $u = u_0 = 0$ give

$$\|u_n - h^*\|_w \leq \theta^n \|h^*\|_w.$$

Thus to get (5.17) it suffices to show that $\|h^*\|_w \leq \bar{c}/(1-\theta)$. To obtain the latter inequality note that (3.1), (3.4), and (5.9) yield

$$\|Nu\|_w \leq \bar{c} + \theta\|u\|_w \qquad \forall u \in \mathbb{B}_{lsc}. \tag{5.22}$$

This inequality and (4.11)–(4.12) give that $\|h^*\|_w \leq \bar{c}/(1-\theta)$, and so (5.17) follows.

Finally, since $\rho^* := \nu(h^*)$, from (5.16) we obtain

$$|\rho_n - \rho^*| \leq \int |u_n - h^*| d\nu \leq \|u_n - h^*\|_w \int w d\nu.$$

This inequality and (5.17) give (5.18) because $\nu(w) := \int w d\nu$. □

REMARK 5.1. *We should mention that although the "contraction approach" is quite standard, the convergence estimates in* (5.17) *and* (5.18) *are new, even for the standard (or one-player) AC control problem. As in the α-DC case, obtaining h^* and ρ^* via u_n and ρ_n is also called the* value iteration *(or* successive approximations*) procedure for the AC problem.*

Of course, (5.17) and (5.18) yield for the AC problem an analogue of Corollary 4.6, as follows.

COROLLARY 5.5. *Suppose that the hypotheses of Corollary* 5.4 *hold, and let $v_n \equiv v_{n,1}$ and f_n be as in* (3.10) *with $\alpha = 1$, for each $n = 1, 2, \ldots$, with $v_0 \equiv 0$. Then*

(a) *$\{f_n\}$ converges in the sense of Schäl to some stationary strategy $f_* \in \mathbb{F}_A$, and*

(b) *f_* is a minimax strategy for the AC problem.*

Proof. Part (a) is obtained as in the paragraph after Definition 4.5. To prove (b) note that (5.15) and (5.16) give

$$m_{n+1} - m_n = \rho_n \qquad \forall n \geq 2.$$

Using this equality and (5.15) we may express (3.10), when $\alpha = 1$, as

$$\rho_{n-1} + u_n(x) = \sup_{b \in B(x, f_n(x))} \left[c(x, f_n(x), b) + \int_X u_{n-1}(y) Q(dy|x, f_n(x), b) \right] \tag{5.23}$$

for all $x \in X$ and $n \geq 3$. Finally, (5.22) and the first equality in (5.21) yield

$$\|u_n\|_w \leq \bar{c} + \theta \|u_{n-1}\|_w \qquad \forall n \geq 1 \tag{5.24}$$

and, therefore,

$$\|u_n\|_w \leq \bar{c}/(1-\theta) \qquad \forall n \geq 0;$$

that is, the sequence $\{u_n\}$ *is bounded in the w-norm* (compare (5.24) with (4.21)). Hence, as (5.23) yields

$$\rho_{n-1} + u_n(x) \geq c(x, f_n(x), b) + \int_X u_{n-1}(y) Q(dy \mid x, f_n(x), b) \qquad \forall b \in B(x, f_n(x)),$$

using (5.17) and (5.18) the proof of (b) can now be completed "exactly" as the proof of Corollary 4.6(b) (after (4.22)), with the obvious changes. □

REMARK 5.2. *Concerning Assumption* 5.1(a)*, with* $\delta \equiv 1$ *as in* (5.14)*, Kurano* [35] *introduced an interesting approach that might be particularly useful if one does not know in advance that the said assumption holds. The idea is to introduce, for each* $0 < \varepsilon < 1$*, an auxiliary game model in which the transition law* Q *is replaced with*

$$Q_\varepsilon(\cdot \mid x, a, b) := (1-\varepsilon) Q(\cdot \mid x, a, b) + \varepsilon \psi(\cdot) \qquad [\geq \varepsilon \psi(\cdot)]$$

for some suitable probability measure ψ *on* X*. As* Q_ε *satisfies the Assumption* 5.1(a) *and* (5.14)*, one can solve (as in Corollary* 5.4*, say) the corresponding minimax control problem, and, finally, one should verify that the original problem is indeed solved in the limit as* $\varepsilon \to 0$*.*

The following proposition gives a couple of restrictive but easy-to-verify conditions that guarantee Assumption 5.1(a) with $\delta \equiv 1$.

PROPOSITION 5.6 (cf. Theorem 3.2 in [19]). *Suppose that* $Q(\cdot \mid x, a, b)$ *has a density* $q(x, a, b, \cdot)$ *with respect to a* σ*-finite measure* m *on* X*, that is,*

$$Q(D \mid x, a, b) = \int_D q(x, a, b, y) m(dy) \qquad \forall (x, a, b) \in \mathbb{K}, D \in \mathfrak{B}(X). \tag{5.25}$$

Then the following two conditions satisfy that (a) $\Rightarrow$ (b) $\Rightarrow$ *Assumption* 5.2(a) *with* $\delta \equiv 1$*.*

(a) *There exists a number* $\varepsilon > 0$ *and a set* $D_0 \in \mathfrak{B}(X)$ *with* $m(D_0) > 0$ *such that*

$$q(x, a, b, y) \geq \varepsilon \qquad \forall (x, a, b) \in \mathbb{K},\ y \in D_0. \tag{5.26}$$

(b) *There exists a nonnegative measurable function* q_0 *on* X *such that*

$$q(x, a, b, y) \geq q_0(y) \qquad \forall (x, a, b) \in \mathbb{K},\ y \in X \tag{5.27}$$

and

$$\int_X q_0(y)m(dy) > 0.$$

Proof. (a) *implies* (b): If (a) holds, then $q_0(y) := \varepsilon I_{D_0}(y)$ satisfies (b), where I_D stands for the indicator function of a set D.

(b) *implies Assumption* 5.1(a) *with* $\delta \equiv 1$: By (5.25) and (5.27), the measure $\nu(D) := \int_D q_0(y)m(dy)$ satisfies the desired conclusion. □

For examples satisfying (5.25)–(5.27), see [11, 17, 23, 35], for instance, or Examples 7.1 and 7.2 below.

REMARK 5.3. *Hypotheses similar to Assumption* 5.1 *have been used to study several classes of AC Markov games and control processes—see, e.g.,* [1, 2, 14, 17, 18, 22, 23, 25, 31, 32, 33, 34, 35, 42, 46, 48, 56, 61]*—but with some important differences. For instance, in* [18] *the inequality* $\int \delta(x,f,g)\nu(dx) > 0$ *in Assumption* 5.1(d) *is supposed to hold* uniformly *in* $f \in \mathbb{F}_A$ *and* $g \in \mathbb{F}_B$*, but Vega-Amaya* [61] *noted that it suffices to take the inequality in our present form, that is, for* each *pair of strategies* $f \in \mathbb{F}_A$*,* $g \in \mathbb{F}_B$*. On the other hand, instead of our Assumptions* 5.1(c) *and* (a)*, Jaśkiewicz and Nowak* [25] *use conditions of the form*

$$\int_X w(y)Q(dy \mid x,a,b) \le \theta w(x) + \eta I_C(x) \tag{5.28}$$

for some fixed Borel subset C *of* X*, and*

$$Q(D \mid x,f,g) \ge \delta\nu_{f,g}(D) \qquad \forall D \subset C, D \in \mathfrak{B}(X), \tag{5.29}$$

respectively, where $\nu_{f,g}$ *is a probability measure on* X *concentrated on* C*, for each* $f \in \mathbb{F}_A$ *and* $g \in \mathbb{F}_B$*. In the latter case, when* $\nu_{f,g}$ *depends on the strategies* f *and* g*, our fixed-point approach to prove Theorem* 5.2 *is* not *applicable. Another different form of* (5.28), (5.29) *has been introduced in* [31]*, allowing multichain Markov games. This basically means that there is a measurable partition* $D_0, \dots, D_m$ *of* X *such that for every pair of stationary strategies* $f \in \mathbb{F}_A$ *and* $g \in \mathbb{F}_B$ *the corresponding Markov chain has the same set* D_0 *of transient states and the same recurrent sets which consist of several cyclic sets* $D_1, \dots, D_m$*. A little more precisely, there is an integer* $k \ge 1$ *and measures* $\nu_1, \dots, \nu_m$ *on* X*, with* ν_i *concentrated on* D_i $(i = 1, \dots, m)$ *such that*

$$Q^k(\cdot \mid x,f,g) \ge \sum_{i=1}^m I_{D_i}(x)\nu_i(\cdot) \tag{5.30}$$

for each $x \in X$*,* $f \in \mathbb{F}_A$*, and* $g \in \mathbb{F}_B$*, where* I_D *denotes the indicator function of the set* D*. When* $k = m = 1$ *and* $D_1 = X$*,* (5.30) *reduces to Assumption* 5.1(a) *with* $\delta \equiv 1$ *as in* (5.14)*; in other words,* (5.30) *is a generalization of our special case* (5.14)*. In general, however, since the right-hand side of* (5.30) *does not depend on the players' actions/strategies, the relation (if any) between* (5.30) *and either* (5.29) *or our Assumption* 5.1(a) *is unclear. For other, detailed comments on Assumption* 5.1*, see* [25] *and* [61].

Finally, we mention also that the special case (5.14) has been used by many authors; see, e.g., [11, 15, 19, 32, 50, 52] and the references therein.

6. Systems with unknown disturbance distribution. In this section we consider the control system

$$x_{t+1} = F(x_t, a_t, \xi_t), \qquad t = 0, 1, \dots, \tag{6.1}$$

in which the ξ_t are independent random variables with values in a Borel space S and with *unknown* probability distributions. More precisely, the distributions may change from period to period, but they are restricted to lie in a certain class B specified below. The initial state x_0 is supposed to be independent of the sequence $\{\xi_t\}$, and F is a measurable function from $X \times A \times S$ to X.

REMARK 6.1. *Control problems in which the random variables are i.i.d. with a common but* unknown *distribution μ are usually referred to as (nonparametric)* adaptive control problems. *They are so named because at each decision time $t = 0, 1, \ldots$, the controller determines an estimate of the distribution of ξ_t, say $\hat{\xi}_t$, and then he/she "adapts" his/her actions to the given estimate. But, of course, this approach requires the disturbances to be "observable" because the estimate $\hat{\xi}_t$ is typically computed using realizations $\xi_0, \ldots, \xi_t$* [15, 21, 42]. *This observability condition occurs in many situations. For example, in control of inventories, queues, and water reservoirs, the disturbances are, respectively, the product's demand, the arrival process, and the water inflow, which can be safely assumed to be observable. However, there are other cases in which the disturbances are "really" random noises and it is impossible to "observe" them. For example, in economics, finance, and control of populations (fisheries, epidemics, and so on) there are so many external factors influencing the system's dynamics that it is practically necessary to model the disturbances as a "real" (unobservable) random noise. This is precisely the kind of situation that we have in mind in this section: the controller's opponent is the "nature" that at each period t picks (from a given set) a distribution for ξ_t.*

We shall denote by $\mathbb{P}(S)$ the family of probability measures on S, endowed with the topology of weak convergence. Thus, a sequence $\{\mu_n\}$ in $\mathbb{P}(S)$ *converges weakly* to μ if

$$\int_S h\, d\mu_n \to \int_S h\, d\mu \qquad \forall h \in C_b(S), \tag{6.2}$$

where $C_b(S)$ is the space of real-valued continuous bounded functions on S. As S is a Borel space, so is $\mathbb{P}(S)$ (see [4]).

We shall consider the game model in (2.1) except that now we suppose the following:

(a) B is a Borel subset of $\mathbb{P}(S)$ such that $B(x, a) = B$ for all $(x, a) \in \mathbb{K}_A$.

(b) The transition law $Q(\cdot \mid x, a, b)$ is the conditional distribution of x_{t+1} given that $(x_t, a_t) = (x, a) \in \mathbb{K}_A$ and that ξ_t has distribution $b \in \mathbb{P}(S)$; hence, by (6.1),

$$Q(D|x, a, b) = \int_S I_D[F(x, a, s)]b(ds) \qquad \forall D \in \mathfrak{B}(X). \tag{6.3}$$

(c) Similarly, for some given measurable function $\hat{c}$ on $\mathbb{K}_A \times S$,

$$c(x, a, b) = \int_S \hat{c}(x, a, s)b(ds). \tag{6.4}$$

We shall refer to this model as a *game against nature*. In the remainder of this section we give conditions under which the latter game satisfies some parts of Assumptions 3.1 and 5.1. The simplest—actually, trivial—is the following.

PROPOSITION 6.1. *Assumption* 3.1(b) *holds for the game against nature if there exists a constant $\bar{c} \geq 0$ and a measurable function $w(\cdot) \geq 1$ on X such that*

$$|\hat{c}(x, a, s)| \leq \bar{c}w(x) \qquad \forall (x, a, s) \in \mathbb{K}_A \times S. \tag{6.5}$$

Proof. The proof is trivial. □

To obtain other parts of Assumption 3.1 the game against nature requires more structure, as in the following proposition.

PROPOSITION 6.2. *Assumption* 3.1(c) *holds for the transition law in* (6.3) *if*

(a) *X, A, and S are locally compact separable metric spaces, and $\mathbb{K}_A$ is a closed subset of $X \times A$; and*

(b) *$F : \mathbb{K}_A \times S \to X$ is continuous.*

Proof. First note that the integrand in (6.3) can be written as a Dirac measure concentrated at $F(x, a, s)$, i.e.,

$$I_D[F(x, a, s)] = \delta_{F(x,a,s)}(D).$$

Suppose now that $(x^n, a^n, b^n) \to (x, a, b)$ in $\mathbb{K}_A \times B$, and define the stochastic kernels

$$\varphi_n(D \mid s) := \delta_{F(x^n,a^n,s)}(D) \qquad \text{and} \qquad \varphi(D \mid s) := \delta_{F(x,a,s)}(D). \tag{6.6}$$

Then, in particular, we can write (6.3) as a so-called *b-mixture of φ*, that is,

$$Q(D \mid x, a, b) = \int_S \varphi(D \mid s) b(ds).$$

Therefore, as the hypothesis (a) ensures that $\mathbb{K}_A \times S \to X$ is a locally compact and separable metric space and, on the other hand, we are assuming that $b^n \to b$ weakly, the proposition will follow from Serfozo's [57] Theorem 4.1 provided that

$$\varphi_n(\cdot \mid s^n) \to \varphi(\cdot \mid s) \qquad \text{weakly if} \qquad s^n \to s.$$

The latter condition, however, is obvious because by (6.6) and the hypothesis (b) we get

$$\int_X u(y)\varphi_n(dy|s^n) = u[F(x^n, a^n, s^n)] \to u[F(x, a, s)] = \int_X u(y)\varphi(dy|s)$$

for any continuous bounded function u on X. □

To get conditions under which the cost function in (6.4) satisfies Assumption 3.1(a) we shall assume that $\hat{c}$ is "separable" in (x, a) and s, that is,

$$\hat{c}(x, a, s) = c_1(x, a) + c_2(s) \qquad \forall (x, a, s) \in \mathbb{K}_A \times S. \tag{6.7}$$

In this case, (6.4) becomes

$$c(x, a, b) = c_1(x, a) + \int_S c_2(s) b(ds). \tag{6.8}$$

It is well known that if $b^n \to b$ weakly and $h : S \to \mathbb{R}$ is l.s.c. and bounded below, then

$$\liminf_{n\to\infty} \int_S h db^n \geq \int_S h db. \tag{6.9}$$

This fact yields the following.

PROPOSITION 6.3. *Assumption* 3.1(a) *holds for the function c in* (6.8) *if c_1 is l.s.c. on $\mathbb{K}_A$ and c_2 is l.s.c. and bounded below on S.*

Proposition 6.6, below, does not require c_2 to be bounded below. To state this, let us first recall the following concept.

DEFINITION 6.4. *Let $\{b^n\}$ be a sequence in $\mathbb{P}(S)$. A function u on S is said to be $\{b^n\}$-*uniformly integrable *if*

$$\lim_{\eta\to\infty} \sup_n \int_{N(u,\eta)} |u(s)| b^n(ds) = 0, \tag{6.10}$$

where, for $\eta > 0$,

$$N(u,\eta) := \{s \in S \mid |u(s)| \geq \eta\}.$$

For example, (6.10) holds if

$$\sup_n \int_S |u(s)|^{1+\varepsilon} b^n(ds) < \infty \tag{6.11}$$

for some $\varepsilon > 0$ because in this case

$$\int_{N(u,\eta)} |u(s)| b^n(ds) \leq (1/\eta^{\varepsilon}) \int_S |u(s)|^{1+\varepsilon} b^n(ds).$$

At any rate, the fact we are interested in is the following, whose proof can be found in [57], for instance.

LEMMA 6.5. *Suppose that $u : S \to \mathbb{R}$ is continuous and that it satisfies* (6.10). *If in addition $b^n \to b$ weakly, then $\int u\, db^n \to \int u\, db$.*

Replacing the function u in Lemma 6.5 with c_2 we obtain the following.

PROPOSITION 6.6. *Assumption* 3.1(a) *holds for the function c in* (6.8) *if c_1 is l.s.c. on $\mathbb{K}_A$ and, furthermore,*

(a) *c_2 is continuous on S;*

(b) *if $(x^n, a^n) \to (x, a)$ and b^n is in B for all n, then c_2 is $\{b^n\}$-uniformly integrable.*

EXAMPLE 6.1. *Suppose that $S = \mathbb{R}$ and let $c_2(s) := s^2$. From* (6.9) *we obtain that if $b^n \to b$ weakly, then*

$$\liminf_{n\to\infty} \int s^2 b^n(ds) \geq \int s^2 b(ds). \tag{6.12}$$

On the other hand, if, for instance,

$$\sup_n \int |s|^3 b^n(ds) < \infty, \tag{6.13}$$

and $b^n \to b$ weakly, then instead of (6.12) *we obtain the stronger result:*

$$\lim_{n\to\infty} \int s^2 b^n(ds) = \int s^2 b(ds). \tag{6.14}$$

This result comes from Lemma 6.5 *and* (6.11) *with $\varepsilon = 1$. In other words,* (6.14) *states that the mapping $b \mapsto \int s^2 db$ is continuous, whereas* (6.12) *gives that it is l.s.c.*

Finally, note that for "separable" functions as in (6.8) we have

$$\inf_{a\in A(x)} \sup_{b\in B} c(x,a,b) = \inf_{a\in A(x)} c_1(x,a) + \sup_{b\in B} \int c_2(s) b(s). \tag{6.15}$$

7. Examples. We will now present two examples on the game against nature introduced in section 6. In both examples, the state space X and the control set A are Borel subsets of $\mathbb{R}$. Moreover, to fix ideas we will choose a particular set B of admissible noise distributions, but the reader should note that such a set B can be replaced with *any family* of compactly supported distributions with zero mean, finite second moment, and strictly positive densities (for instance, uniform distributions). Further, the choice of zero-mean distributions is simply to facilitate calculations.

Throughout this section, the disturbance set S in (6.3), (6.4) is the compact interval $S := [-\hat{s}, \hat{s}]$ for some $\hat{s} > 0$, and the set $B \subset \mathbb{P}(S)$ of admissible disturbance distributions is the family $N(0, \sigma^2; \underline{\sigma}, \bar{\sigma})$ of truncated Gaussian distributions on S, with zero mean and standard deviation σ in $[\underline{\sigma}, \bar{\sigma}]$, where $0 < \underline{\sigma} \leq \bar{\sigma} < \infty$. In other words, a probability measure, denoted by b_σ, in B is of the form $b_\sigma(ds) = g_\sigma(s)ds$, where g_σ is the density function given by

$$g_\sigma(s) := k(\sigma)^{-1} e^{-s^2/2\sigma^2} I_S(s) \qquad \forall s \in \mathbb{R}, \tag{7.1}$$

with

$$k(\sigma) := \int_{-\hat{s}}^{\hat{s}} e^{-s^2/2\sigma^2} ds. \tag{7.2}$$

We shall denote by $\xi(\sigma)$ a generic random variable with distribution b_σ in B. It is evident that the second moments

$$E[\xi(\sigma)^2] = \int_{-\hat{s}}^{\hat{s}} s^2 g_\sigma(s) ds \qquad \text{for} \qquad \sigma \in [\underline{\sigma}, \bar{\sigma}] \tag{7.3}$$

are continuous in σ and uniformly bounded above. On the other hand, as $S = [-\hat{s}, \hat{s}]$ is a compact set, so is $\mathbb{P}(S)$ and also $B \subset \mathbb{P}(S)$. Therefore, there exists $\xi_*^2 > 0$ such that

$$\max_{\underline{\sigma} \leq \sigma \leq \bar{\sigma}} E[\xi(\sigma)^2] = \xi_*^2. \tag{7.4}$$

EXAMPLE 7.1. *This example was motivated by the mold level control problem briefly described in section* 1. *The underlying idea is that the state variable x stands for the height of a certain object that we wish to keep as close as possible to a nominal height x_*. Thus we consider the discrete-time model*

$$x_{t+1} = x_t + a_t + \xi_t \qquad for \qquad t = 0, 1, \dots, \tag{7.5}$$

with independent disturbances $\xi_t \equiv \xi(\sigma)$ as in the previous paragraph, and state space $X := [x_* - \underline{x}, x_* + \bar{x}]$, where $\underline{x}$ and $\bar{x}$ are given positive numbers. For each state x, the control set is

$$A(x) := [x_* - x, 0] \quad if \quad x \geq x_*, \quad and \quad A(x) := [0, x_* - x] \quad if \quad x \leq x_*. \tag{7.6}$$

The rationale is that if $x \geq x_*$, then we choose a control action $a \in A(x)$ to decrease x down to a point $x + a$ in $[x_*, x]$, whereas if $x \leq x_*$, then $a \in A(x)$ increases x up to a point $x + a$ in $[x, x_*]$. Moreover, to ensure that x_t is indeed in X for all $t \geq 0$ whenever x_0 is in X, we will assume that the disturbance set $S = [-\hat{s}, \hat{s}]$ is such that

$$0 < \hat{s} \leq \min\{\underline{x}, \bar{x}\}. \tag{7.7}$$

On the other hand, observe that as the disturbances $\xi(\sigma)$ have zero mean, (7.5) *gives that $E(x_{t+1}) = x + a$ if $x_t = x$ and $a_t = a$. Therefore it is natural to consider the cost function*

$$\hat{c}(x, a, s) := (x + a - x_*)^2 + s^2, \tag{7.8}$$

so that, by (7.3) *and* (6.8),

$$c(x, a, b_\sigma) = (x + a - x_*)^2 + E[\xi(\sigma)^2]. \tag{7.9}$$

We next verify that the corresponding minimax control problems satisfy Assumptions 3.1 *and* 5.1.

Verification of Assumption 3.1. Assumption 3.1(a) obviously follows from Proposition 6.3. In fact, it also follows from Proposition 6.6 because $c_2(s) := s^2$ is continuous on the *compact* set S. To verify Assumption 3.1(b) note that

$$-\underline{x} \le x + a - x_* \le \bar{x} \qquad \forall x \in X,\ a \in A(x). \tag{7.10}$$

This inequality and (7.4) yield that the *nonnegative* cost c in (7.9) is *bounded above* by $\bar{c} := \max\{\underline{x}^2, \bar{x}^2\} + \xi_*^2$. This yields Assumption 3.1(b) and also (d) and (e) with $w(\cdot) \equiv 1$ and $\beta = 1$. Observe that $\beta = 1$ also satisfies (4.6). On the other hand, Proposition 6.2 yields Assumption 3.1(c), and finally, parts (f) and (g) in Assumption 3.1 follow from the definitions of X, $A(x)$ and $B(x, a) \equiv B$—see Remark 3.1(c).

Verification of Assumption 5.1. Recall that $w(\cdot) \equiv 1$. We now wish to verify Assumption 5.1 in the special case (5.14), that is, $\delta \equiv 1$, assuming that $\underline{x} = \bar{x} =: \hat{x}$ and that, moreover, in lieu of (7.7) we have

$$\hat{s} < \hat{x} < 2\hat{s}. \tag{7.11}$$

Now, using (7.5), (7.1), and (6.3), we get the transition law

$$Q(D \mid x, a, b_\sigma) = \int_S I_D(x + a + s) g_\sigma(s) ds \qquad \forall D \in \mathfrak{B}(X). \tag{7.12}$$

Hence, as $S = [-\hat{s}, \hat{s}]$, using the change of variable $y := x + a + s$ we can write Q as in (5.25), i.e.,

$$Q(D \mid x, a, b_\sigma) = \int_D q(x, a, b_\sigma, y) dy,$$

where q is the transition density

$$q(x, a, b_\sigma, y) := I_{D(x,a)}(y) g_\sigma(y - (x + a)), \tag{7.13}$$

with $D(x, a) := [x + a - \hat{s}, x + a + \hat{s}]$. Furthermore, as

$$-\hat{s} \le y - (x + a) \le \hat{s} \qquad \forall y \in D(x, a),\ x \in X,\ a \in A(x), \tag{7.14}$$

$D(x, a)$ contains the interval $D_1 := [x_* - \hat{s}, x_* - \hat{x} + \hat{s}]$ if $x \le x_*$, and the interval $D_2 := [x_* + \hat{x} - \hat{s}, x_* + \hat{s}]$ if $x \ge x_*$. Hence

$$I_{D(x,a)} \ge I_{D_1} + I_{D_2}.$$

On the other hand, by (7.1), there exists $\varepsilon > 0$ such that $g_\sigma(y-(x+a)) \geq \varepsilon g_\sigma(y)$ for all x, a, y as in (7.14) and all σ in $[\underline{\sigma}, \bar{\sigma}]$. Therefore, the transition density in (7.13) satisfies an inequality of the form (5.27) with

$$q_0(y) := \varepsilon \left[I_{D_1}(y) + I_{D_2}(y) \right] g_\sigma(y).$$

Consequently, by Proposition 5.6, we get Assumption 5.1(a) with $\delta(x,a,b) \equiv 1$.

Summarizing, Assumptions 3.1 and 5.1 hold in the present example, and therefore all of the optimality results in sections 3, 4, and 5 are satisfied. In fact, as we show next, getting the optimal cost function turns out to be surprisingly simple because there exists a *myopic* minimax strategy! (See Remark 7.1 below.)

Computation of the minimax strategy and the optimal costs. Let f_n and $v_{n,\alpha}$ be as in (3.10) with $v_{0,\alpha} \equiv 0$. Then from (7.9), (7.4), and (6.15) we get that

$$f_1(x) := x_* - x, \qquad \text{and} \qquad v_{1,\alpha}(x) = \xi_*^2 \qquad \forall x \in X. \tag{7.15}$$

Similarly, a straightforward induction argument using (3.11) yields that

$$f_n(\cdot) \equiv f_1(\cdot) \qquad \text{and} \qquad v_{n,\alpha}(\cdot) \equiv \xi_*^2 \sum_{j=0}^{n-1} \alpha^j \qquad \forall n \geq 1, \alpha \in (0,1]. \tag{7.16}$$

It follows from Theorem 3.1 that the *constant function* $v_{n,\alpha}$ is the optimal n-stage cost and that the *stationary strategy* f_1 is minimax *for all* $n \geq 1$ *and all* α *in* (0,1].

In turn, from (7.16) and (4.7) it follows that the optimal α-DC is the *constant function* $v_\alpha^*(\cdot) \equiv \xi_*^2/(1-\alpha)$ *for all* $0 < \alpha < 1$, and so, by Theorem 4.2, $f_\alpha^*(\cdot) \equiv f_1(\cdot)$ is an α-DC minimax strategy *for all* $0 < \alpha < 1$.

Finally, from (5.15) and (5.17) we obtain that also the function h^* in (5.4) is constant, and so (5.4) reduces to $\rho^* = v_1(\cdot) \equiv \xi_*^2$. That is, the optimal AC is $\rho^* = \xi_*^2$, and $f^*(\cdot) \equiv f_1(\cdot)$ is a minimax strategy for the AC criterion.

REMARK 7.1. *A minimax strategy is said to be* myopic *if it can be obtained by solving a static (i.e., a one-period) optimization problem. Myopic policies, if they exist, are usually "problem-dependent"—for instance, in the* α*-DC case they depend on* α*. Thus, it is truly surprising that the myopic strategy* f_1 *in* (7.15) *is minimax for all of the problems associated to* (7.5), (7.6), (7.8)*, finite- or infinite-horizon, discounted or average. It is worth noting, on the other hand, that myopic behavior has been observed by many authors (see, e.g.,* [7, 8, 36, 53, 63]*), but as far as we can tell, it has been determined only for* particular *classes of problems, and* a posteriori*, that is, after one has actually solved the problem, as in Example* 7.1. *An interesting* open problem *would be to find a priori conditions—i.e., exclusively based on the problem data—for the existence of myopic strategies. The existing results in that direction* [58, 59, 60] *are not sufficiently general to include, for instance, the minimax model in Example* 7.1.

On the other hand, one might be willing to conjecture that the myopic property of f_1 is due to the fact that the system in (7.5) is linear and that the cost (7.9) is quadratic. These conditions, however, in general are not sufficient to have myopic behavior: for instance, the following example shows another LQ (i.e., linear system, quadratic cost) problem in which the minimax strategies are not myopic.

EXAMPLE 7.2. *Consider the linear system*

$$x_{t+1} = k_1 x_t + k_2 a_t + \xi_t \tag{7.17}$$

with state space $X := \mathbb{R}$*, and disturbances* $\xi_t \equiv \xi(\sigma) \in B$*, as at the beginning of this section. The coefficients* k_1 *and* k_2 *are given positive constants. For each state* x*, the control set* $A(x) \subset A := \mathbb{R}$ *is the interval*

$$A(x) := [-k_1|x|/k_2, k_1|x|/k_2]. \tag{7.18}$$

(Actually, in (7.18) *we may take* $A(x) := [-g_1(x), g_2(x)]$*, where* g_1 *and* g_2 *are nonnegative continuous functions such that* $g_i(x) \geq |k_1 x/k_2|$ *for all* $x \in X$ *and* $i = 1, 2$*. However, the choice in* (7.18) *greatly simplifies the calculations below.) To complete the specification of the minimax model we introduce the cost-per-stage*

$$\hat{c}(x, a, s) := c_1 x^2 + c_2 a^2 + s^2 \qquad \forall (x, a) \in \mathbb{K},\ s \in S,$$

where c_1 *and* c_2 *are positive constants. Thus, by* (7.3) *and* (6.8)*,*

$$c(x, a, b_\sigma) = c_1 x^2 + c_2 a^2 + E[\xi(\sigma)^2]. \tag{7.19}$$

The plan now is as in Example 7.1*: first we will verify the Assumptions* 3.1 *and* 5.1*, and then we will compute the optimal cost function and the minimax strategies for each of the problems in sections* 3*,* 4*,* 5*.*

Verification of Assumption 3.1. As in Example 7.1, part (a) in Assumption 3.1 follows from either Proposition 6.3 or Proposition 6.6. Now let ξ_*^2 be as in (7.4), and define

$$w_1 := c_1 + c_2(k_1/k_2)^2, \qquad \overline{w} := \max\{1, \xi_*^2, w_1\}$$

and, for some $\gamma \geq 2$,

$$w(x) := \overline{w} e^{\gamma|x|} \qquad \forall x \in X. \tag{7.20}$$

Therefore, by (7.18) and (7.19),

$$0 \leq c(x, a, b_\sigma) \leq w_1 x^2 + \xi_*^2 \leq w(x),$$

and so Assumption 3.1(b) holds with $\bar{c} := 1$. Furthermore, Assumption 3.1(c) follows from Proposition 6.2, whereas, as $c(x, a, b_\sigma)$ is nonnegative, Assumption 3.1(d) is not required in the present case (see Remark 3.1(a)). On the other hand, parts (f) and (g) are obvious (see Remark 3.1(c)), and so it only remains to verify Assumption 3.1(e), which can be done as follows.

LEMMA 7.1. *Let* $w(x)$ *be as in* (7.20)*, and suppose that*

$$0 < k_1 < 1/2. \tag{7.21}$$

Then Assumption 3.1(e) *holds with* $\beta := e^{\gamma \hat{s}}$*, i.e.,*

$$\hat{w}(x, a, b_\sigma) \leq \beta w(x), \tag{7.22}$$

where $\hat{w}(x, a, b_\sigma) := \int w(y) Q(dy|x, a, b_\sigma)$*. If in addition* $\gamma \hat{s} < -\log \alpha$*, then we also obtain* (4.6)*, i.e.,* $1 \leq \beta < 1/\alpha$*.*

Proof. By (7.18),

$$|k_1 x + k_2 a| \leq 2k_1|x| \qquad \forall (x, a) \in \mathbb{K}_A. \tag{7.23}$$

On the other hand, by (7.20),

$$\begin{aligned}\hat{w}(x,a,b_\sigma) &= E[w(k_1x+k_2a+\xi_0)]\\ &\le \overline{w}e^{\gamma|k_1x+k_2a|}E(e^{\gamma|\xi_0|})\\ &\le \overline{w}e^{2\gamma k_1|x|}e^{\gamma\hat{s}} \qquad \text{[by (7.23)]}\\ &= w(x)e^{-\gamma|x|}e^{\gamma|x|}\beta \qquad \text{[by (7.21) and (7.20)]}.\end{aligned} \tag{7.24}$$

This gives (7.22). The last statement in the lemma is obvious. □

Verification of Assumption 5.1. Let us replace (7.21) with the stronger condition

$$0 < k_1 < 1/6. \tag{7.25}$$

Consider the intervals $D_0 := [-\hat{s}/2, \hat{s}/2], D_1 := [-\hat{s}/4k_1, \hat{s}/4k_1]$, and let $\delta(x)$ be the indicator function of D_1. Observe that, by (7.23),

$$D_2(x,a) \supset D_0 \qquad \text{if} \qquad x \in D_1, \tag{7.26}$$

with $D_2(x,a) := [k_1x + k_2a - \hat{s}, k_1x + k_2a + \hat{s}]$. Moreover, as in Example 7.1 (see (7.12), (7.13)), one can see that the transition law Q for the system (7.17) has the density

$$q(x,a,b_\sigma,y) = I_{D_2(x,a)}(y)g_\sigma(y-(k_1x+k_2a)).$$

We can now verify Assumption 5.1 with

$$\delta(x,a,b_\sigma) \equiv \delta(x) \qquad \text{and} \qquad \nu(B) := \varepsilon \cdot m(B \cap D_0), \tag{7.27}$$

where $m(dy) = dy$ denotes the Lebesgue measure on X, and $\varepsilon > 0$ is a lower bound for $g_\sigma(\cdot)$ on S. (Observe that a point y is in $D_2(x,a)$ if and only if $y-(k_1x+k_2a)$ is in $S = [-\hat{s}, \hat{s}]$.) Indeed, by (7.26), Assumption 5.1(a) follows because

$$\begin{aligned}Q(B|x,a,b_\sigma) &= \int_B I_{D_2(x,a)}(y)g_\sigma(y-(k_1x+k_2a))dy\\ &\ge \delta(x)\int_B I_{D_0}(y)\cdot\varepsilon dy\\ &= \delta(x)\nu(B).\end{aligned}$$

Further, parts (b) and (d) in Assumption 5.1 are obvious. To verify part (c) we shall use the notation $\hat{w}(x,a,b_\sigma)$ in (7.22). Let us first consider the case $\delta(x) = 0$, that is, $|x| > \hat{s}/4k_1$. Then as in (7.24) we obtain

$$\begin{aligned}\hat{w}(x,a,b_\sigma) &\le w(x)e^{-\gamma|x|(1-2k_1)}e^{\gamma\hat{s}}\\ &\le w(x)e^{-\gamma\hat{s}(1-2k_1)/4k_1}e^{\gamma\hat{s}}\\ &=: \theta w(x)\end{aligned}$$

with $\theta := \exp[-\gamma\hat{s}(1-6k_1)/4k_1] < 1$. A similar calculation gives us

$$\hat{w}(x,a,b_\sigma) \le \theta w(x) + \nu(w)$$

if $\delta(x) = 1$, that is, $|x| \le \hat{s}/4k_1$. This completes the verification of Assumption 5.1.

We next compute the optimal cost functions and the minimax strategies for the system (7.17)–(7.19).

Finite-horizon case. Let $v_{n,\alpha}$ and f_n be as (3.10), (3.11) for $n \geq 1$, with $v_{0,\alpha}(\cdot) \equiv 0$.

PROPOSITION 7.2. *For all $n = 1, 2, \ldots$ and $x \in X$,*

$$v_{n,\alpha}(x) = v_1(n,\alpha)x^2 + v_2(n,\alpha) \qquad \text{and} \qquad f_n(x) = -f(n,\alpha)x \tag{7.28}$$

with coefficients given by $v_1(0,\alpha) = v_2(0,\alpha) = 0$, and for $n \geq 0$,

$$f(n+1,\alpha) := [c_2 + \alpha v_1(n,\alpha)k_2^2]^{-1}\alpha v_1(n,\alpha)k_1k_2, \tag{7.29}$$

$$v_1(n+1,\alpha) := c_1 + c_2 f(n+1,\alpha)^2 + \alpha v_1(n,\alpha)(k_1 - k_2 f(n+1,\alpha))^2, \tag{7.30}$$

$$v_2(n+1,\alpha) := (1 + \alpha v_1(n,\alpha))\xi_*^2 + \alpha v_2(n,\alpha). \tag{7.31}$$

As $c_2 > 0$, (7.29) implies that $|f(n,\alpha)| \leq k_1/k_2$, and therefore $f_n(x)$ is indeed in the interval $A(x)$ (in (7.18)) for all $x \in X$ and $n \geq 1$.

Proof. For $n = 1$, from (3.11), (6.15), and (7.19) we get $v_{1,\alpha}(x) = c_1x^2 + \xi_*^2$ and $f_1(x) = 0$ for all $x \in X$, with ξ_*^2 as in (7.4). That is, (7.28)–(7.31) hold with $v_1(1,\alpha) = c_1, v_2(1,\alpha) = \xi_*^2$, and $f(1,\alpha) = 0$. Now, by induction, let us suppose that (7.28) holds for some $n \geq 1$. Thus, by (7.23),

$$\int_X v_{n,\alpha}(y)Q(dy|x,a,b_\sigma) = v_1(n,\alpha)(k_1x + k_2a)^2 + v_1(n,\alpha)E[\xi(\sigma)^2] + v_2(n,\alpha),$$

and replacing this expression in (3.11) and using (6.15) we obtain

$$v_{n+1,\alpha}(x) = \min_{a \in A(x)} [c_1x^2 + c_2a^2 + \alpha v_1(n,\alpha)(k_1x + k_2a)^2] + \xi_*^2(1 + \alpha v_1(n,\alpha)) + \alpha v_2(n,\alpha).$$

Finally, a straightforward calculation gives

$$v_{n+1,\alpha}(x) = v_1(n+1,\alpha)x^2 + v_2(n+1,\alpha) \qquad \text{and} \qquad f_{n+1}(x) = -f(n+1,\alpha)x$$

with coefficients as in (7.29)–(7.31). □

The infinite-horizon α-discounted case ($0 < \alpha < 1$). Let us suppose that the number β in (7.25) satisfies (4.6). Then, by (7.28) and (4.7), we may "guess" that there is a unique function

$$v^*(x) = v_1(\alpha)x^2 + v_2(\alpha) \qquad \forall x \in X \tag{7.32}$$

in $\mathbb{B}_{lsc}$ that satisfies (4.4), i.e., using (7.28) and (7.19),

$$\begin{aligned} v^*(x) = \min_a \max_b \big\{ & c_1x^2 + c_2a^2 + E[\xi(\sigma)^2] + \alpha v_1(\alpha)(k_1x + k_2a)^2 \\ & + \alpha v_1(\alpha)E[\xi(\sigma)^2] + \alpha v_2(\alpha) \big\}. \end{aligned} \tag{7.33}$$

Moreover, from (6.15), a direct calculation shows that the minimum in (7.33) is attained when $a = f_\alpha(x)$ is the minimax strategy given by

$$f_\alpha(x) = -f(\alpha)x \qquad \forall x \in X, \tag{7.34}$$

with coefficient (cf. (7.29))

$$f(\alpha) := [c_2 + \alpha v_1(\alpha)k_2^2]^{-1}\alpha v_1(\alpha)k_1k_2. \tag{7.35}$$

As $c_2 > 0$, we have $|f(\alpha)| \leq |k_1/k_2|$, and so $f_\alpha(x)$ is indeed in $A(x)$ for all $x \in X$. Substituting (7.34) and (7.4) in (7.33), and then comparing the result with (7.32), we conclude that the coefficients $v_1(\alpha)$ and $v_2(\alpha)$ in (7.32) are given by the equations (similar to (7.30), (7.31))

$$v_1(\alpha) = c_1 + c_2 f(\alpha)^2 + \alpha v_1(\alpha)(k_1 - k_2 f(\alpha))^2, \tag{7.36}$$

$$v_2(\alpha) = (1 + \alpha v_1(\alpha))\xi_*^2 + \alpha v_2(\alpha). \tag{7.37}$$

Observe that (7.37) can be solved for $v_2(\alpha)$ in terms of $v_1(\alpha)$, i.e.,

$$v_2(\alpha) = (1 - \alpha)^{-1}(1 + \alpha v_1(\alpha))\xi_*^2, \tag{7.38}$$

whereas inserting (7.35) in (7.36) we obtain a quadratic equation for $v_1(\alpha)$:

$$\alpha k_2^2 v_1(\alpha)^2 + (c_2 - \alpha c_1 k_2^2 - \alpha c_2 k_1^2)v_1(\alpha) - c_1 c_2 = 0. \tag{7.39}$$

This equation has a *unique positive solution*, which is the value of the coefficient $v_1(\alpha)$ in (7.32), (7.35), and (7.38). In other words, the latter equations and (7.39) give an explicit solution for the α-discounted minimax problem in terms of the coefficients in (7.17) and (7.19).

The average cost (AC) case. We shall assume that (7.25) holds. To solve the minimax problem in the AC case we shall proceed as in the discounted cost problem. Namely, in light of (5.17) and (5.18), taking $\alpha = 1$ in (7.28)–(7.31) we "guess" that the solution $\rho^*, h^*(x)$ of (5.4) consists of some constant ρ^* and a quadratic function $h^*(x) = h_* x^2$ for some constant h_*. (Adding a constant to $h^*(x)$ does not alter the fact that $h^*(x)$ is a solution to (5.4); hence we take such a constant to be zero.) Now, substitution of

$$\rho^* \qquad \text{and} \qquad h^*(x) = h_* x^2 \tag{7.40}$$

in (5.4) yields, with the usual calculations,

$$\rho^* + h_* x^2 = \min_{a \in A(x)} [c_1 x^2 + c_2 a^2 + h_*(k_1 x + k_2 a)^2] + (1 + h_*)\xi_*^2 \tag{7.41}$$

and that the minimum is realized when $a = f^*(x)$ is the minimax strategy

$$f^*(x) = -f_* x, \qquad \text{with} \qquad f_* := (c_2 + h_* k_2^2)^{-1} h_* k_1 k_2. \tag{7.42}$$

Substitution of (7.42) in (7.41) yields that

$$\rho^* = (1 + h_*)\xi_*^2, \tag{7.43}$$

$$h_* = c_1 + c_2 f_*^2 + h_*(k_1 - k_2 f_*)^2. \tag{7.44}$$

Moreover, substitution of (7.42) in (7.44) gives the following quadratic equation for h_*:

$$k_2^2 h_*^2 + (c_2 - c_1 k_2^2 - c_2 k_1^2)h_* - c_1 c_2 = 0, \tag{7.45}$$

and the *unique positive solution* of (7.45) is the coefficient h_* in (7.40), (7.42), and (7.43). Thus, the optimal cost and the minimax strategy for the AC problem are as in (7.43) and (7.42), respectively.

8. Concluding remarks. We have presented in this paper a unified, self-contained study of minimax control problems for discrete-time stochastic systems in Borel spaces, allowing unbounded cost functions. Our results can be extended to maximin problems by modifying Assumption 3.1 in the obvious manner. Our study includes finite- and infinite-horizon problems. In fact, Theorem 4.2 (on the α-discounted case) and Corollary 5.4 (for AC problems) give precise estimates for the convergence of finite-horizon optimal cost functions to their infinite-horizon counterparts. Moreover, for the special case of stochastic systems with unknown disturbance distribution, in which the minimax problem is a "game against nature," we have presented (in section 6) sufficient conditions for the assumptions that ensure our optimality results in sections 3, 4, and 5. These sufficient conditions are used in section 7 to make a detailed analysis of two particular minimax control problems (Examples 7.1 and 7.2).

Finally, in connection with Example 7.1 and Remark 7.1 we should mention once again an important problem that remains open: determining conditions for the existence of *myopic* minimax strategies.

REFERENCES

[1] E. Altman and A. Hordijk, *Zero-sum Markov games and worst-case optimal control of queueing systems*, Queueing Syst. Theory Appl., 21 (1995), pp. 415–447.

[2] E. Altman, A. Hordijk, and F. M. Spieksma, *Contraction conditions for average and α-discount optimality in countable state Markov games with unbounded rewards*, Math. Oper. Res., 22 (1997), pp. 588–618.

[3] M. A. Barrón, R. Aguilar, J. González, and E. Meléndez, *Model-based control of mold level in a steel continuous caster under model uncertainties*, Control Engrg. Practice, 6 (1998), pp. 191–198.

[4] D. P. Bertsekas and S. E. Shreve, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.

[5] G. Bocher, R. Obermann, B. Winkler, G. Kruger, and P. Patte, *Slab quality improvement by means of advanced mould level control*, in Proceedings of the 1st European Conference on Continuous Casting, Florence, Italy, Institute of Materials, 1991, pp. 1205–1214.

[6] S. P. Coraluppi and S. I. Marcus, *Mixed risk-neutral/minimax control of discrete-time finite-state Markov decision process*, IEEE Trans. Automat. Control, 45 (2000), pp. 528–532.

[7] Y. M. I. Dirickx and L. P. Jennergren, *On the optimality of myopic policies in sequential decision problems*, Manage. Sci., 21 (1975), pp. 550–556.

[8] C. A. J. M. Dirven and O. J. Vrieze, *Advertising models, stochastic games and myopic strategies*, Oper. Res., 34 (1986), pp. 645–649.

[9] M. Dussud, S. Galichet, and L. P. Foulloy, *Application of fuzzy logic control for continuous casting mould level control*, IEEE Trans. Control Systems Tech., 6 (1998), pp. 246–256.

[10] J. Filar and K. Vrieze, *Competitive Markov Decision Processes*, Springer-Verlag, New York, 1997.

[11] M. K. Ghosh and A. Bagchi, *Stochastic games with average payoff criterion*, Appl. Math. Optim., 38 (1998), pp. 283–301.

[12] X. P. Guo and O. Hernández-Lerma, *Zero-Sum Games for Continuous-Time Markov Games with a Discounted Payoff Criterion*, preprint, 2000.

[13] X. P. Guo and O. Hernández-Lerma, *Nonzero-Sum Games for Continuous-Time Markov Chains with Unbounded Discounted Payoffs*, preprint, 2001.

[14] X. P. Guo and O. Hernández-Lerma, *Zero-Sum Games for Nonhomogenous Markov Chains with Expected Average Payoff Criterion*, preprint, 2002.

[15] O. Hernández-Lerma, *Adaptive Markov Control Processes*, Springer-Verlag, New York, 1989.

[16] O. Hernández-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer-Verlag, New York, 1996.

[17] O. Hernández-Lerma and J. B. Lasserre, *Further Topics in Discrete-Time Markov Control Processes*, Springer-Verlag, New York, 1999.

[18] O. Hernández-Lerma and J. B. Lasserre, *Zero-sum stochastic games in Borel spaces: Average payoff criteria*, SIAM J. Control Optim., 39 (2001), pp. 1520–1539.

[19] O. Hernández-Lerma, R. Montes-de-Oca, and R. Cavazos-Cadena, *Recurrence conditions*

for Markov decision processes with Borel state space: A survey, Ann. Oper. Res., 28 (1991), pp. 29–46.

[20] T. Hesketh, D. J. Clements, and R. Williams, *Adaptive mould level control for continuous steel slab casting*, Automatica, 29 (1993), pp. 851–864.

[21] N. Hilgert and J. A. Minjárez-Sosa, *Adaptive policies for time-varying stochastic systems under discounted criterion*, Math. Methods Oper. Res., 54 (2001), pp. 491–505.

[22] A. Hordijk, O. Passchier, and F. M. Spieksma, *Optimal control against worst case admission policies: A multichained stochastic game*, Math. Methods Oper. Res., 45 (1997), pp. 281–301.

[23] A. Hordijk and A. A. Yushkevich, *Blackwell optimality in the class of all policies in Markov decision chains with Borel state space and unbounded rewards*, Math. Methods Oper. Res., 49 (1999), pp. 421–448.

[24] R. Jagannathan, *A minimax ordering policy for the infinite stage dynamic inventory problem*, Manage. Sci., 24 (1978), pp. 1138–1149.

[25] A. Jaśkiewicz and A. S. Nowak, *On the optimality equation for zero-sum ergodic stochastic games*, Math. Methods Oper. Res., 54 (2001), pp. 291–301.

[26] H. Kato and M. Yamasita, *New automation and control technology of slab caster*, in Proceedings of the IFAC Symposium on Automation in Mining, Mineral and Metal Processing (MMM'86), Buenos Aires, Argentina, International Federation of Automatic Control (IFAC), 1986, pp. 253–258.

[27] R. M. C. de Keyser, *Improved mold-level control in a continuous steel casting line*, Control Engrg. Practice, 5 (1997), pp. 231–237.

[28] H. Kitada, O. Kondo, H. Kusachi, and K. Sasame, *H^∞ Control of molten steel level in continuous caster*, IEEE Trans. Control Systems Tech., 6 (1998), pp. 200–207.

[29] N. Kiupel, P. M. Frank, and J. Wochnik, *Improvement of mold-level control using fuzzy logic*, Eng. Appl. Artif. Intell., 7 (1994), pp. 493–499.

[30] G. Kruger, *Advanced mould level control for continuous casting plants*, Metallurgical Plant and Technol., 3 (1985), pp. 42–49.

[31] H.-U. Küenle, *On multichain Markov games*, in Advances in Dynamic Games and Applications, Ann. Internat. Soc. Dynam. Games 6, Birkhäuser, Boston, 2001, pp. 147–163.

[32] H.-U. Küenle, *Stochastiche Spiele und Entscheidungsmodelle*, B. G. Teubner, Leipzig, 1986.

[33] H.-U. Küenle, *On the optimality of (s,S)-strategies in a minimax inventory model with average cost criterion*, Optimization, 22 (1991), pp. 123–138.

[34] H.-U. Küenle, *Stochastic games with complete information and average cost criteria*, in Advances in Dynamic Games and Applications, J. A. Filar, V. Gaitsgory, and K. Mizukami, eds., Birkhäuser, Boston, 2000, pp. 325–338.

[35] M. Kurano, *Minimax strategies for average cost stochastic games with an application to inventory models*, J. Oper. Res. Soc. Japan, 30 (1987), pp. 232–247.

[36] W. S. Lovejoy, *Myopic policies for some inventory models with uncertain demand distributions*, Manage. Sci., 36 (1990), pp. 724–738.

[37] A. Maitra and W. Sudderth, *Finitely additive and measurable stochastic games*, Internat. J. Game Theory, 22 (1993), pp. 201–223.

[38] A. Maitra and W. Sudderth, *Borel stochastic games with lim sup payoff*, Ann. Probab., 21 (1993), pp. 861–885.

[39] A. Maitra and W. Sudderth, *Discrete Gambling and Stochastic Games*, Springer-Verlag, New York, 1996.

[40] A. Maitra and W. Sudderth, *Finitely additive stochastic games with Borel measurable payoffs*, Internat. J. Game Theory, 27 (1998), pp. 257–267.

[41] D. A. Martin, *The determinacy of Blackwell games*, J. Symbolic Logic, 63 (1998), pp. 1565–1581.

[42] J. A. Minjárez-Sosa, *Nonparametric adaptive control for discrete-time Markov processes with unbounded costs under the average criterion*, Appl. Math. (Warsaw), 26 (1999), pp. 267–280.

[43] A. S. Nowak, *Minimax selection theorems*, J. Math. Anal. Appl., 103 (1984), pp. 106–116.

[44] A. S. Nowak, *Measurable selection theorems for minimax stochastic optimization problems*, SIAM J. Control Optim., 23 (1985), pp. 466–476.

[45] A. S. Nowak, *Zero-sum average payoff stochastic games with general state space*, Games Econ. Behavior, 7 (1994), pp. 221–232.

[46] A. S. Nowak, *Optimal strategies in a class of zero-sum ergodic stochastic games*, Math. Methods Oper. Res., 50 (1999), pp. 399–419.

[47] J. Paiuk, A. Zanini, M. Remorino, and O. Frola, *The automatic mould level control for a continuous casting process*, in Proceedings of the IFAC Symposium on Automation in

Mining, Mineral and Metal Processing (MMM'89), Buenos Aires, Argentina, International Federation of Automatic Control (IFAC), 1989, pp. 205–208.

[48] O. Passchier, *The Theory of Markov Games and Queueing Control*, Ph.D. thesis, Dept. of Mathematics and Computer Science, Leiden University, The Netherlands, 1996.

[49] U. Rieder, *Measurable selection theorems for optimization problems*, Manuscripta Math., 24 (1978), pp. 115–131.

[50] U. Rieder, *On Semi-Continuous Dynamic Games*, Technical Report, University of Karlsruhe, Germany, 1978.

[51] U. Rieder, *Non-cooperative dynamic games with general utility functions*, in Stochastic Games and Related Topics, T. E. S. Raghavan et al., ed., Kluwer, Dordrecht, The Netherlands, 1991, pp. 161–174.

[52] U. Rieder, *Average optimality in Markov games with general state space*, in Proceedings of the 3rd International Conference on Approximation Theory and Optimization, Puebla, Mexico, 1995. Available online from www.emis.de/proceedings/.

[53] D. B. Rosenfield, *Optimality of myopic policies in disposing excess inventory*, Oper. Res., 40 (1992), pp. 800–803.

[54] Y. Sasabe, S. Kubota, A. Koyama, and H. Miki, *Real-time expert system applied to mould bath level control of continuous caster*, ISIJ International, 30 (1990), pp. 136–141.

[55] M. Schäl, *Conditions for optimality and for the limit of n-stage optimal policies to be optimal*, Z. Wahrsch. Verw. Gebiete, 32 (1975), pp. 179–196.

[56] L. I. Sennott, *Zero-sum stochastic games with unbounded cost: Discounted and average cost cases*, Z. Oper. Res., 39 (1994), pp. 209–225.

[57] R. Serfozo, *Convergence of Lebesgue integrals with varying measures*, Sankhyā Ser. A, 44 (1982), pp. 380–402.

[58] M. J. Sobel, *Myopic solutions of Markov decision processes and stochastic games*, Oper. Res., 29 (1981), pp. 945–1009.

[59] M. J. Sobel, *Myopic solutions of affine dynamic models*, Oper. Res., 38 (1990), pp. 847–853.

[60] L. Tesfatsion, *Global and approximate global optimality of myopic economic decisions*, J. Econom. Dynamics Control, 2 (1980), pp. 135–160.

[61] O. Vega-Amaya, *The Average Cost Optimality Equation: A Fixed Point Approach*, preprint; available from http://fractus.mat.uson.mx/˜tedi/reportes

[62] J. H. Watters, G. S. Jiang, C. J. Treadgold, and A. Kumar, *Model studies of surface wave phenomena in the continuous casting mould*, in Proceedings of the 2nd European Continuous Casting Conference, Vol. 1, Düsseldorf, Germany, Institute of Materials, 1994, pp. 95–102.

[63] J. Weishaupt, *Optimal myopic policies and index policies for stochastic scheduling problems*, Z. Oper. Res., 40 (1994), pp. 75–89.

[64] J. Wessels, *Markov games with unbounded rewards*, in Dynamische Optimierung, Bonner Math. Schriften 98, M. Schäl, ed., University of Bonn, Germany, 1977, pp. 133–147.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 6, pp. 1661–1695

ROBUST DISSIPATIVITY OF INTERVAL UNCERTAIN LINEAR SYSTEMS*

FLORIN DAN BARB†, AHARON BEN-TAL‡, AND ARKADI NEMIROVSKI‡

Abstract. In this paper we are concerned with the problem of robust dissipativity of linear systems with parameters affected by box uncertainty; our major goal is to evaluate the largest uncertainty level for which all perturbed instances share a common dissipativity certificate. While it is NP-hard to compute this quantity exactly, we demonstrate that under favorable circumstances one can build an $O(1)$-tight lower bound of this "intractable" quantity by solving an explicit semidefinite program of the size polynomial in the size of the system. We consider a number of applications, including the robust versions of the problems of extracting nearly optimal available storage, providing nearly optimal required supply, Lyapunov stability analysis, and linear-quadratic control.

Key words. interval matrices, dissipative linear systems, interval uncertain LMIs, positive-real systems, contractive systems, Riccati equation, semidefinite programming

AMS subject classifications. 15A15, 15A09, 15A23

PII. S0363012901398174

1. Introduction and motivation.

An important requirement of any modern control system is its robustness. In many system theory and control applications, the concept of robustness is related to the stability of the closed-loop system and its performance measured with respect to a certain objective function.

In this paper, we focus on robustness with respect to unknown-but-bounded (and possibly time-varying) perturbations of the entries in the matrix $\Sigma = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$ of a continuous-time linear dynamical system

$$\begin{aligned} \dot{z} &= Az + Bu, \\ y &= Cz + Du. \end{aligned}$$

For the time being, we assume the simplest *interval* model of perturbations—every entry Σ_{ij} in Σ, independently of all other entries, can vary in the interval $\Sigma_{ij} \pm \rho d\Sigma_{ij}$, where Σ_{ij} are the nominal data, $d\Sigma_{ij}$ are given scale factors, and ρ is the uncertainty level. The set of matrices just defined will be referred to as *interval matrix* and will be denoted by $\mathcal{U}_\rho$.

The question we are addressing is as follows:

(?) *What is the supremum $\rho^\star$ of those uncertainty levels ρ under which all perturbations of level ρ preserve a particular property of the system, such as stability, passivity, contractiveness, etc.?*

Typically, it is computationally intractable to give a *precise* answer to such a question. For example, it is known to be NP-hard to check the stability of all instances of an interval matrix $\mathcal{U}_\rho$ [7]. In other words, we do not know how to check efficiently whether every one of the Lyapunov linear matrix inequalities (LMIs)

*Received by the editors November 14, 2001; accepted for publication (in revised form) August 9, 2002; published electronically February 4, 2003.
http://www.siam.org/journals/sicon/41-6/39817.html

†Delft University of Technology, Faculty of Information Technology and Systems, Mekelweg 4, 2628 CD Delft, The Netherlands (f.d.barb@its.tudelft.nl).

‡William Davidson Faculty of Industrial Engineering and Management, Technion—Israel Institute of Technology, Technion City, Haifa 32000, Israel (abental@ie.technion.ac.il, nemirovs@ie.technion.ac.il).

$$\text{find } X \text{ such that } X \succ 0 \text{ and } A^T X + XA \prec 0^1$$

corresponding to the instances A of an interval square matrix is solvable.

The situation does not improve at all when we pass from the question "whether all instances of an interval matrix are stable" to the seemingly simpler question "whether all instances of an interval matrix admit a common quadratic Lyapunov stability certificate," or, which is the same, whether the aforementioned LMIs have a *common* solution. Although in the new form the question is to check the solvability of a *finite* system of LMIs

$$X \succ 0, \quad A^T X + XA \prec 0 \quad \forall (A \in \mathcal{V}),$$

where $\mathcal{V}$ is the (finite!) set of the extreme points of the original interval matrix, the number of LMIs in this system blows up exponentially with the size of the matrix (unless the number of uncertain entries in the matrix remains once and for ever fixed). It turns out that in general it is already NP-hard to check whether a given candidate solution X is feasible for the above system of LMIs.[2]

The difficulty arising when checking stability of (all instances of) an interval matrix is typical for other problems of the aforementioned type: the property of interest is equivalent to the solvability of certain LMI $\mathcal{L}_\Sigma(X) \succ 0$ with the data coming from the matrix Σ of the system in question. When Σ is subject to interval uncertainty, both of the following tasks become NP-hard:

(1.A) Checking whether every one of the LMIs

$$\mathcal{L}_\Sigma(X) \succ 0 \tag{1}$$

with $\Sigma \in \mathcal{U}_\rho$ is solvable (i.e., to verify that the desired property is possessed by all instances), and

(1.B) Checking whether the infinite system of LMIs

$$\mathcal{L}_\Sigma(X) \succ 0 \qquad \forall (\Sigma \in \mathcal{U}_\rho)$$

is solvable, i.e., whether all instances of our interval matrix share a common certificate for the property of interest (which normally is a *sufficient* condition for the property to be preserved also by dynamic perturbations).

Now, in light of the fact that it is NP-hard to answer questions (1.A), (1.B) *exactly*, a natural course of action is to *relax* the questions in order to make them tractable. We are not aware of any good relaxation of question (1.A). In contrast to this, recent progress in what is called robust semidefinite programming [1, 5, 6] (specifically, the matrix cube theorem [3]) leads to "tight" tractable relaxations of question (1.B). It turns out that *under favorable circumstances* (which do take place for a wide family of "properties of interest") *one can build efficiently a lower bound* $\widehat{\rho}$ *on the supremum* $\rho^\star$ *of those uncertainty levels* ρ *for which the answer to the question* (1.B) *is affirmative, and this lower bound is tight within an absolute constant factor* (the latter is in most of the cases $\frac{\pi}{2} = 1.57\ldots$). The goal of this paper is to justify the above claim.

[1] We write $A \succeq B$ ($A \succ B$) to express that A, B are symmetric matrices of the same size such that $A - B$ is positive semidefinite (respectively, positive definite).

[2] This "analysis" problem is not simpler than checking whether all instances of a given interval symmetric matrix are positive semidefinite; it is shown in [7] that the latter problem is NP-hard already in the case when all entries in the interval matrix, except for those from the first two rows and columns, are fixed.

A convenient general framework for our study is the dissipativity-based approach, as developed in the seminal papers of Willems [9, 10]. The notion of dissipativity is one of the most important concepts in systems and control theory, both from the theoretical point of view as well as from the practical perspective. In many mechanical and electrical engineering applications, dissipativity is related to the notion of energy. Here, a dissipative system is characterized by the following property: at any moment of time, the amount of energy which the system can supply to its environment cannot exceed the amount of energy that has been supplied to it. However, the dissipativity-based framework is not restricted to the energy-related issues; it allows us to investigate stability analysis and linear-quadratic control as well.

The rest of the paper is organized as follows. In section 2, we review basic notions and results from dissipativity theory. In section 3, we present the box model of uncertainty (which is slightly more general than the simple interval model) and pose and motivate three basic dissipativity-related versions of question (1.B): finding a common dissipativity certificate for all instances of a given uncertain system (a particular case of this problem is the Lyapunov stability analysis under box uncertainty); extracting available storage/providing required supply in the face of uncertainty (this covers, in particular, the optimal linear-quadratic control of uncertain systems). In the central section 4, we develop "tractable tight relaxations" of the problems posed in section 3. Finally, in section 5, we present several illustrating numerical examples.

On many occasions in this paper we use the term "efficient computability" of various quantities. An appropriate definition of this notion does exist,[3] but for our purposes here it suffices to agree that all "LMI-representable" quantities—those which can be represented as optimal values in semidefinite programs

$$\min_x \left\{ c^T x : A_0 + \sum_{i=1}^N x_i A_i \succeq 0 \right\}$$

or generalized eigenvalue problems

$$\min_{x,\omega} \left\{ \omega : \begin{array}{c} A(x) \equiv A_0 + \sum_{i=1}^N x_i A_i \succeq 0 \\ B(x) \equiv B_0 + \sum_{i=1}^N x_i B_i \preceq \omega A(x) \\ C(x) \equiv C_0 + \sum_{i=1}^N x_i C_i \succeq 0 \end{array} \right\}$$

—are efficiently computable functions of the data $c, \{A_i \in \mathbf{S}^n\}_{i=0}^N$, respectively, $\{A_i, B_i, C_i \in \mathbf{S}^n\}_{i=0}^N$; where $\mathbf{S}^n$ is the space of real symmetric $n \times n$ matrices. From now on, missing blocks in block matrices are assumed to be zero.

2. Dissipative systems. In this section, we shall briefly review the dissipativity theory for linear systems with quadratic storage and supply functions as developed in [10]. The readers less familiar with the topic are referred to [8] for details.

Consider a continuous-time linear time-invariant dynamical system given by

$$\begin{array}{ll} \dot{z}(t) = Az(t) + Bu(t), & z(0) = \zeta, \\ y(t) = Cz(t) + Du(t), & \end{array} \tag{2}$$

[3]For a definition which fits best of all the contents of the paper, see [2, Chapter 5].

where $\Sigma \equiv \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right] \in \mathbf{R}^{(n+p)\times(n+m)}$ is the matrix of system coefficients, $u(\cdot) \in \mathbf{R}^m$ is the *input* (which henceforth is assumed to be locally square integrable), $z(\cdot) \in \mathbf{R}^n$ is the *state*, and $y(\cdot) \in \mathbf{R}^p$ is the *output*. In what follows we refer to system (2) given by a matrix Σ as "system Σ."

Let us fix a quadratic *supply function*

$$\mathfrak{S} : \mathbf{R}^{p+m} \to \mathbf{R}, \qquad \mathfrak{S}(y,u) = \begin{bmatrix} y \\ u \end{bmatrix}^T \begin{bmatrix} Q & L \\ L^T & R \end{bmatrix} \begin{bmatrix} y \\ u \end{bmatrix}; \tag{3}$$

here

$$\mathfrak{P} = \begin{bmatrix} Q & L \\ L^T & R \end{bmatrix}$$

is a symmetric *supply matrix* (Q is $p\times p$, R is $m\times m$). Given a trajectory $(z(\cdot), y(\cdot), u(\cdot))$ of (2) and two time instants $t_0 \le t_1$, we interpret the corresponding *supply*

$$\int_{t_0}^{t_1} \mathfrak{S}(y(t), u(t))dt$$

as the work carried on the system in the time interval $[t_0, t_1]$ along the trajectory in question, if the supply is nonnegative, and as minus the energy extracted from the system, if the supply is negative.

Note that along a trajectory of (2) the supply can be expressed in terms of the state and the input:

$$\begin{aligned} \mathcal{S}_\Sigma(z,u) &\equiv \mathfrak{S}(Cz+Du,u) \\ &= \begin{bmatrix} z \\ u \end{bmatrix}^T \underbrace{\begin{bmatrix} C^TQC & C^T(L+QD) \\ (L+QD)^TC & D^TQD + L^TD + D^TL + R \end{bmatrix}}_{\mathbf{S}_\Sigma} \begin{bmatrix} z \\ u \end{bmatrix}. \end{aligned} \tag{4}$$

Definition 2.1. *System Σ is called dissipative with respect to supply $\mathfrak{S}$, if there exists a nonnegative storage function $V(z)$, $V(0) = 0$, such that*

$$V(z(0)) + \int_0^T \mathfrak{S}(y(t), u(t))dt \ge V(z(T)) \tag{5}$$

for all $T \ge 0$ and all trajectories $(z(\cdot), y(\cdot), u(\cdot))$ of the system.

The standard interpretation of a storage function is that $V(z)$ is the internal energy stored by system in state z; with this interpretation, (5) means that the work W on the system needed to move it from one state to another is at least the resulting change ΔV in the internal energy stored by the system; the excess $W - \Delta V \ge 0$ is thought of to be dissipated by the system.

The summary of facts on dissipativity we need in what follows is as follows. *Assume that system Σ is controllable*, and let $\mathfrak{S}$ be a quadratic supply:

D.1. $(\Sigma, \mathfrak{S})$ is dissipative if and only if $(\Sigma, \mathfrak{S})$ admits a quadratic storage function $V(z) = z^TZz$, where $Z \in \mathbf{S}^n_+$ (from now on, $\mathbf{S}^n_+$ is the cone of positive semidefinite matrices from $\mathbf{S}^n$).

D.2. A quadratic function $V(z) = z^TZz$ is a storage function for $(\Sigma, \mathfrak{S})$ if and only if $Z \in \mathbf{S}^n_+$ and

$$\mathfrak{S}(y(t), u(t)) - \frac{d}{dt}(z^T(t)Zz(t)) \ge 0$$

for all trajectories $(z(\cdot), y(\cdot), u(\cdot))$, or, which is the same, if and only if $Z \in \mathbf{S}^n$ solves the system of matrix inequalities (MIs)

$$Z \succeq 0, \tag{6a}$$

$$\mathbb{D}_\Sigma[Z] \equiv \mathbf{S}_\Sigma - \begin{bmatrix} A^T Z + ZA & ZB \\ B^T Z & \end{bmatrix} \succeq 0 \tag{6b}$$

(for notation, see (4)). Note that MI (6b) expresses a very transparent requirement that

$$\mathcal{S}_\Sigma(z(t), u(t)) \geq \frac{d}{dt}(z^T(t) Z z(t)) \tag{7}$$

for all t and all trajectories $(z(t), y(t), u(t))$ of Σ. In what follows, we call the solutions of (6) the *dissipativity certificates* for $(\Sigma, \mathfrak{S})$.

D.3. If $(\Sigma, \mathfrak{S})$ is dissipative, then we have the following:

(a) among the associated storage functions there exist the (pointwise) minimal one,

$$V_{av}(z) = \sup_{(z(\cdot),y(\cdot),u(\cdot))} \left\{ -\int_0^{t_1} \mathfrak{S}(y(t), u(t))dt : \begin{array}{l} (z(\cdot), y(\cdot), u(\cdot)) \text{ is a trajectory,} \\ z(0) = z \end{array} \right\}$$

("available storage"), and the (pointwise) maximal one,

$$V_{req}(z) = \inf_{(z(\cdot),y(\cdot),u(\cdot))} \left\{ \int_0^{t_1} \mathfrak{S}(y(t), u(t))dt : \begin{array}{l} (z(\cdot), y(\cdot), u(\cdot)) \text{ is a trajectory,} \\ z(0) = 0, z(t_1) = z \end{array} \right\}$$

("required supply"). Every storage function $V(\cdot)$ for $(\Sigma, \mathfrak{S})$ satisfies the relations $V_{av}(z) \leq V(z) \leq V_{req}(z)$ for all z, and every convex combination of $V_{av}(\cdot)$ and $V_{req}(\cdot)$ is a storage function for $(\Sigma, \mathfrak{S})$.

(b) Both the available storage and the required supply are quadratic functions of the state:

$$\begin{array}{rcl} V_{av}(z) & = & z^T Z_{av} z, \\ V_{req}(z) & = & z^T Z_{req} z, \end{array}$$

where the positive semidefinite matrices Z_{av}, Z_{req} are, respectively, the $\succeq$-minimal and the $\succeq$-maximal solutions of (6). The set of solutions to (6) is exactly the "matrix interval" $\{Z : Z_{av} \preceq Z \preceq Z_{req}\}$.

D.4. Assume that $(\Sigma, \mathfrak{S})$ is dissipative and that the matrix $D^T Q D + L^T D + D^T L + R$ is positive definite. Then the state feedback

$$u = F_{av} z, \quad F_{av} = -(D^T Q D + L^T D + D^T L + R)^{-1}(B^T Z_{av} - (L + QD)^T C)$$

stabilizes the system (i.e., the real parts of all eigenvalues of the matrix $A + BF_{av}$ of the closed-loop system are negative), and with this feedback, the energy extracted from the system, the initial state of the system being $\zeta \in \mathbb{R}^n$, is exactly the available storage $V_{av}(\zeta)$:

$$-\int_0^\infty \mathfrak{S}(y(t), u(t))dt = \zeta^T Z_{av} \zeta,$$

where $(y(t), u(t))$ are given by

$$\begin{aligned} \dot{z}(t) &= Az(t) + Bu(t), \qquad z(0) = \zeta, \\ u(t) &= F_{av} z(t), \\ y(t) &= Cz(t) + Du(t). \end{aligned}$$

Similarly, the state feedback

$$u = F_{req} z, \quad F_{req} = -(D^T Q D + L^T D + D^T L + R)^{-1}(B^T Z_{req} - (L + QD)^T C)$$

stabilizes the "backward time" system (i.e., the real parts of all eigenvalues of the matrix $-(A + BF_{req})$ of the closed-loop system with backward time are negative), and with this feedback the supply required to move the system from the origin to a state ζ is exactly the required supply $V_{req}(\zeta)$:

$$\int_0^\infty \mathfrak{S}(y(t), u(t))dt = \zeta^T Z_{req} \zeta,$$

where $(y(t), u(t))$ are given by

$$\begin{aligned} \dot{z}(t) &= -[Az(t) + Bu(t)], \qquad z(0) = \zeta, \\ u(t) &= F_{req} z(t), \\ y(t) &= Cz(t) + Du(t). \end{aligned}$$

Let us list several important examples of supply functions.

EXAMPLE 1 (positive-real systems). *Here $m = p$, and the supply matrix is $\mathfrak{P} = \left[\begin{smallmatrix} 0 & I \\ I & 0 \end{smallmatrix}\right]$*, i.e.,

$$\mathfrak{S}(y, u) = 2y^T u.$$

Assuming that A is stable and (A, B, C) is minimal, the pair $(\Sigma, \mathfrak{S})$ is dissipative if and only if (2) is passive, i.e., $\int_0^T y^T(t) v(t) dt \geq 0$ for all $T \geq 0$ and all trajectories $(z(t), y(t), v(t))$ with $z(0) = 0$. Under the same assumptions on A, B, C, the frequency domain characterization of passivity is that the transfer function

$$H(s) = C(sI - A)^{-1} B + D$$

of the system is such that

$$\Re(s) \geq 0 \Rightarrow H(s) + H^*(s) \succeq 0,$$

where $H^*(s)$ is the Hermitian conjugate of $H(s)$ and $\Re(s)$ is the real part of $s \in \mathbb{C}$.

EXAMPLE 2 (nonexpansive systems [4]). *Here the supply matrix is $\mathfrak{P} = \left[\begin{smallmatrix} I_p & 0 \\ 0 & -I_m \end{smallmatrix}\right]$ (I_k is the $k \times k$ unit matrix), i.e.,*

$$\mathfrak{S}(y, u) = u^T u - y^T y.$$

Assuming again that A is stable and (A, B, C) is minimal, dissipativity of $(\Sigma, \mathfrak{S})$ is equivalent to the fact that

$$\int_0^T y^T(t) y(t) dt \leq \int_0^T u^T(t) u(t) dt$$

for all $T \geq 0$ and trajectories $(z(t), y(t), u(t))$ of (2) with $z(0) = 0$. Under the same assumption on A, B, C, the frequency domain characterization of nonexpansivity is that the transfer function $H(s)$ of the system is such that

$$\Re(s) \geq 0 \Rightarrow H^*(s)H(s) \preceq I.$$

EXAMPLE 3 (linear-quadratic control [4]). *Here the supply matrix $\mathfrak{P}$ is positive semidefinite. Assuming that (A, B) is controllable, the pair $(\Sigma, \mathfrak{S})$ is always dissipative, with the available storage $V_{av}(z) \equiv 0$. The required supply $V_{req}(z)$ is the optimal value in the problem of optimal control where the goal is to minimize $\int_0^T \mathfrak{S}(y(t), u(t))dt$ when moving the system from the origin at time 0 to the state z at time T (to be chosen).*

3. Dissipativity under uncertainty. Now assume that the linear dynamic system in question is *uncertain*, so that all we know about the matrix Σ is that Σ belongs to a given *uncertainty set* $\mathcal{U}_\rho$ in the space of $(n+p) \times (m+p)$ real matrices. In this paper we focus on the case of *box uncertainty*:

$$\mathcal{U}_\rho = \left\{ \Sigma = \Sigma + \sum_{\ell=1}^{L} u_\ell d\Sigma_\ell : -\rho \leq u_\ell \leq \rho,\ \ell = 1, \ldots, L \right\}, \tag{8}$$

where

- $\Sigma = \left[\begin{smallmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{smallmatrix}\right]$ is the *nominal* system;
- $d\Sigma_\ell = \left[\begin{smallmatrix} dA_\ell & dB_\ell \\ dC_\ell & dD_\ell \end{smallmatrix}\right]$, $\ell = 1, \ldots, L$, are *basic perturbation matrices*;
- $\rho > 0$ is the *uncertainty level*.

In what follows, we refer to matrices $\Sigma \in \mathcal{U}_\rho$ as to *instances* of the uncertain system associated with the uncertainty set $\mathcal{U}_\rho$.

Let us fix a quadratic supply function (3); in what follows, when speaking about the dissipativity of a certain system, we mean the dissipativity with respect to this supply function. We assume from now on that the nominal pair $(\mathbf{A}, \mathbf{B})$ is controllable, and the nominal system Σ is dissipative, with the minimal and maximal dissipativity certificates $\mathbf{Z}_{av}$, $\mathbf{Z}_{req}$, respectively.

We intend to focus on three dissipativity-related problems for uncertain systems, specifically, the following problems:

1. Common dissipativity certificate. Find a common dissipativity certificate for all instances of the uncertain system.
2. Extracting available storage. Given $\epsilon \in (0, 1)$, find a feedback which stabilizes all instances of the uncertain system and allows us to extract from the initial state ζ of any instance energy at least $(1 - \epsilon)\zeta^T \mathbf{Z}_{av} \zeta$.
3. Providing required supply. Given $\delta > 0$, find a feedback which stabilizes in backward time all instances of the uncertain system and allows to move every instance from the origin to a given state ζ with total supply at most $(1 + \delta)\zeta^T \mathbf{Z}_{req} \zeta$.

Our next goal is to motivate and to model the outlined problems.

3.1. Common dissipativity certificate. The problem of finding a common dissipativity certificate for all instances of an uncertain system is as follows.

PROBLEM 1. *Given a supply $\mathfrak{S}$, a convex set $\mathcal{I}$ in the cone $\mathbf{S}^n_+$, and the data specifying $\mathcal{U}_\rho$, find the supremum of those $\rho \geq 0$ for which all instances from $\mathcal{U}_\rho$ admit a common dissipativity certificate in $\mathcal{I}$, or, which is the same in view of* D.2*, find the*

supremum of those $\rho \geq 0$ for which the system of constraints

$$Z \in \mathcal{I}, \tag{9a}$$

$$\mathbb{D}_\Sigma[Z] \equiv \begin{bmatrix} C^TQC - A^TZ - ZA & C^T(L+QD) - ZB \\ (L+QD)^TC - B^TZ & D^TQD + L^TD + D^TL + R \end{bmatrix} \succeq 0 \tag{9b}$$

$$\forall \Sigma = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \in \mathcal{U}_\rho$$

(see (6)*) in matrix variable Z is solvable.*

The motivation behind Problem 1 is quite transparent: there are cases when the dissipativity is a highly desirable property, and in these cases it is worthy of knowing what are the largest perturbations which for sure preserve this property. With this motivation, however, it remains unclear why we should be interested in a *common* dissipativity certificate for all $\Sigma \in \mathcal{U}_\rho$ rather than to ask what is the largest ρ for which every instance from $\mathcal{U}_\rho$ admits a dissipativity certificate (perhaps depending on the instance). The motivation behind seeking a common dissipativity certificate comes from the fact that such a certificate ensures dissipativity of the uncertain *time-varying* system

$$\begin{array}{l} \dot{z}(t) = A(t)z(t) + B(t)u(t), \\ y(t) = C(t)z(t) + D(t)u(t), \end{array} \tag{10}$$

where the dependence of $\Sigma(t) \equiv \left[\begin{smallmatrix} A(t) & B(t) \\ C(t) & D(t) \end{smallmatrix}\right]$ on t is not known in advance; all we know is that $\Sigma(t)$ is a measurable function of t taking values in $\mathcal{U}_\rho$. The precise meaning of the claim "existence of a common dissipativity certificate for all instances $\Sigma \in \mathcal{U}_\rho$ implies dissipativity of the uncertain time-varying system (10)" is given by the following simple statement.

PROPOSITION 3.1. *Let Z be a common dissipativity certificate for all instances $\Sigma \in \mathcal{U}_\rho$, i.e., let $Z \succeq 0$ satisfy* (9b)*. Then for every $T \geq 0$ and every trajectory $(z(t), y(t), u(t))$ of the time-varying system* (10) *with $\Sigma(t) \in \mathcal{U}_\rho$ for all t, one has*

$$z^T(0)Zz(0) + \int_0^T \mathfrak{S}(y(t), u(t))dt \geq z^T(T)Zz(T). \tag{11}$$

Proof. It is immediately seen that (9b) implies that

$$\mathfrak{S}(y(t), u(t)) \geq \frac{d}{dt}(z^T(t)Zz(t))$$

for all t. Integrating this inequality, we arrive at (11). ∎

EXAMPLE 4 (Lyapunov stability analysis under box uncertainty). *Assume that we have designed a controller for a linear dynamical system, and let*

$$\dot{z} = Az$$

be the description of the closed-loop system (so that some components of z represent states of the plant, while the remaining components of z represent states of the controller). After the design is completed, a natural question is how the performance of the system can be affected by perturbations in A (i.e., in the parameters of the plant and of the controller). Assuming a box model of perturbations

$$A \in \mathcal{V}_\rho = \left\{ A = \mathbf{A} + \sum_{\ell=1}^{L} u_\ell dA_\ell : -\rho \leq u_\ell \leq \rho,\ \ell = 1, \ldots, L \right\},$$

an important component of the above question is, What is the supremum $\rho^\star$ of those uncertainty levels ρ for which all instances $A \in \mathcal{V}_\rho$ remain stable, moreover, such that

$$z^T(t)\mathbf{Z}z(t) \le \beta \exp\{-\alpha t\} z^T(0)\mathbf{Z}z(0) \qquad \forall t \ge 0 \tag{12}$$

for all trajectories $z(\cdot)$ of all perturbed instances? Here $\mathbf{Z} \succ 0$, $\beta > 1$, and $\alpha > 0$ are given in advance. A well-known sufficient condition for (12) *is the existence of an appropriate quadratic Lyapunov stability certificate, namely, a matrix Z satisfying the relations*

$$\beta^{-1}\mathbf{Z} \preceq Z \preceq \mathbf{Z}, \tag{13a}$$

$$A^T Z + ZA \preceq -\alpha \mathbf{Z} \qquad \forall A \in \mathcal{V}_\rho. \tag{13b}$$

Indeed, if Z satisfies (13) *and $z(t)$ is a trajectory of the time-varying system*

$$\dot{z}(t) = A(t)z(t) \qquad (A(t) \in \mathcal{V}_\rho \qquad \forall t),$$

then

$$\begin{aligned} \tfrac{d}{dt}(z^T(t)Zz(t)) &= z^T(t)[A^T(t)Z + ZA(t)]z(t) \\ &\le -\alpha z^T(t)\mathbf{Z}z(t) && \text{(cf. (13b))} \\ &\le -\alpha z^T(t)Zz(t) && \text{(cf. (13a));} \end{aligned}$$

hence

$$\begin{aligned} z^T(t)Zz(t) &\le \exp\{-\alpha t\} z^T(0)Zz(0) \\ &\le \exp\{-\alpha t\} z^T(0)\mathbf{Z}z(0) \quad \text{(cf. (13a))} \end{aligned}$$

and therefore

$$\begin{aligned} z^T(t)\mathbf{Z}z(t) &\le \beta z^T(t)Zz(t) && \text{(cf. (13a))} \\ &\le \beta \exp\{-\alpha t\} z^T(0)\mathbf{Z}z(0). \end{aligned}$$

On the other hand, it is immediately seen that relations (13) *say exactly that Z is a common dissipativity certificate, belonging to the matrix interval $\mathcal{I} = \{Z : \beta^{-1}\mathbf{Z} \preceq Z \preceq \mathbf{Z}\}$ for all instances $\Sigma \in \mathcal{U}_\rho$ of the system*

$$\begin{aligned} \dot{z} &= Az + 0_{n\times 1} \cdot u, \\ y &= z + 0_{n\times 1} \cdot u \end{aligned} \tag{14}$$

when the supply matrix is specified as

$$\mathfrak{P} = \begin{bmatrix} -\alpha\mathbf{Z} & \\ & I \end{bmatrix}; \tag{15}$$

here $\mathcal{U}_\rho$ is the box uncertainty given by

$$d\Sigma_\ell = \begin{bmatrix} dA_\ell & 0_{n\times 1} \\ 0_{n\times n} & 0_{n\times 1} \end{bmatrix}, \qquad \ell = 1, \dots, L.$$

We see that Problem 1 *can be used to find the largest uncertainty level ρ for which the validity of* (13) *can be guaranteed by a quadratic Lyapunov stability certificate.*

3.2. Extracting available storage. Assume that we are interested in retrieving the energy stored in the initial state ζ. If there were no perturbations, the maximal amount of energy we could retrieve would be the nominal available storage $\zeta^T \mathbf{Z}_{av} \zeta$, and the corresponding control could be chosen in the state feedback form (see D.4). With perturbations, we hardly could guarantee the same amount of retrieved energy; however, it is reasonable to look for a state feedback which stabilizes all instances of the uncertain system in question and allows us to retrieve, whatever is an instance and an initial state ζ, at least a given fraction $(1-\epsilon)\zeta^T \mathbf{Z}_{av} \zeta$ of the nominal available storage. To model this target mathematically, we start with the following simple observation.

PROPOSITION 3.2. *Assume that $0 \prec \mathbf{Z}_{av}$, and let $\epsilon \in [0,1)$, $\rho \geq 0$ be given. Assume that matrices $G, H \in \mathbf{S}^n_+$ and a state feedback $u = Fz$ are such that*

1.

(16)
$$\begin{bmatrix} C^TQC & C^T(L+QD) \\ (L+QD)^TC & D^TQD + L^TD + D^TL + R \end{bmatrix} - \begin{bmatrix} A^TG + GA & GB \\ B^TG & \end{bmatrix} \succeq 0 \quad \forall \begin{bmatrix} A & B \\ C & D \end{bmatrix} \in \mathcal{U}_\rho,$$

i.e., G is a common dissipativity certificate for all instances of $\mathcal{U}_\rho$;

2.

(17)
$$\begin{bmatrix} I & F^T \end{bmatrix} \begin{bmatrix} C^TQC & C^T(L+QD) \\ (L+QD)^TC & D^TQD + L^TD + D^TL + R \end{bmatrix} \begin{bmatrix} I \\ F \end{bmatrix} \prec \left[(A+BF)^TH + H(A+BF)\right] \quad \forall \begin{bmatrix} A & B \\ C & D \end{bmatrix} \in \mathcal{U}_\rho;$$

3.

(18)
$$(1-\epsilon)\mathbf{Z}_{av} \preceq H \prec G.$$

Then all instances of the uncertain time-varying closed-loop system

(19)
$$\begin{array}{rcl} \dot{z}(t) & = & A(t)z(t) + B(t)u(t), \\ y(t) & = & C(t)z(t) + D(t)u(t), \\ u(t) & = & Fz(t), \end{array} \qquad \begin{bmatrix} A(t) & B(t) \\ C(t) & D(t) \end{bmatrix} \in \mathcal{U}_\rho \qquad \forall t$$

share a common quadratic Lyapunov function $z^T(G-H)z$. Moreover, for every initial state $\zeta = z(0)$ of (19), one has

(20)
$$-\int_0^\infty \mathfrak{S}(y(t),u(t))dt \geq \zeta^T H \zeta \geq (1-\epsilon)\zeta^T \mathbf{Z}_{av} \zeta,$$

i.e., the state feedback F allows us to extract at least $(1-\epsilon)$ times the nominal available storage $\zeta^T \mathbf{Z}_{av} \zeta$.

Proof. Consider a time-invariant instance $\Sigma = \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$ of (19), and let $(z(t), y(t), u(t))$ be a trajectory of this instance. By (16), the quadratic function $V(z) = z^TGz$ is a storage function for $(\Sigma, \mathfrak{S})$; hence for every $t_0 \leq t_1$

(21)
$$z^T(t_0)Gz(t_0) + \int_{t_0}^{t_1} \mathfrak{S}(y(t),u(t))dt \geq z^T(t_1)Gz(t_1).$$

On the other hand, (17) implies that

$$\mathfrak{S}(y(t),u(t)) \le \frac{d}{dt}\left(z^T(t)Hz(t)\right) - \theta z^T(t)z(t)$$

for certain $\theta > 0$; hence

$$\int_{t_0}^{t_1} \mathfrak{S}(y(t),u(t))dt \le \left[z^T(t_1)^T Hz(t_1) - z^T(t_0)^T Hz(t_0)\right] - \theta \int_{t_0}^{t_1} z^T(t)z(t)dt.$$

Substituting this inequality into (21), we see that for every trajectory of every time-invariant instance of (19) and every pair $t_0 \le t_1$ of time instants one has

$$z^T(t_0)[G-H]z(t_0) - \theta \int_{t_0}^{t_1} z^T(t)z(t)dt \ge z^T(t_1)[G-H]z(t_1);$$

hence $\frac{d}{dt}(z^T(t)[G-H]z(t)) \le -\theta z^T(t)z(t)$ for all $t \ge 0$ and all trajectories, so that

$$(A+BF)^T[G-H] + [G-H](A+BF) \preceq -\theta I.$$

Since this relation is valid for all $\Sigma \in \mathcal{U}_\rho$, and since $G - H \succ 0$ by (18), $G - H$ is indeed a quadratic Lyapunov stability certificate for (19).

Now consider a trajectory $(z(t), y(t), u(t))$ of (19). Same as above, we have

$$\mathfrak{S}(y(t),u(t)) \le \frac{d}{dt}\left(z^T(t)Hz(t)\right).$$

Integrating both sides of this inequality from 0 to ∞ and taking into account that $(z(t), y(t), u(t))$ converges exponentially fast to 0 as $t \to \infty$ (we have seen that (19) admits quadratic Lyapunov stability certificate!), we get

$$\int_0^\infty \mathfrak{S}(y(t),u(t))dt \le -z^T(0)Hz(0),$$

as required in the first inequality in (20); the second inequality in (20) is readily given by (18). □

In view of Proposition 3.2, we could pose the problem of extracting available storage as the problem of finding the supremum of those uncertainty levels ρ for which the semi-infinite system of MIs (16), (17), (18) in matrix variables G, H, F is solvable. This problem, however, is too difficult; it is completely unclear how to check efficiently the solvability of this *nonlinear* in F, H MI even in the nominal case $\rho = 0$. This is why we are forced to simplify our task by assuming that either F or H are given in advance. With this simplification, we arrive at the following pair of problems.

PROBLEM 2A. *Given a supply $\mathfrak{S}$, a feedback matrix F, parameter $\epsilon \in (0,1)$, and the data specifying $\mathcal{U}_\rho$, find the supremum of those $\rho \ge 0$ for which the system of MIs* (16)–(18) *in matrix variables G, H is solvable.*

With F specified as the ideal nominal feedback F_{av}, see D.4, Problem 2A becomes a quite natural question of finding the largest uncertainty level for which we can certify the fact that whatever is an initial state ζ of an instance of the uncertain system, the nominal feedback allows us to extract at least the fraction $(1-\epsilon)$ of the corresponding nominal available storage $\zeta^T \mathbf{Z}_{av}\zeta$.

PROBLEM 2B. *Given a supply $\mathfrak{S}$, parameter $\epsilon \in (0,1)$, an $n \times n$ positive definite matrix $H \succeq (1-\epsilon)\mathbf{Z}_{av}$, and the data specifying $\mathcal{U}_\rho$, find the supremum of those $\rho \ge 0$ for which the system of MIs* (16), (18) *in matrix variables G and F is solvable.*

A simple choice for the matrix H in Problem 2B is the solution of Problem 2A.

3.3. Providing required supply. The motivation behind this problem is completely similar to the one for the extracting storage problem; the only difference is that now we want to drive the system from the origin to a given state ζ and we are interested in achieving this target with the total supply not exceeding $(1+\delta)$ times the nominal required supply $\zeta^T\mathbf{Z}_{req}\zeta$. We have the following analogy of Proposition 3.2.

PROPOSITION 3.3. *Assume that $0 \prec \mathbf{Z}_{req}$, and let $\delta \in [0,1)$, $\rho \geq 0$ be given. Assume that matrices $G, H \in \mathbf{S}^n_+$ and a state feedback $u = Fz$ are such that the conditions* (16), (17) *and the condition*

$$G \prec H \preceq (1+\delta)\mathbf{Z}_{req} \tag{22}$$

are satisfied.

Then all instances of the uncertain time-varying closed-loop system

$$\begin{array}{l} \dot{z}(t) = -[A(t)z(t) + B(t)u(t)], \\ y(t) = C(t)z(t) + D(t)u(t), \\ u(t) = Fz(t), \end{array} \qquad \begin{bmatrix} A(t) & B(t) \\ C(t) & D(t) \end{bmatrix} \in \mathcal{U}_\rho \qquad \forall t \tag{23}$$

(which is the backward time version of system (19)*) share a common quadratic Lyapunov function $z^T[H-G]z$. Moreover, for every initial state $\zeta = z(0)$ of* (23)*, one has*

$$\int_0^\infty \mathfrak{S}(y(t),u(t))dt \leq (1+\delta)\zeta^T\mathbf{Z}_{req}\zeta, \tag{24}$$

i.e., the state feedback F allows us to move system (19) *from the origin to a given state ζ with total supply at most $(1+\delta)$ times the nominal required supply $\zeta^T\mathbf{Z}_{req}\zeta$.*

The proof is similar to the one of Proposition 3.2.

In view of Proposition 3.3, a natural way to model the providing required supply problem would be to look for the largest ρ for which the semi-infinite system (16), (17), (22) in matrix variables F, G, H is solvable; however, "tractability reasons" similar to those in section 3.2 force us to simplify the setting and restrict ourselves to the following pair of problems.

PROBLEM 3A. *Given a supply $\mathfrak{S}$, a feedback matrix F, parameter $\delta > 0$, and the data specifying $\mathcal{U}_\rho$, find the supremum of those $\rho \geq 0$ for which the system of MIs* (16), (17), (22) *in matrix variables G, H is solvable.*

PROBLEM 3B. *Given a supply $\mathfrak{S}$, parameter $\delta > 0$, an $n \times n$ positive definite matrix $H \preceq (1+\delta)\mathbf{Z}_{req}$, and the data specifying $\mathcal{U}_\rho$, find the supremum of those $\rho \geq 0$ for which the system of MIs* (16), (17) *in matrix variables G, $0 \preceq G \prec H$, and F is solvable.*

In contrast to the situation of section 3.2, now there exists a particular "tractable case" where one can treat in the system of interest (which is now the system (16), (17), (22)) both F and H as design variables; this is the case of positive semidefinite supply matrix $\left[\begin{smallmatrix} Q & L \\ L^T & R \end{smallmatrix}\right]$ (as it happens in linear-quadratic control).[4] In this case it makes sense to specify the common dissipativity certificate G of the perturbed instances as the zero matrix; this choice ensures the validity of (16) and is "ideal" from the viewpoint of the constraint (22). Setting $G = 0$ and treating F, H as the design variables in the

[4]Note that this case makes no sense in the extracting storage problem, since there it would imply that "there is nothing to extract" – $\mathbf{Z}_{av} = 0$

system (16), (17), (22), we arrive at the following version of the problem of providing required supply.

PROBLEM 3C. *Given a supply* $\mathfrak{S}$ *such that the supply matrix* $\mathfrak{P}$ *is positive semidefinite, parameter* $\delta > 0$*, and the data specifying* $\mathcal{U}_\rho$*, find the supremum of those* $\rho \geq 0$ *for which the system comprised of semi-infinite MI* (17) *and the LMI*

$$0 \prec H \preceq (1+\delta)\mathbf{Z}_{req} \tag{25}$$

in matrix variables F, H *is solvable.*

4. Processing the problems. Every one of Problems 1, 2A, 2B, 3A, 3B, and 3C asks for finding the largest ρ such that a given system of MIs (depending on ρ as on a parameter) is solvable. The systems in question are *semi-infinite*—they involve infinitely many MIs with the data running through the uncertainty sets. It is well known that semi-infinite systems of MIs are, in general, NP-hard; it is easy to show that in general this is the case with the specific semi-infinite systems arising in Problems 1, 2A, 2B, 3A, 3B, and 3C. What we intend to do is to replace these NP-hard systems with their computationally tractable *conservative approximations*, the latter notion being defined as follows.

DEFINITION 4.1. *Let* $\mathcal{S}$ *be a system of constraints on a design vector* x*. We say that a system* $\mathcal{A}$ *of constraints on* x *and a vector of additional variables* y *is a conservative approximation of* $\mathcal{S}$ *if the* x*-component of every feasible solution* (x, y) *of the approximating system* $\mathcal{A}$ *is a feasible solution of the original system* $\mathcal{S}$*.*

Our plan for processing Problems 1, 2A, 2B, 3A, 3B, and 3C is as follows: we start with reviewing the basic results we intend to use when building computationally tractable approximations of the problems and then apply these results to the problems of interest.

4.1. The matrix cube theorem. Consider an uncertain LMI with affine box uncertainty

$$\mathcal{A}^0(x) + \sum_{\ell=1}^{L} u_\ell \mathcal{A}^\ell(x) \succeq 0 \qquad \forall (u : \|u\|_\infty \leq \rho), \tag{26}$$

where

- $x \in \mathbf{R}^d$ is the vector of decision variables;
- $\mathcal{A}^\ell(x)$, $\ell = 0, 1, \ldots, L$, are symmetric $m \times m$ matrices affinely depending on x;
- $u_1, \ldots, u_L$ are *perturbations*, and $\rho \geq 0$ is the *uncertainty level*.

It is known that in general, it is NP-hard to solve (26) or even to check whether a given candidate solution x is feasible. However, (26) admits a computationally tractable conservative approximation which is a system of LMIs in original variables x and additional symmetric matrix variables $X_1, \ldots, X_L$. Let us write $X \succeq \pm Y$ as a shortcut for the system of two matrix inequalities $X \succeq Y$, $X \succeq -Y$. The aforementioned conservative approximation of (26) is as follows:

$$X_\ell \succeq \pm \mathcal{A}^\ell(x), \qquad \ell = 1, \ldots, L; \tag{27a}$$

$$\rho \sum_{\ell=1}^{L} X_\ell \preceq \mathcal{A}^0(x). \tag{27b}$$

The fact that (27) is indeed a conservative approximation of (26) is evident: if x can be extended by appropriately chosen $X_1, \dots, X_L$ to a feasible solution of (27), then from (27a) it follows that $u_\ell \mathcal{A}^\ell(x) \succeq -\rho X_\ell$ for all u_ℓ such that $|u_\ell| \le \rho$; hence

$$\mathcal{A}^0(x) + \sum_{\ell=1}^{L} u_\ell \mathcal{A}^\ell(x) \succeq \mathcal{A}^0(x) - \rho \sum_{\ell=1}^{L} X_\ell \qquad \forall (u : \|u\|_\infty \le \rho);$$

the right-hand side matrix in the latter relation is $\succeq 0$ by (27b), so that x indeed satisfies (26).

It turns out that the "level of conservativeness" of the approximation (27) is not too big, provided that the matrices $\mathcal{A}^1(x), \dots, \mathcal{A}^L(x)$ are of small ranks.

PROPOSITION 4.1 (matrix cube theorem [3]). *Let* $\mu = \max_x \max_{\ell \ge 1} \operatorname{Rank}(\mathcal{A}^\ell(x))$. *(Note* $\ell \ge 1$ *in the* max*!). Then the relation between the feasible sets of* (26) *and* (27) *is as follows:*

1. *If* x *can be extended to a feasible solution of* (27)*, then* x *is feasible for* (26).
2. *If* x *cannot be extended to a feasible solution of* (27)*, then* x *is infeasible for* (26) *with* ρ *replaced by* $\vartheta(\mu)\rho$*, where* $\vartheta(\cdot)$ *is certain universal function such that* $\vartheta(\mu) \le \frac{\pi\sqrt{\mu}}{2}$ *for all* μ *and*

$$\vartheta(1) = 1, \quad \vartheta(2) = \frac{\pi}{2} = 1.57\dots, \quad \vartheta(3) = 1.73\dots, \quad \vartheta(4) = 2.$$

In particular, for every set $\mathcal{X} \subset \mathbf{R}^d$ *one has*

$$1 \le \frac{\sup\{\rho : (26) \text{ has a solution in } \mathcal{X}\}}{\sup\{\rho : (27) \text{ has a solution in } \mathcal{X}\}} \le \vartheta(\mu)$$

provided that the numerator in the fraction is positive.

Remark 1. Sometimes we shall be interested in a sufficient condition for the strict version

$$\mathcal{A}^0(x) + \sum_{\ell=1}^{L} u_\ell \mathcal{A}^\ell(x) \succ 0 \qquad \forall (u : \|u\|_\infty \le \rho)$$

of the semi-infinite LMI (26). Such a sufficient condition can be obtained from (27) by replacing the nonstrict LMI (27b) with its strict version. For the resulting pair of conditions, a statement completely similar to the matrix cube theorem takes place.

4.2. Approximating Problem 1. Let

$$Q = Q_+ - Q_-$$

be the representation of Q as a difference of two positive semidefinite symmetric matrices with orthogonal image spaces, and let

$$S_+ = Q_+^{1/2}, \qquad S_- = Q_-^{1/2}.$$

From now on, we assume that the set $\mathcal{I}$ in Problem 1 is *LMI-representable*, i.e., it can be specified by LMI $\{Z : \mathcal{Z}[Z] \succeq 0\}$, where $\mathcal{Z}[\cdot]$ is an affine function taking values in the space of symmetric matrices. With this assumption, Problem 1 becomes the problem of finding the supremum $\rho_1^\star$ of those $\rho > 0$ for which the system of LMIs

$$\mathcal{Z}[Z] \succeq 0, \tag{28a}$$

$$\begin{bmatrix} -A^T Z - ZA & C^T L - ZB \\ L^T C - B^T Z & L^T D + D^T L + R \end{bmatrix} + \begin{bmatrix} C^T \\ D^T \end{bmatrix} Q \begin{bmatrix} C & D \end{bmatrix} \succeq 0 \tag{28b}$$

$$\forall (A, B, C, D) \in \mathcal{U}_\rho$$

in symmetric matrix variable Z has a solution. This system can be equivalently rewritten as

(29a)
$$\mathcal{Z}[Z] \succeq 0,$$

$$\left[\begin{array}{c|c} \delta C^T Q\mathbf{C} + \mathbf{C}^T Q\delta C + \mathbf{C}^T Q\mathbf{C} & \delta C^T Q\mathbf{D} + \mathbf{C}^T Q\delta D + \mathbf{C}^T Q\mathbf{D} \\ -A^T Z - ZA & +C^T L - ZB \\ \hline \mathbf{D}^T Q\delta C + \delta D^T Q\mathbf{C} + \mathbf{D}^T Q\mathbf{C} & \delta D^T Q\mathbf{D} + \mathbf{D}^T Q\delta D + \mathbf{D}^T Q\mathbf{D} \\ +L^T C - B^T Z & +L^T D + D^T L + R \end{array}\right]$$

(29b)
$$-\left[\begin{array}{c} \delta C^T S_- \\ \delta D^T S_- \end{array}\right]\left[\begin{array}{cc} S_-\delta C & S_-\delta D \end{array}\right] + \left[\begin{array}{c} \delta C^T \\ \delta D^T \end{array}\right] Q_+ \left[\begin{array}{cc} \delta C & \delta D \end{array}\right] \succeq 0$$

$$\forall \left[\begin{array}{cc} A = \mathbf{A} + \delta A & B = \mathbf{B} + \delta B \\ C = \mathbf{C} + \delta C & D = \mathbf{D} + \delta D \end{array}\right] \in \mathcal{U}_\rho.$$

Since $Q_+ \succeq 0$, the last term in the left-hand side of (29b) is positive semidefinite. Eliminating this term, we pass from (29) to a conservative approximation of this system. By the Schur complement lemma,[5] this approximation is equivalent to the system of LMIs

(30a)
$$\mathcal{Z}[Z] \succeq 0,$$

(30b)
$$\left[\begin{array}{c|c|c} \delta C^T Q\mathbf{C} + \mathbf{C}^T Q\delta C + \mathbf{C}^T Q\mathbf{C} & \delta C^T Q\mathbf{D} + \mathbf{C}^T Q\delta D + \mathbf{C}^T Q\mathbf{D} & \delta C^T S_- \\ -A^T Z - ZA & +C^T L - ZB & \\ \hline \mathbf{D}^T Q\delta C + \delta D^T Q\mathbf{C} + \mathbf{D}^T Q\mathbf{C} & \delta D^T Q\mathbf{D} + \mathbf{D}^T Q\delta D & \\ +L^T C - B^T Z & +\mathbf{D}^T Q\mathbf{D} & \delta D^T S_- \\ & +L^T D + D^T L + R & \\ \hline S_-\delta C & S_-\delta D & I_p \end{array}\right] \succeq 0$$

$$\forall \left[\begin{array}{cc} A = \mathbf{A} + \delta A & B = \mathbf{B} + \delta B \\ C = \mathbf{C} + \delta C & D = \mathbf{D} + \delta D \end{array}\right] \in \mathcal{U}_\rho.$$

Taking into account (8), we see that the latter semi-infinite system of LMIs is in the form of (26), and we can use the construction from section 4.1 to build a computationally tractable conservative approximation of this system (and thus of (28)). The approximation is the following system of LMIs in matrix variables $Z, \{X_\ell\}$:

(31)
$$\mathcal{Z}[Z] \succeq 0,$$

$$X_\ell \succeq \pm \overbrace{\left[\begin{array}{c|c|c} dC_\ell^T Q\mathbf{C} + \mathbf{C}^T Q dC_\ell & dC_\ell^T [L + Q\mathbf{D}] & dC_\ell^T S_- \\ -dA_\ell^T Z - Z dA_\ell & +\mathbf{C}^T Q dD_\ell - Z dB_\ell & \\ \hline [L + Q\mathbf{D}]^T dC_\ell & dD_\ell^T [L + Q\mathbf{D}] & dD_\ell^T S_- \\ +dD_\ell^T Q\mathbf{C} - dB_\ell^T Z & +[L + Q\mathbf{D}]^T dD_\ell & \\ \hline S_- dC_\ell & S_- dD_\ell & 0_{pp} \end{array}\right]}^{\mathcal{A}^\ell[Z]}, \quad \ell = 1, \ldots, L,$$

$$\rho \sum_{\ell=1}^{L} X_\ell \preceq \left[\begin{array}{c|c|c} \mathbf{C}^T Q\mathbf{C} - \mathbf{A}^T Z - Z\mathbf{A} & \mathbf{C}^T [L + Q\mathbf{D}] - Z\mathbf{B} & \\ \hline [L + Q\mathbf{D}]^T \mathbf{C} - \mathbf{B}^T Z & L^T \mathbf{D} + \mathbf{D}^T L & \\ & +\mathbf{D}^T Q\mathbf{D} + R & \\ \hline & & I_p \end{array}\right].$$

[5] The Schur complement lemma (see, e.g., [2, Chapter 4]) states that a symmetric block matrix $\left[\begin{smallmatrix} P & L \\ L^T & Q \end{smallmatrix}\right]$ with $Q \succ 0$ is positive definite (positive semidefinite) if and only if the matrix $P - LQ^{-1}L^T$ is positive definite (positive semidefinite).

Note that the supremum $\widehat{\rho}_1$ of those $\rho \geq 0$ for which system (31) is solvable is efficiently computable—it is the optimal value in the problem

$$\max_{\rho,\{X_\ell\},Z} \{\rho : (\rho, \{X_\ell\}, Z) \text{ solves (31)}\}.$$

The latter is a generalized eigenvalue problem, so that its optimal value is efficiently computable. We intend to use the efficiently computable quantity $\widehat{\rho}$ as a bound for the "quantity of interest" $\rho_1^\star$. The properties of this bound are described in the following statement.

PROPOSITION 4.2. (i) *System* (31) *is a conservative approximation of* (28)*, so that the Z-component of a feasible solution to* (31) *is a feasible solution of* (28)*. In particular, $\widehat{\rho}_1$ is a lower bound for $\rho_1^\star$.*

(ii) *If either*

(a) *$Q \preceq 0$ (i.e., $Q_+ = 0$) (as it is the case, e.g., in Examples* 1, 2, 4*)*

or

(b) *D and C are certain (i.e., $dC_\ell = 0$, $dD_\ell = 0$ for all ℓ),*

then

$$1 \leq \frac{\rho_1^\star}{\widehat{\rho}_1} \leq \vartheta(\mu) \tag{32}$$

provided that $\rho_1^\star > 0$. Here $\vartheta(\mu)$ is the function from Proposition 4.1 *and*

$$\mu = \max_{\ell=1,\dots,L} \max_{Z} \operatorname{Rank}(\mathcal{A}^\ell[Z]);$$

see (31)*.*

Proof. The validity of the first claim is readily given by the origin of (31). To justify the second claim, note that in the case of $Q \preceq 0$, same as in the case when C, D are certain, system (30) is solvable if and only if (28) is solvable, so that $\rho_1^\star$ is the supremum of those $\rho \geq 0$ for which (30) is solvable; with this observation, (32) is readily given by Proposition 4.1. □

Unfortunately, we cannot bound from above fraction (32) in the case of uncertain C, D and $Q_+ \neq 0$, since here the derivation of the approximating system includes a step (passing from (28) to (30)) with an unknown "level of conservativeness."

4.3. Approximating Problems 2A and 3A. It suffices to process Problem 2A, since Problem 3A can be treated in a completely similar fashion. The semi-infinite LMI (16), similar to the semi-infinite LMI (28), admits the conservative approximation (cf. (30))

(33)

$$\left[\begin{array}{c|c|c} \delta C^T Q\mathbf{C} + \mathbf{C}^T Q\delta C + \mathbf{C}^T Q\mathbf{C} \\ -A^T G - GA & \delta C^T Q\mathbf{D} + \mathbf{C}^T Q\delta D + \mathbf{C}^T Q\mathbf{D} \\ & +C^T L - GB & \delta C^T S_- \\ \hline \mathbf{D}^T Q\delta C + \delta D^T Q\mathbf{C} + \mathbf{D}^T Q\mathbf{C} \\ +L^T C - B^T G & \delta D^T Q\mathbf{D} + \mathbf{D}^T Q\delta D \\ & +\mathbf{D}^T Q\mathbf{D} \\ & +L^T D + D^T L + R & \delta D^T S_- \\ \hline S_-\delta C & S_-\delta D & I_p \end{array}\right] \succeq 0$$

$$\forall \left[\begin{array}{cc} A = \mathbf{A} + \delta A & B = \mathbf{B} + \delta B \\ C = \mathbf{C} + \delta C & D = \mathbf{D} + \delta D \end{array}\right] \in \mathcal{U}_\rho,$$

which is equivalent to (16) in the case of $Q_+ = 0$, as well as in the case of certain C, D. The semi-infinite LMI (17) can be rewritten as

(34)
$$\begin{bmatrix} I & F^T \end{bmatrix} \begin{bmatrix} \mathbf{C}^TQ\mathbf{C} + \delta C^TQ\mathbf{C} + \mathbf{C}^TQ\delta C & \mathbf{C}^T(L+Q\mathbf{D}) + \delta C^T(L+Q\mathbf{D}) \\ (L+QD)^T\mathbf{C} + (L+Q\mathbf{D})^T\delta C & \delta D^T(L+Q\mathbf{D}) + (L+Q\mathbf{D})^T\delta D + \mathbf{D}^TQ\mathbf{D} + \mathbf{D}^TL + L^T\mathbf{D} + R \end{bmatrix} \begin{bmatrix} I \\ F \end{bmatrix}$$
$$+ (\delta C + \delta DF)^T S_+^2 (\delta C + \delta DF)$$
$$- (\delta C + \delta DF)^T S_-^2 (\delta C + \delta DF) \prec \big[(A+BF)^TH + H(A+BF)\big]$$
$$\forall \begin{bmatrix} A = \mathbf{A} + \delta A & B = \mathbf{B} + \delta B \\ C = \mathbf{C} + \delta C & D = \mathbf{D} + \delta D \end{bmatrix} \in \mathcal{U}_\rho.$$

The third term in the left-hand side of this MI is negative semidefinite; eliminating this term, we get a conservative approximation of (34), and this approximation, by the Schur complement lemma, is equivalent to the following semi-infinite LMI, where we set

$$\mathcal{F} = \left[\begin{array}{c|c} I_n & \\ F & \\ \hline & I_p \end{array}\right]:$$

(35)
$$\mathcal{F}^T \left[\begin{array}{cc|c} A^TH + HA - \mathbf{C}^TQ\mathbf{C} & HB - \mathbf{C}^T(L+QD) & \delta C^TS_+ \\ -\delta C^TQ\mathbf{C} - \mathbf{C}^TQ\delta C & -\delta C^T(L+Q\mathbf{D}) & \\ \hline B^TH - (L+QD)^T\mathbf{C} & -\delta D^T(L+Q\mathbf{D}) - (L+Q\mathbf{D})^T\delta D & \delta D^TS_+ \\ -(L+Q\mathbf{D})^T\delta C & -\mathbf{D}^TL + L^T\mathbf{D} - \mathbf{D}^TQ\mathbf{D} - R & \\ \hline S_+\delta C & S_+\delta D & I_p \end{array}\right] \mathcal{F} \succ 0$$
$$\forall \begin{bmatrix} A = \mathbf{A} + \delta A & B = \mathbf{B} + \delta B \\ C = \mathbf{C} + \delta C & D = \mathbf{D} + \delta D \end{bmatrix} \in \mathcal{U}_\rho$$

in matrix variable H. Thus, the system of semi-infinite LMIs (33), (35) in matrix variables G, H is a conservative approximation of (16), (17); in the cases when $Q = 0$ and/or C, D are certain, the former system in fact is equivalent to the latter one. The semi-infinite system (33), (35) is in the form of (26). Applying the construction from section 4.1, we end up with computationally tractable conservative approximation of the system (16), (17), (18). The approximation is the following system of LMIs in matrix variables $G, H, \{X_\ell, Y_\ell\}$:

(36a)
$$X_\ell \succeq \pm \overbrace{\left[\begin{array}{cc|c} dC_\ell^TQ\mathbf{C} + \mathbf{C}^TQdC_\ell - dA_\ell^TG - GdA_\ell & dC_\ell^T(L+Q\mathbf{D}) + \mathbf{C}^TQdD_\ell - GdB_\ell & dC_\ell^TS_- \\ \hline (L+Q\mathbf{D})^TdC_\ell + dD_\ell^TQ\mathbf{C} - dB_\ell^TG & dD_\ell^T(L+Q\mathbf{D}) + (L+Q\mathbf{D})^TdD_\ell & dD_\ell^TS_- \\ \hline S_-dC_\ell & S_-dD_\ell & \end{array}\right]}^{B^\ell[G]}, \quad \ell = 1, \dots, L,$$

(36b)
$$\rho \sum_{\ell=1}^{L} X_\ell \preceq \left[\begin{array}{cc|c} \mathbf{C}^TQ\mathbf{C} - \mathbf{A}^TG - G\mathbf{A} & \mathbf{C}^T(L+Q\mathbf{D}) - G\mathbf{B} & \\ \hline (L+Q\mathbf{D})^T\mathbf{C} - \mathbf{B}^TG & \mathbf{D}^TQ\mathbf{D} + L^T\mathbf{D} + \mathbf{D}^TL + R & \\ \hline & & I_p \end{array}\right],$$

(36c)

$$Y_\ell \succeq \pm \mathcal{F}^T \overbrace{\left[\begin{array}{c|c|c} dA_\ell^T H + H dA_\ell & H dB_\ell & dC_\ell^T S_+ \\ -dC_\ell^T Q\mathbf{C} - \mathbf{C}^T Q dC_\ell & -dC_\ell^T(L+Q\mathbf{D}) - \mathbf{C}^T Q dD_\ell & \\ \hline dB_\ell^T H & -dD_\ell^T(L+Q\mathbf{D}) & dD_\ell^T S_+ \\ -(L+Q\mathbf{D})^T dC_\ell - dD_\ell^T Q\mathbf{C} & -(L+Q\mathbf{D})^T dD_\ell & \\ \hline S_+ dC_\ell & S_+ dD_\ell & \end{array}\right]}^{\mathcal{C}^\ell[H]} \mathcal{F},$$

(36d)

$$\rho \sum_{\ell=1}^{L} Y_\ell \prec \mathcal{F}^T \left[\begin{array}{c|c|c} \mathbf{A}^T H + H\mathbf{A} - \mathbf{C}^T Q\mathbf{C} & H\mathbf{B} - \mathbf{C}^T(L+Q\mathbf{D}) & \\ \hline \mathbf{B}^T H - (L+Q\mathbf{D})^T\mathbf{C} & -\mathbf{D}^T Q\mathbf{D} - L^T\mathbf{D} - \mathbf{D}^T L - R & \\ \hline & & I_p \end{array}\right] \mathcal{F},$$

(36e)

$$(1-\epsilon)\mathbf{Z}_{av} \preceq H \prec G.$$

The supremum $\widehat{\rho}_{2A}$ of those $\rho \geq 0$ for which system (36) is solvable is efficiently computable, and this efficiently computable quantity can be used as a bound for the optimal value $\rho^\star_{2\mathrm{A}}$ in Problem 2A. The properties of this bound are described in the following.

PROPOSITION 4.3. (i) *System* (36) *is a conservative approximation of* (16), (17), (18) *so that the* G, H*-components of a feasible solution to* (36) *are a feasible solution of* (16), (17), (18). *In particular,* $\widehat{\rho}_{2A}$ *is a lower bound for* $\rho^\star_{2\mathrm{A}}$.

(ii) *If either*

(a) $Q = 0$ *(i.e.,* $Q_+ = Q_- = 0$*),*

or

(b) D *and* C *are certain (i.e.,* $dC_\ell = 0$, $dD_\ell = 0$ *for all* ℓ*),*

then

$$1 \leq \frac{\rho^\star_{2\mathrm{A}}}{\widehat{\rho}_{2A}} \leq \vartheta(\mu) \tag{37}$$

provided that $\rho^\star_{2\mathrm{A}} > 0$. *Here* $\vartheta(\mu)$ *is the function from Proposition* 4.1 *and*

$$\mu = \max\left[\max_{\ell\geq 1, G} \mathrm{Rank}(\mathcal{B}^\ell[G]), \max_{\ell\geq 1, H} \mathrm{Rank}(\mathcal{C}^\ell[H])\right];$$

see (36).

Tractable conservative approximation of Problem 3A looks exactly as (36), up to the constraint (36e), which should be replaced with the constraint

$$0 \preceq G \prec H \preceq (1+\delta)\mathbf{Z}_{req}.$$

The properties of this approximation are completely similar to those established in Proposition 4.3.

4.4. Approximating Problems 2B and 3B. Our current goal is to build a tractable conservative approximation of the semi-infinite system of MIs associated with Problems 2B and 3B. Both problems have the same structure, so that it suffices to consider the system associated with Problem 2B, i.e., the system (16), (17), (22) in variables G, F (H now is fixed). We have already built a tractable conservative approximation of the semi-infinite MI (16); it is given by system of LMIs (36a), (36b) in matrix variables $G, \{X_\ell\}$. Let us focus on the semi-infinite MI (17) in variable F.

We can rewrite this inequality equivalently as

(38)
$$(S_+\delta C + S_+\delta DF)^T(S_+\delta C + S_+\delta DF) - (S_-\delta C + S_-\delta DF)^T(S_-\delta C + S_-\delta DF)$$
$$+\begin{bmatrix} I & F^T \end{bmatrix}\left[\begin{array}{c|c} \delta C^T Q\mathbf{C} + \mathbf{C}^T Q\delta C + \mathbf{C}^T Q\mathbf{C} & \mathbf{C}^T(L+Q\mathbf{D}) + \delta C^T(L+Q\mathbf{D}) + \mathbf{C}^T Q\delta D \\ \hline (L+Q\mathbf{D})^T\mathbf{C} + (L+Q\mathbf{D})^T\delta C + \delta D^T Q\mathbf{C} & \mathbf{D}^T Q\mathbf{D} + L^T\mathbf{D} + \mathbf{D}^T L + R + \delta D^T Q\mathbf{D} + \mathbf{D}^T Q\delta D + L^T\delta D + \delta D^T L \end{array}\right]\begin{bmatrix} I \\ F \end{bmatrix}$$
$$\prec \left([A+BF]^T H + H[A+BF]\right)$$
$$\forall \begin{bmatrix} A = \mathbf{A} + \delta A & B = \mathbf{B} + \delta B \\ C = \mathbf{C} + \delta C & D = \mathbf{D} + \delta D \end{bmatrix} \in \mathcal{U}_\rho.$$

The second term in the left-hand side of the latter MI always is negative semidefinite; eliminating this term, we come to a conservative approximation of (38) as follows:

(39)
$$(S_+\delta C + S_+\delta DF)^T(S_+\delta C + S_+\delta DF)$$
$$+\begin{bmatrix} I & F^T \end{bmatrix}\left[\begin{array}{c|c} \overbrace{\delta C^T Q\mathbf{C} + \mathbf{C}^T Q\delta C + \mathbf{C}^T Q\mathbf{C}}^{J_{00}[\Sigma]} & \overbrace{\mathbf{C}^T(L+Q\mathbf{D}) + \delta C^T(L+Q\mathbf{D}) + \mathbf{C}^T Q\delta D}^{J_{01}[\Sigma]} \\ \hline \underbrace{(L+Q\mathbf{D})^T\mathbf{C} + (L+Q\mathbf{D})^T\delta C + \delta D^T Q\mathbf{C}}_{J_{10}[\Sigma]} & \underbrace{\mathbf{D}^T Q\mathbf{D} + L^T\mathbf{D} + \mathbf{D}^T L + R + \delta D^T Q\mathbf{D} + \mathbf{D}^T Q\delta D + L^T\delta D + \delta D^T L}_{J_{11}[\Sigma]} \end{array}\right]\begin{bmatrix} I \\ F \end{bmatrix}$$
$$\prec \left([A+BF]^T H + H[A+BF]\right)$$
$$\forall\Sigma \equiv \begin{bmatrix} A = \mathbf{A} + \delta A & B = \mathbf{B} + \delta B \\ C = \mathbf{C} + \delta C & D = \mathbf{D} + \delta D \end{bmatrix} \in \mathcal{U}_\rho.$$

Note that the matrices $J_{ij}[\Sigma]$ are affine in Σ.

Observe that (39) is exactly the semi-infinite MI

$$\begin{aligned} &[A+BF]^T H + H[A+BF] - J_{00}[\Sigma] - F^T J_{10}[\Sigma] - J_{01}[\Sigma]F \\ &\quad -(S_+\delta C + S_+\delta DF)^T(S_+\delta C + S_+\delta DF) - F^T J_{11}[\Sigma]F \succ 0 \\ &\qquad \forall\Sigma \equiv \begin{bmatrix} A = \mathbf{A} + \delta A & B = \mathbf{B} + \delta B \\ C = \mathbf{C} + \delta C & D = \mathbf{D} + \delta D \end{bmatrix} \in \mathcal{U}_\rho. \end{aligned} \tag{40}$$

Now assume that $J_{11}[\Sigma] \succ 0$. Note that this assumption is quite natural—the matrix $J_{11}[\Sigma]$ should be positive semidefinite already to make feasible (16) with $\rho = 0$. Let

$$\mathbf{K} = J_{11}^{-1}[\Sigma], \qquad \delta J_{11}[\delta\Sigma] = J_{11}[\Sigma + \delta\Sigma] - J_{11}[\Sigma].$$

We claim that the following relations hold true:

$$\mathbf{K} - \mathbf{K}\delta J_{11}[\delta\Sigma]\mathbf{K} \succ 0 \qquad \forall\Sigma \equiv \Sigma + \delta\Sigma \in \mathcal{U}_\rho \tag{41a}$$

$$\Updownarrow$$

$$J_{11}[\Sigma] \succ 0 \qquad \forall\Sigma \in \mathcal{U}_\rho \tag{41b}$$

$$\Downarrow$$

$$[\mathbf{K} - \mathbf{K}\delta J_{11}[\delta\Sigma]\mathbf{K}]^{-1} \succeq [J_{11}[\Sigma]]^{-1} \succ 0 \qquad \forall \Sigma \equiv \Sigma + \delta\Sigma \in \mathcal{U}_\rho. \tag{41c}$$

Indeed, the equivalence between (41a) and (41b) follows from the identity

$$\mathbf{K} - \mathbf{K}\delta J_{11}[\delta\Sigma]\mathbf{K} = \mathbf{K}J_{11}[\Sigma - \delta\Sigma]\mathbf{K}$$

(which is readily given by the definition of $\mathbf{K}$), combined with the fact that $\mathcal{U}_\rho$ is symmetric with respect to Σ. To see that (41b) implies (41c), observe, first, that

$$X \succ \pm Y \Rightarrow [X^{-1} - X^{-1}YX^{-1}]^{-1} \succeq X + Y \succ 0. \tag{42}$$

Indeed, assuming $X \succ \pm Y$ and setting $Z = X^{-1/2}YX^{-1/2}$ (so that $I \succ \pm Z$), we have

$$\begin{aligned}(X+Y)^{-1} - [X^{-1} - X^{-1}YX^{-1}] &= [X^{1/2}(I+Z)X^{1/2}]^{-1} - X^{-1/2}[I-Z]X^{-1/2} \\ &= X^{-1/2}[(I+Z)^{-1} - (I-Z)]X^{-1/2} \\ &= X^{-1/2}Z(I+Z)^{-1}ZX^{-1/2} \succeq 0,\end{aligned}$$

hence $(X+Y)^{-1} \succeq [X^{-1} - X^{-1}YX^{-1}]$ and thus $[X^{-1} - X^{-1}YX^{-1}]^{-1} \succeq X + Y \succ 0$, as required in (42). Now let $\delta\Sigma$ be such that $\Sigma + \delta\Sigma \in \mathcal{U}_\rho$. Since $\mathcal{U}_\rho$ is symmetric with respect to Σ, we have $\Sigma - \delta\Sigma \in \mathcal{U}_\rho$ as well. In the case of (41b) it follows that $J_{11}[\Sigma \pm \delta\Sigma] \succ 0$ or, which is the same, $J_{11}[\Sigma] \succ \pm\delta J_{11}[\delta\Sigma]$. Applying (42), we arrive at (41c).

By (41), in the case of (41a) the semi-infinite MI

$$\begin{gathered}[A+BF]^T H + H[A+BF] - J_{00}[\Sigma] - F^T J_{10}[\Sigma] - J_{01}[\Sigma]F \\ - (S_+\delta C + S_+\delta DF)^T(S_+\delta C + S_+\delta DF) - F^T[\mathbf{K} - \mathbf{K}\delta J_{11}[\delta\Sigma]\mathbf{K}]^{-1}F \succ 0 \\ \forall \Sigma \equiv \Sigma + \delta\Sigma \equiv \left[\begin{array}{ll} A = \mathbf{A} + \delta A & B = \mathbf{B} + \delta B \\ C = \mathbf{C} + \delta C & D = \mathbf{D} + \delta D \end{array}\right] \in \mathcal{U}_\rho\end{gathered}$$

is a conservative approximation of (40), which in turn is a conservative approximation of (17). Applying the Schur complement lemma, the resulting semi-infinite MI can be rewritten as

$$\left[\begin{array}{c|c|c} [A+BF]^T H + H[A+BF] & (S_+\delta C + S_+\delta DF)^T & F^T \\ -J_{00}[\Sigma] - F^T J_{10}[\Sigma] - J_{01}[\Sigma]F & & \\ \hline (S_+\delta C + S_+\delta DF) & I & \\ \hline F & & \mathbf{K} - \mathbf{K}\delta J_{11}[\delta\Sigma]\mathbf{K} \end{array}\right] \succ 0 \tag{43}$$

$$\forall \Sigma \equiv \Sigma + \delta\Sigma \equiv \left[\begin{array}{ll} A = \mathbf{A} + \delta A & B = \mathbf{B} + \delta B \\ C = \mathbf{C} + \delta C & D = \mathbf{D} + \delta D \end{array}\right] \in \mathcal{U}_\rho.$$

Note that the validity of this semi-infinite LMI automatically implies (41a). Further, the matrix in the left-hand side of the resulting semi-infinite LMI is affine in Σ, so that we can apply the scheme from section 4.1 to build a computationally tractable conservative approximation of this semi-infinite LMI. The approximation is the following

system of LMIs in matrix variables $F, \{Y_\ell\}$:

(44)

$$Y_\ell \succeq \pm \overbrace{\left[\begin{array}{c|c|c} [dA_\ell + dB_\ell F]^T H + H[dA_\ell + dB_\ell F] & & \\ -dC_\ell^T Q\mathbf{C} - \mathbf{C}^T Q dC_\ell & dC_\ell^T S_+ & \\ -\left\{dC_\ell^T[L+Q\mathbf{D}] + \mathbf{C}^T Q dD_\ell\right\} F & +F^T dD_\ell^T S_+ & \\ -F^T\left\{dC_\ell^T[L+Q\mathbf{D}] + \mathbf{C}^T Q dD_\ell\right\}^T & & \\ \hline S_+ dC_\ell & & \\ +S_+ dD_\ell F & & \\ \hline & & -\mathbf{K}dD_\ell^T[L+Q\mathbf{D}]\mathbf{K} \\ & & -\mathbf{K}[L+Q\mathbf{D}]^T dD_\ell \mathbf{K} \end{array}\right]}^{\mathcal{D}^\ell[F]},$$

$$\ell = 1, \ldots, L,$$

$$\rho \sum_{\ell=1}^{L} Y_\ell \prec \left[\begin{array}{c|c|c} [\mathbf{A}+\mathbf{B}F]^T H + H[\mathbf{A}+\mathbf{B}F] - \mathbf{C}^T Q\mathbf{C} & & F^T \\ -F^T[L+Q\mathbf{D}]^T\mathbf{C} - \mathbf{C}^T[L+Q\mathbf{D}]F & & \\ \hline & I & \\ \hline F & & \mathbf{K} \end{array}\right].$$

We arrive at the following result.

PROPOSITION 4.4. *Assume that the matrix*

$$\mathbf{K}^{-1} \equiv \mathbf{D}^T Q\mathbf{D} + L^T\mathbf{D} + \mathbf{D}^T L + R \tag{45}$$

is positive definite. Then

(i) *The system of LMIs* (36a), (36b), (44) *and the LMI*

$$H \prec G \tag{46}$$

in matrix variables $G, F, \{X_\ell, Y_\ell\}$ is a conservative approximation of the system (16), (17), (18) *associated with Problem* 2B. *In particular, the efficiently computable supremum $\widehat{\rho}$ of those $\rho \geq 0$ for which the approximating system is solvable is a lower bound on the optimal value $\rho^\star_{2\mathrm{B}}$ of Problem* 2B.

(ii) *If either*

(a) *C, D are certain (i.e., $dC_\ell = 0$, $dD_\ell = 0$ for all ℓ)*

or

(b) *$Q = 0$ and D is certain,*

then

$$1 \leq \frac{\rho^\star_{2\mathrm{B}}}{\widehat{\rho}} \leq \vartheta(\mu), \tag{47}$$

provided that $\rho^\star_{2\mathrm{B}} > 0$. Here $\vartheta(\mu)$ is the function from Proposition 4.1 *and*

$$\mu = \max\left[\max_{\ell\geq 1, G} \mathrm{Rank}(\mathcal{B}^\ell[G]), \max_{\ell\geq 1, F} \mathrm{Rank}(\mathcal{D}^\ell[F])\right];$$

see (36a), (44) *for the definitions of $\mathcal{B}^\ell[G]$ and $\mathcal{D}^\ell[F]$.*

Tractable conservative approximation of Problem 3B looks exactly like the one for Problem 2B, with the only difference that the LMI (46) should now be replaced with the LMIs

$$0 \preceq G \prec H.$$

The properties of this approximation are completely similar to those established in Proposition 4.4.

4.5. Approximating Problem 3C. Now consider Problem 3C. The system of MIs to be approximated is now comprised of the semi-infinite MI (17) in matrix variables F, H and the LMI $0 \prec H \preceq (1+\delta)\mathbf{Z}_{req}$. The system in question can be rewritten equivalently as

(48)
$$\begin{array}{rcl}
\left[\begin{array}{cc} I & F^T \end{array}\right]\left[\begin{array}{cc} C^T & \\ D^T & I \end{array}\right]\left[\begin{array}{cc} Q & L \\ L^T & R \end{array}\right]\left[\begin{array}{cc} C & D \\ & I \end{array}\right]\left[\begin{array}{c} I \\ F \end{array}\right] & \prec & [(A+BF)^T H + H(A+BF)] \\
& & \forall \left[\begin{array}{cc} A & B \\ C & D \end{array}\right] \in \mathcal{U}_\rho, \\
0 \prec H & \preceq & (1+\delta)\mathbf{Z}_{req}.
\end{array}$$

We can assume that $\mathbf{Z}_{req} \succ 0$—otherwise the system clearly is unsolvable. Let us use the standard change of variables $(H, F) \mapsto (U = H^{-1}, V = FH^{-1})$. Multiplying both sides of (48) from the right and from the left by H^{-1}, we rewrite (48) in the new variables as

(49)
$$\begin{array}{rcl}
\left[\begin{array}{cc} U & V^T \end{array}\right]\left[\begin{array}{cc} C^T & \\ D^T & I \end{array}\right]\underbrace{\left[\begin{array}{cc} Q & L \\ L^T & R \end{array}\right]}_{\mathfrak{P}}\left[\begin{array}{cc} C & D \\ & I \end{array}\right]\left[\begin{array}{c} U \\ V \end{array}\right] & \prec & [AU + UA^T + BV + V^T B^T] \\
& & \forall \left[\begin{array}{cc} A & B \\ C & D \end{array}\right] \in \mathcal{U}_\rho, \\
U & \succeq & (1+\delta)^{-1}\mathbf{Z}_{req}^{-1}.
\end{array}$$

Setting $M \equiv \begin{bmatrix} M_{yy} & M_{yu} \\ M_{yu}^T & M_{uu} \end{bmatrix} = \mathfrak{P}^{1/2}$ (recall that we are in the case of $\mathfrak{P} \succeq 0$) and applying the Schur complement lemma, we can rewrite the latter system equivalently as

(50)
$$\left[\begin{array}{c|c|c}
\begin{array}{c} AU + UA^T \\ +BV + V^T B^T \end{array} & \begin{array}{c} UC^T M_{yy} \\ +V^T[M_{yy}D + M_{yu}]^T \end{array} & \begin{array}{c} UC^T M_{yu} \\ +V^T[M_{yu}^T D + M_{uu}]^T \end{array} \\ \hline
\begin{array}{c} M_{yy}CU \\ +[M_{yy}D + M_{yu}]V \end{array} & I_p & \\ \hline
\begin{array}{c} M_{yu}^T CU \\ +[M_{yu}^T D + M_{uu}]V \end{array} & & I_m
\end{array}\right] \succeq 0$$
$$\forall \left[\begin{array}{cc} A & B \\ C & D \end{array}\right] \in \mathcal{U}_\rho,$$
$$U \succeq (1+\delta)^{-1}\mathbf{Z}_{req}^{-1}.$$

System (50) is in the form of (26); applying the construction from section 4.1, we end up with a tractable conservative approximation of (49), which is the following system of LMIs in matrix variables $U, V, \{X_\ell\}$:

(51)
$$U \succeq (1+\delta)^{-1}\mathbf{Z}_{req}^{-1},$$
$$X_\ell \succeq \pm \underbrace{\left[\begin{array}{c|c|c}
\begin{array}{c} dA_\ell U + UdA_\ell^T \\ +dB_\ell V + V^T dB_\ell^T \end{array} & \begin{array}{c} UdC_\ell^T M_{yy} \\ +V^T dD_\ell^T M_{yy} \end{array} & \begin{array}{c} UdC_\ell^T M_{yu} \\ +V^T dD_\ell^T M_{yu} \end{array} \\ \hline
\begin{array}{c} M_{yy}dC_\ell U \\ +M_{yy}dD_\ell V \end{array} & 0_{p\times p} & \\ \hline
\begin{array}{c} M_{yu}^T dC_\ell U \\ +M_{yu}^T dD_\ell V \end{array} & & 0_{m\times m}
\end{array}\right]}_{\mathcal{E}^\ell[U,V]}, \quad \ell = 1, \ldots, L,$$

$$\rho \sum_{\ell=1}^{L} X_\ell \preceq \left[\begin{array}{c|c|c} \mathbf{A}U + U\mathbf{A}^T & UC^T M_{yy} & UC^T M_{yu} \\ +\mathbf{B}V + V^T\mathbf{B}^T & +V^T[M_{yy}\mathbf{D} + M_{yu}]^T & +V^T[M_{yu}^T\mathbf{D} + M_{uu}]^T \\ \hline M_{yy}\mathbf{C}U & I_p & \\ +[M_{yy}\mathbf{D} + M_{yu}]V & & \\ \hline M_{yu}^T\mathbf{C}U & & I_m \\ +[M_{yu}^T\mathbf{D} + M_{uu}]V & & \end{array}\right].$$

We arrive at the following.

PROPOSITION 4.5. *Assume that the supply matrix* $\mathfrak{P} = \left[\begin{smallmatrix} Q & L \\ L^T & R \end{smallmatrix}\right]$ *is positive semidefinite and that* $\mathbf{Z}_{req} \succ 0$. *Then the system of LMIs* (51) *in matrix variables* $U, V, \{X_\ell\}$ *is a conservative approximation of the system associated with Problem* 3C. *In particular, the efficiently computable supremum* $\widehat{\rho}$ *of those* $\rho \geq 0$ *for which the approximating system is solvable is a lower bound on the optimal value* $\rho^\star_{3C}$ *of Problem* 3C. *For this lower bound, one has*

$$1 \leq \frac{\rho^\star_{3C}}{\widehat{\rho}} \leq \vartheta(\mu), \tag{52}$$

provided that $\rho^\star_{3C} > 0$. *Here* $\vartheta(\mu)$ *is the function from Proposition* 4.1 *and*

$$\mu = \max_{\ell \geq 1, U, V} \mathrm{Rank}(\mathcal{E}^\ell[U, V]);$$

see (51).

4.6. Simplifying approximating systems. A severe practical disadvantage of the tractable approximations of Problems 1, 2A, 2B, 3A, 3B, and 3C we have built is that the sizes of these approximations, although polynomial in the sizes m, n, p, L of the underlying dynamical system and uncertainty set, are quite large. For example, approximation (31) has a single $(m+n+p) \times (m+n+p)$ symmetric matrix variable X_ℓ and two $(m+n+p) \times (m+n+p)$ LMIs per every basic perturbation in the data, so that the design dimension of the approximation is of order of $L(m+n+p)^2$, a quantity which typically is prohibitively large for practical computations. We are about to demonstrate that under favorable circumstances the sizes of the approximating systems can be reduced dramatically. For the sake of simplicity, we restrict our considerations to the case of the approximation (31) associated with Problem 1; the approximations associated with other problems can be processed in a completely similar fashion.

System (31) is of the generic form

$$\mathcal{P}(x) \succeq 0, \tag{53a}$$

$$U_\ell \succeq \pm \mathcal{Q}_\ell(x), \qquad \ell = 1, \ldots, M, \tag{53b}$$

$$V_\ell \succeq \pm R_\ell, \qquad \ell = 1, \ldots, N, \tag{53c}$$

$$\rho \left[\sum_\ell U_\ell + \sum_\ell V_\ell\right] \preceq \mathcal{S}(x), \tag{53d}$$

where

- x is the collection of the original design variables (for (31), $x = Z$);
- U_ℓ, V_ℓ are additional $K \times K$ matrix variables (for (31), $K = m + n + p$, $M + N = L$, the U-variables are those of X_ℓ for which $\mathcal{A}^\ell[Z]$ indeed depends on Z, while the V-variables correspond to those of X_ℓ for $\mathcal{A}^\ell[Z]$ in fact does not depend on Z);

- $\mathcal{P}(x)$, $\mathcal{Q}_\ell(x)$, $\mathcal{S}(x)$ are affine functions of x taking values in the spaces of symmetric matrices of appropriate sizes, and R_ℓ are given $K \times K$ symmetric matrices.

Note that in the situations we are interested in, the ranks of the matrices $\mathcal{Q}_\ell(x)$, R_ℓ are small, provided that the ranks of basic perturbation matrices $dA_\ell, dB_\ell, dC_\ell, dD_\ell$ are small (as indeed is the case in applications). The undesirable large sizes of the approximating system (53) come exactly from the necessity to introduce large-size "matrix bounds" U_ℓ, V_ℓ on the small rank matrices $\mathcal{Q}_\ell(x)$, R_ℓ.

Note that in our applications all we are interested in are the x-components of the feasible solutions of (53). Thus, for our purposes (53) can be replaced with any *x-equivalent* system of LMIs—a system of LMIs $\mathcal{L}(x,y) \succeq 0$ in the original variables x and additional variables y such that the set of x-components of feasible solutions to the latter system is exactly the same as the set of x-components of feasible solutions of (53). What we intend to do is to demonstrate that under favorable circumstances we can build a system of LMIs which is x-equivalent to (53), while being "much smaller" than the latter system. The key to our construction is given by the following two observations.

LEMMA 4.6 (see [3, Lemma 3.1 and Proposition 2.1]). (i) *Let a, b be two nonzero vectors. A symmetric matrix X satisfies the relation*

$$X \succeq \pm[ab^T + ba^T]$$

if and only if there exists positive λ such that

$$X \succeq \lambda aa^T + \frac{1}{\lambda} bb^T.$$

(ii) *Let A be a $n \times n$ symmetric matrix of rank $k > 0$, so that $A = P^T \widehat{A} P$ for appropriately chosen $k \times k$ matrix $\widehat{A}$ and $k \times n$ matrix P of rank k. A symmetric matrix X satisfies the relation*

$$X \succeq \pm A$$

if and only if there exists $k \times k$ symmetric matrix $\widehat{X}$ such that

$$\begin{array}{l} X \succeq P^T \widehat{X} P, \\ \widehat{X} \succeq \pm \widehat{A}. \end{array} \tag{54}$$

Now assume that the matrices $\mathcal{Q}_\ell(x)$ are of the from

$$\mathcal{Q}_\ell(x) = a_\ell b_\ell^T(x) + b_\ell(x) a_\ell^T, \tag{55}$$

where $a_\ell \neq 0$, $b_\ell(x) \not\equiv 0$ are, respectively, a vector and an affine vector-valued function of x. Let also

$$R_\ell = P_\ell^T \widehat{R}_\ell P_\ell : \quad \widehat{R}_\ell = \widehat{R}_\ell^T \in \mathbf{S}^{k_\ell}, \qquad k_\ell = \mathrm{Rank}(R_\ell) > 0.$$

Applying Lemma 4.6, we see that (53) is x-equivalent to the following system of constraints in the original variables x and the additional variables $\lambda_\ell \geq 0$, $\widehat{V}_\ell \in \mathbf{S}^{k_\ell}$:

$$\begin{array}{c} \mathcal{P}(x) \succeq 0, \\ \widehat{V}_\ell \succeq \pm \widehat{R}_\ell, \qquad \ell = 1, \ldots, N, \\ \rho\left[\sum_\ell \left[\lambda_\ell a_\ell a_\ell^T + \frac{1}{\lambda_\ell} b_\ell(x) b_\ell^T(x)\right] + \sum_\ell P_\ell^T \widehat{V}_\ell P_\ell\right] \preceq \mathcal{S}(x) \end{array}$$

(where $\frac{1}{0}bb^T$ is 0 for $b = 0$ and is undefined for $b \neq 0$). The resulting system, via the Schur complement lemma, is x-equivalent to the system of LMIs

$$\mathcal{P}(x) \succeq 0, \tag{56a}$$

$$\left[\begin{array}{c|cccc} X - \sum\limits_{\ell=1}^{M} \lambda_\ell a_\ell a_\ell^T & b_1(x) & b_2(x) & \dots & b_M(x) \\ \hline b_1^T(x) & \lambda_1 & & & \\ b_2^T(x) & & \lambda_2 & & \\ \vdots & & & \ddots & \\ b_M^T(x) & & & & \lambda_M \end{array}\right] \succeq 0, \tag{56b}$$

$$\widehat{V}_\ell \succeq \pm \widehat{R}_\ell, \qquad \ell = 1, \dots, N, \tag{56c}$$

$$\rho\left[X + \sum_\ell P_\ell^T \widehat{V}_\ell P_\ell\right] \preceq \mathcal{S}(x) \tag{56d}$$

in the original variables x and additional scalar variables $\{\lambda_\ell\}_{\ell=1}^M$ and matrix variables X, $\{\widehat{V}_\ell\}_{\ell=1}^N$.

System (56) is x-equivalent to our original system (53) and is usually much better suited for numerical processing than the original system. Indeed, as compared to (53), in (56) there are

- a single $K \times K$ matrix variable X and M scalar variables $\{\lambda_\ell\}_{\ell=1}^M$ instead of M $K \times K$ matrix variables U_ℓ;
- $k_\ell \times k_\ell$ matrix variables $\widehat{V}_\ell$ instead of $K \times K$ matrix variables V_ℓ, and $k_\ell \times k_\ell$ LMIs (56c) instead of $K \times K$ LMIs (53c) (recall that k_ℓ are assumed to be small as compared to K);
- a single LMI (56b) instead of M LMIs (53b). Although the size of LMI (56b) is larger than those of LMIs (53b), the LMI is of very simple arrow structure and is extremely sparse.

It remains to understand what should be required from the uncertainty set $\mathcal{U}_\rho$ in order to ensure that the approximations associated with Problems 1, 2A, 2B, 3A, 3B, and 3C possess property (55) and thus admit the outlined simplification. The corresponding requirements are as follows:

A. In the case of Problems 1, 2A, 3A, it suffices to assume the following:
   - A.1. The parts $[A, B]$ and $[C, D]$ of the matrix $\Sigma = \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$ are perturbed independently (i.e., for every ℓ exactly one of the matrices $[dA_\ell, dB_\ell]$, $[dC_\ell, dD_\ell]$ is nonzero).
   - A.2. The basic perturbations of the part $[A, B]$ of Σ are of ranks ≤ 1.

   Note that under these assumptions the quantity μ in Propositions 4.2, 4.3 and the above quantities k_ℓ satisfy the relation

$$k_\ell \leq \mu \leq 2 \max\left[1, \max_\ell (\mathrm{Rank}(dC_\ell) + \mathrm{Rank}(dD_\ell))\right].$$

B. In the case of Problems 2B, 3B, it suffices to assume the following.
   - B.1. The parts A, B, C, D of Σ are perturbed independently (i.e., for every ℓ exactly one of the matrices dA_ℓ, dB_ℓ, dC_ℓ, dD_ℓ is nonzero).
   - B.2. The basic perturbations of the parts A, B, C, D of Σ are of ranks ≤ 1 and
     - i. either $Q = 0$

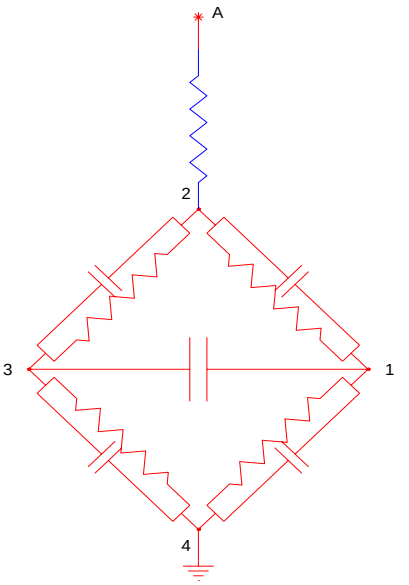


FIG. 1. *"Bridge."*

ii. or D is certain.

Note that under these assumptions the quantity μ in Proposition 4.4 and the above quantities k_ℓ satisfy the relation

$$k_\ell \le \mu \le 2.$$

C. In the case of Problem 3C, it suffices to assume that

C.1. the basic perturbations $d\Sigma_\ell$ are of ranks ≤ 1.

Note that under these assumptions the quantity μ in Proposition 4.5 equals 2.

Note that the sets A.1–A.2, B.1–B.2, C.1 of the assumptions are satisfied in the simplest case of the *interval uncertainty*—every entry in Σ, independently of other entries, runs through a given interval. In this case, $k_\ell \le \mu = 2$, and the corresponding "tightness bound" $\vartheta(\mu)$ (see (32), (37), (47), (52)) becomes $\frac{\pi}{2}$.

5. Illustrating examples. Here we present three simple illustrations of the proposed approach. The first two of them correspond to the positive-real case, while the third has to do with the linear-quadratic case.

5.1. Positive-real case. Consider the simple RC circuit ("bridge") presented in Figure 1. The input is the outer voltage applied between the node A and the ground, the output is the current through the circuit. The state variables are the potentials at the nodes 1, 2, 3 (normalized by the condition that the potential of the ground is identically zero). Applying the Kirchoff laws, the description of the system becomes

$$\begin{aligned} \dot z(t) &= A_{c,r} z(t) + B_{c,r} u(t), \\ y(t) &= C_r z(t) + D_r u(t), \end{aligned} \tag{57}$$

where we have the following:

- $c \in \mathbf{R}^{10}$ is the vector of capacitances of the capacitors in the 10 arcs of the circuit (9 "visible arcs" and the external arc from node 2 via point A to the ground; for arc i with no capacitor, $c_i = 0$).
- $r \in \mathbf{R}^{10}$ is the vector of conductances of the resistors in the 10 arcs of the circuit (for arc i with no resistor, $r_i = 0$).
- the matrix $\Sigma = \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$ is given by

$$\Sigma = \Sigma_{c,r} \equiv \left[\begin{array}{c|c} -[P^T\mathrm{Diag}\{c\}P]^{-1}[P^T\mathrm{Diag}\{r\}P] & [P^T\mathrm{Diag}\{c\}P]^{-1}[P^T\mathrm{Diag}\{r\}J], \\ \hline -[P^T\mathrm{Diag}\{r\}J], & J^T\mathrm{Diag}\{r\}J, \end{array}\right],$$

where $\mathrm{Diag}\{p\}$ denotes the diagonal matrix with diagonal entries given by vector p and

- P is the *incidence matrix*. The rows of P are indexed by the 10 arcs in the circuit, the columns are indexed by the 3 nonground nodes 1, 2, 3 and the element P_{ij} is equal to $+1$, -1 or 0 depending on whether node # j starts arc # i, ends this arc, or is not incident to the arc. For our circuit, P is as follows (R stands for arcs with resistors, C for arcs with capacitors):

Arcs			Nodes		
Origin	Destination	Type	1	2	3
1	2	R	1	−1	0
1	2	C	1	−1	0
2	3	R	0	1	−1
2	3	C	0	1	−1
3	4	R	0	0	1
3	4	C	0	0	1
4	1	R	−1	0	0
4	1	C	−1	0	0
1	3	C	1	0	−1
2	→ A → 4	R	0	1	0

- $J = (0, \ldots, 0, 1)^T \in \mathbf{R}^{10}$ "points" to the external arc (which in our enumeration is the last of the 10 arcs of the circuit).

We treat as the uncertain parameters the capacitances of the capacitors and the conductances of the resistors (except for the "outer" resistor in the external arc; it represents the inner resistance of the outer supply and is assumed to be certain) and assume that every one of these parameters can vary, independently of others, by at most ρ times the nominal value of the parameter, where ρ is the uncertainty level in question. The nominal values of the data are given in Table 1. Here is the nominal instance (entries are rounded to 4 digits after the dot):

$$\Sigma = \left[\begin{array}{rrr|r} -0.5005 & -50.0000 & -0.4995 & 50.0000 \\ 0.1000 & -101.1000 & 0.0000 & 100.0000 \\ -0.4995 & -50.0000 & -0.5005 & 50.0000 \\ \hline 0 & -100.0000 & 0 & 100.0000 \end{array}\right].$$

The elements of the matrix $\Sigma_{c,r}$ are nonlinear functions of the "physical data" c, r, so that an interval uncertainty in the latter data is not equivalent to a box uncertainty in $\Sigma_{c,r}$. We neglect this phenomenon by linearizing $\Sigma_{r,c}$ at the nominal data, thus

TABLE 1
Nominal values for the bridge circuit.

Element	Nominal value	Element	Nominal value
R_{12}	1.2	C_{12}	1.0
R_{23}	1.0	C_{23}	1.0
R_{34}	1.0	C_{34}	1.0
R_{41}	1.0	C_{41}	1.0
R_{2A}	100	C_{13}	1000

arriving at a box uncertainty set with $L = 9$ basic perturbation matrices, according to the number of uncertain capacitances and conductances in the circuit. Note that for our particular circuit, the resulting uncertainty affects only the $[A, B]$-part of Σ, and the basic perturbation matrices $[dA_\ell, dB_\ell]$ are of rank 1.

Recall that the supply in the SISO positive-real case is $2yu$, i.e.,

$$\mathfrak{P} = \left[\begin{array}{cc} Q = 0 & L = 1 \\ L^T = 1 & R = 0 \end{array} \right];$$

for our RC circuit, the supply is nothing but (twice) the electrical power pumped into the circuit by the external voltage.

We have carried out two experiments with the outlined system: the first deals with extracting the energy stored in the circuit, and the second with moving the circuit from the zero initial state to a given state.

Extracting available energy. The question we are addressing is to find the largest level $\rho^\star_{av}$ of uncertainty for which the "performance" Θ of the "ideal extracting feedback" $\mathbf{F}_{av}$ (see D.4) corresponding to the nominal instance is at least $1 - \epsilon$, i.e., this feedback still allows, for every perturbed instance and every initial state ζ of the circuit, to extract at least $(1 - \epsilon)$-part of the nominal available storage $\zeta^T \mathbf{Z}_{av} \zeta$. In our experiment, we set $\epsilon = 0.1$. Solving the conservative approximation

$$\max_{\rho, G, H, \{X_\ell, Y_\ell\}} \{\rho : \ (\rho, G, H, \{X_\ell, Y_\ell\}) \text{ satisfies (36)}\}$$

of the associated Problem 2A, we end up with a lower bound

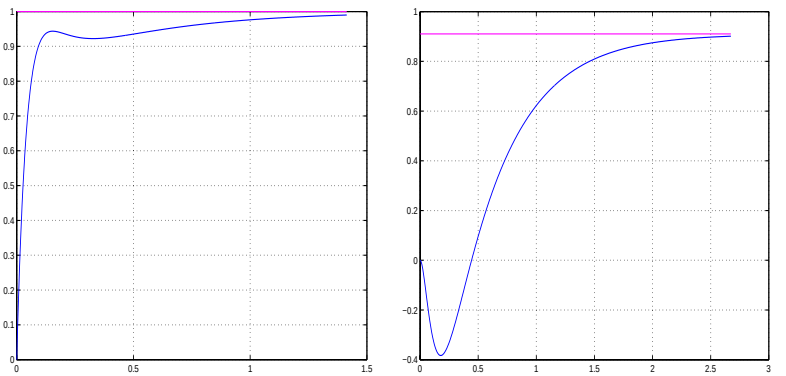
$$\widehat{\rho} = 1.1\mathsf{e}{-3}$$

on $\rho^\star_{av}$; in other words, we can be sure that with 0.11% perturbations of the uncertain capacitances and conductances, the nominal feedback $\mathbf{F}_{av}$ still allows us to extract at least 90% of the nominal available storage, whatever is the initial state of the circuit. A natural question arises, How conservative is our bound? Recall that there are two reasons for it to be conservative:

- First, the bound comes from solving a conservative approximation of Problem 2A rather than from solving the problem itself; according to Proposition 4.3, the true optimal value in the problem is at most $\frac{\pi}{2}$ times larger than the bound (recall that we are in the situation of $Q = 0$ and $\mu = 2$).
- Second, and worse, even the true optimal value in Problem 2A is a lower bound on $\rho^\star_{av}$, since the problem comes from the *sufficient* condition, stated by Proposition 3.2, for "good" performance of the nominal feedback $\mathbf{F}_{av}$ under data perturbations. Note that we have no idea how conservative this sufficient condition is.

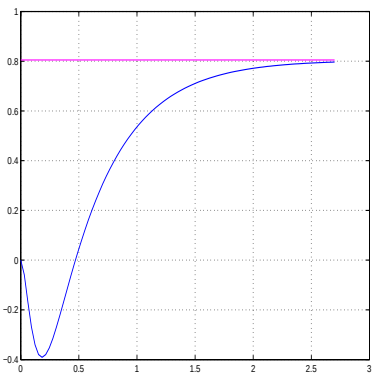
TABLE 2
Performance of the nominal feedback $\mathbf{F}_{av}$ *versus uncertainty level.*

ρ	$1.2\widehat{\rho} = 1.3\mathsf{e}{-3}$	$2.2\widehat{\rho} = 2.3\mathsf{e}{-3}$	$3\widehat{\rho} = 3.2\mathsf{e}{-3}$
Θ	0.893	0.805	0.736



$\rho = 0$
$\Theta = 1$

$\rho = \widehat{\rho} = 1.1\mathsf{e}{-3}$
$\Theta = 0.910$



$\rho = 2.2\widehat{\rho} = 2.3\mathsf{e}{-3}$
$\Theta = 0.805$

FIG. 2. *Sample plots of* $\frac{E_{av}(t)}{z^T(0)\mathbf{Z}_{av}z(0)}$.

In spite of these pessimistic considerations, the experiment shows that our bound is pretty tight. Looking through all $2^L = 512$ "extreme" perturbations of the data, and playing with the initial state of the circuit, we found out that the worst-case (with respect to relative perturbations of the uncertain entries in c, r of level ρ and initial states) performance Θ of the ideal nominal feedback is *at most* as given in Table 2. In particular, we see that with the level of perturbations $1.2\widehat{\rho}$, the worst-case performance of the ideal nominal feedback is less than 0.9 ($\equiv 1-\epsilon$) times the nominal available storage. It follows that $\rho^\star_{av} \leq 1.2\widehat{\rho}$, i.e., our bound $\widehat{\rho}$ is within 20% margin of the quantity of interest.

Figure 2 represents three sample plots of the extracted energy $E_{av}(t)$ as a function of time for the feedback $\mathbf{F}_{av}$.

TABLE 3
Price of the nominal feedback $\mathbf{F}_{req}$ versus uncertainty level.

ρ	$1.2\widehat{\rho} = 5.5\mathtt{e}{-4}$	$2.2\widehat{\rho} = 1.0\mathtt{e}{-3}$	$3\widehat{\rho} = 1.4\mathtt{e}{-3}$
Γ	1.056	1.105	1.148

Moving the circuit to a given state. Now let us try to find the largest uncertainty level $\rho^\star_{req}$ for which the "price" Γ of the "ideal driving feedback" $\mathbf{F}_{req}$ (see D.4) corresponding to the nominal instance is at most $1+\delta$, i.e., this feedback still allows, for every perturbed instance and every target state ζ of the circuit, to move the circuit from the zero initial state to the state ζ while pumping into the circuit at most $(1+\delta)$ times the nominal required energy $\zeta^T\mathbf{Z}_{req}\zeta$. In our experiment, we set $\delta = 0.1$. Solving the conservative approximation of the associated Problem 3A (see the end of section 4.3), we end up with a lower bound

$$\widehat{\rho} = 4.6\mathtt{e}{-4}$$

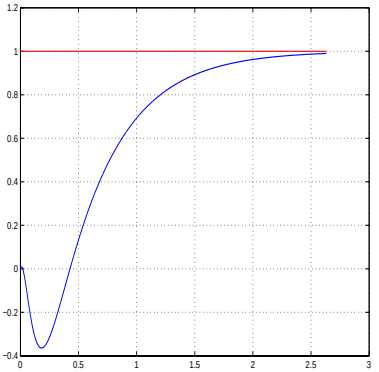
on $\rho^\star_{req}$; thus, we can be sure that with 0.046% perturbations of the uncertain capacitances and conductances, the ideal nominal feedback $\mathbf{F}_{req}$ still allows us to move the circuit from the zero state to (any) target one while pumping into the circuit at most 110% of the nominal required energy. It turns out that our bound is perhaps not as tight as in the previous case, but still is good enough. Indeed, looking at the data in Table 3, which represent *lower* bounds on the price of the ideal nominal driving feedback $\mathbf{F}_{req}$ under different levels of perturbations, we see that with the perturbations of the level $2.2\widehat{\rho}$ the price of moving the circuit to certain target state ζ by the feedback $\mathbf{F}_{req}$ can be larger than 1.1 ($\equiv 1+\delta$) times the nominal required energy $\zeta^T\mathbf{Z}_{req}\zeta$; hence $\rho^\star_{req} \le 2.2\widehat{\rho}$. Note that, in the case in question, the conservative approximation of Problem 3A contributes to the ratio $\rho^\star_{req}/\widehat{\rho} \approx 2.2$ a factor $\le \frac{\pi}{2} = 1.57$; the remaining factor in the ratio (which is at least $2.2/1.57 \approx 1.4$) comes from the conservativeness of the sufficient condition expressed in Proposition 3.3 and underlying Problem 3A.

Figure 3 presents three sample plots of the pumped energy $E_{req}(t)$ as a function of time for the feedback $\mathbf{F}_{req}$.

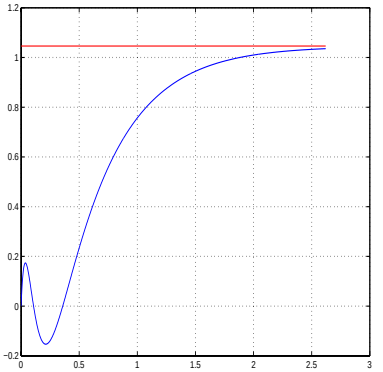
5.2. Linear-quadratic case. Consider the mechanical system shown on Figure 4; it consists of 5 material points in a two-dimensional plane linked to each other by elastic springs as shown on the figure; the points can slide without friction along the respective axes $01, \dots, 05$. The nominal data for the system are given in Table 4. The system is controlled by two external forces acting at the masses 1 and 5. The first 5 components of the state vector are the shifts x_i of the points from their equilibrium positions along the lines of motion, and the next 5 components are the linear velocities $\dot{x}_i$ of the points; these velocities are the outputs of the system. With respect to these states, the dynamical system in question is

$$\begin{array}{rcl} \frac{d}{dt}\left[\begin{array}{c} x \\ \dot{x} \end{array}\right] &=& \left[\begin{array}{c|c} & I_5 \\ \hline -M^{-1}E & \end{array}\right]\left[\begin{array}{c} x \\ \dot{x} \end{array}\right] + Bu, \\ y &=& \dot{x}, \end{array} \tag{58}$$

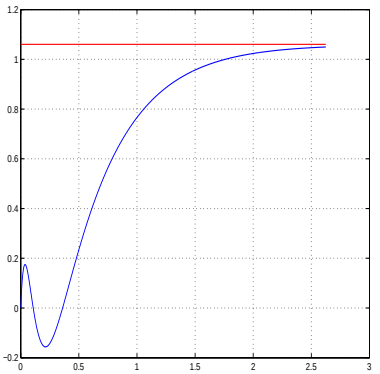
where M is the diagonal matrix with the masses $m(i)$ of the points as the diagonal entries, E is the *stiffness matrix* readily given by the rigidities of the springs and the equilibria positions of the points, and B is the 10×2 matrix with two nonzero



$\rho = 0$
$\Gamma = 1$



$\rho = \widehat{\rho} = 4.6\mathsf{e}{-4}$
$\Gamma = 1.046$



$\rho = 2.2\widehat{\rho} = 1.0\mathsf{e}{-3}$
$\Gamma = 1.105$

FIG. 3. *Sample plots of* $\frac{E_{req}(t)}{z^T(0)\mathbf{Z}_{req}z(0)}$.

entries $B_{5,1} = m^{-1}(1)$ and $B_{10,2} = m^{-1}(5)$. Here is the nominal instance (entries are rounded to 4 digits after the dot):

$$\Sigma = \left[\begin{array}{ccccc|cccccc|c|c} & & & & & 1 & & & & & & & \\ & & & & & & 1 & & & & & & \\ & & & & & & & 1 & & & & & \\ & & & & & & & & 1 & & & & \\ & & & & & & & & & 1 & & & \\ -2.5647 & -1.0797 & & & -1.0890 & & & & & & & 1.9637 & \\ -0.6038 & -0.8206 & -0.4766 & & & & & & & & & & \\ & -0.6009 & -1.5044 & -0.4808 & & & & & & & & & \\ & & -0.4300 & -1.1142 & -0.5131 & & & & & & & & \\ -0.6190 & & & -0.4626 & -0.8352 & & & & & & & 1.1161 & \\ \hline & & & & & 1 & & & & & & & \\ & & & & & & 1 & & & & & & \\ & & & & & & & 1 & & & & & \\ & & & & & & & & 1 & & & & \\ & & & & & & & & & 1 & & & \end{array}\right].$$

We are interested to bring the system from the equilibrium to a given state while

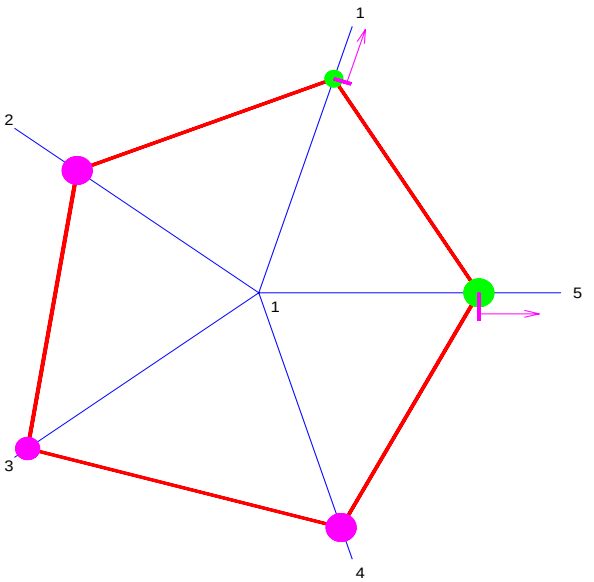


FIG. 4. 5 *masses linked by elastic springs*

TABLE 4
The nominal data.

Point	Mass	Distance to the origin at equilibrium	Spring	Rigidity
1	0.5093	0.8034	1 - 2	1.461
2	0.9107	0.7430	2 - 3	1.369
3	0.7224	0.9456	3 - 4	1.088
4	0.8077	0.8810	4 - 5	1.203
5	0.8960	0.7282	5 - 1	1.468

minimizing the cost functional

$$\int_0^\infty \left[\sum_{i=1}^{5}(\dot{x}_i)^2(t) + \sum_{i=1}^{2} u_i^2(t)\right] dt,$$

which is equivalent to the providing required supply problem with the supply matrix

$$\mathfrak{P} = \left[\begin{array}{c|c} Q = I_5 & L = 0_{5\times 2} \\ \hline L^T = 0_{2\times 5} & R = I_2 \end{array}\right].$$

In our experiment, we treat as uncertain parameters the masses of the points and the rigidities of the springs and assume that every one of these parameters can vary, independently of others, by at most ρ times the nominal value of the parameter. Note that the perturbations affect only the $[A, B]$-part of the matrix Σ of the system and that the dependence of Σ on the masses and rigidities is nonlinear (although both M and E in (58) are affine in the parameters). As in the previous example, we neglect this phenomenon by linearizing Σ at the nominal data, and end up with a box

uncertainty set with $L = 10$ basic perturbation matrices, according to the number of uncertain parameters; all these perturbation matrices turn out to be of rank 1. The outlined model underlies two numerical experiments we are about to report.

Designing robust feedback with "nearly optimal" performance. For the nominal system, there exists the ideal state feedback $u = \mathbf{F}z$ which moves the system from the equilibrium to (any) given initial state ζ at the minimum possible cost $\zeta^T \mathbf{Z}_{req}\zeta$. What we are interested in now is to find the largest uncertainty level for which there still exists an instance-independent state feedback with a given *performance index* $1+\delta$; the latter means that the feedback allows to move every instance of the perturbed system from the equilibrium to (any) given state ζ at the cost at most $(1+\delta)$ times the "ideal nominal cost" $\zeta^T \mathbf{Z}_{req}\zeta$. In our experiment, we set $\delta = 0.1$ and get the desired feedback by solving the conservative approximation (50) of Problem 3C associated with the outlined model. As a result, we get

(a) state feedback with the matrix

$$F = \begin{bmatrix} -0.0396 & 0.0220 & -0.3685 & -0.8069 & -0.4099 & 0.0152 & -0.3694 & 0.0647 & -0.0498 & 1.3167 \\ -0.3993 & -0.6453 & -0.4886 & -0.2269 & -0.0322 & 1.1859 & -0.5896 & -0.2165 & -0.3263 & 0.0268 \end{bmatrix},$$

which is slightly different from the ideal nominal feedback

$$\mathbf{F} = \begin{bmatrix} -0.0281 & 0.0289 & -0.4196 & -0.8948 & -0.4551 & 0.0063 & -0.3897 & 0.0628 & -0.0558 & 1.3826 \\ -0.4467 & -0.7133 & -0.5466 & -0.2423 & -0.0311 & 1.2269 & -0.6375 & -0.2570 & -0.3520 & 0.0111 \end{bmatrix},$$

and

(b) the "safe" uncertainty level $\widehat{\rho} = 0.0048$, which is a lower bound on the optimal value $\rho^\star_{3C}$ in Problem 3C.

What we know about F and $\widehat{\rho}$ from their origin is the following:

- The performance index of the state feedback $u = Fz$ is no worse than $1 + \delta$, provided that the level of perturbations does not exceed 0.48% (which is our $\widehat{\rho}$). Note that this statement remains true even for dynamical perturbations.
- The true optimal value $\rho^\star_{3C}$ in Problem 3C is at most $\frac{\pi}{2}$ times larger than $\widehat{\rho}$ (see Proposition 4.5; note that our basic perturbation matrices are of rank 1, so that the quantity μ in (52) equals 2 by item C of Section 4.6).

What we are interested in now is how conservative are our results, specifically, what is the actual value of the ratio $\rho^\star_{3C}/\widehat{\rho}$. An even more important question is as follows. The optimal value $\rho^\star_{3C}$ of Problem 3C is itself no more than a lower bound on the supremum $\rho^\star$ of those perturbation levels for which there still exists a state feedback with performance index $1 + \delta = 1.1$ (since what underlies Problem 3C is no more than a sufficient condition for good performance under uncertainty). How large is the ratio $\rho^\star/\widehat{\rho}$, or, in other words, how far is the robustness of our feedback F from the "ideal" robustness compatible with the prescribed performance index 1.1? It turns out that the answers to these questions are quite assuring. Indeed, looking at a large enough number of randomly perturbed instances with different perturbation levels and computing the required supply for these instances, one can find out that already at the perturbation level $1.2\widehat{\rho} = 0.0058$ there exist perturbed instances Σ and target states ζ such that Σ cannot be moved from the equilibrium to the state ζ at the cost $\leq 1.1\zeta^T \mathbf{Z}_{req}\zeta$. It follows that

$$\rho^\star_{3C} \leq \rho^\star < 1.2\widehat{\rho},$$

which is much better than we could expect.

Lyapunov stability analysis. Here we use the data yielded by the previous experiment for illustrating another application of the proposed approach, namely,

estimating the level of perturbations which keep the closed-loop system stable. This problem was the subject of Example 4 in section 3.1, where it was shown that the problem can be posed as the one of finding the supremum of those uncertainty levels for which all perturbed instances of the system share a common dissipativity certificate. As our sample closed-loop system, we used the outlined mechanical system equipped with the state feedback F found in the previous experiment. Our uncertainty model for the matrix

$$\widehat{A} = A + BF$$

of the closed-loop system is as follows: we use the aforementioned "physical" model of perturbations in $[A, B]$ and assume, in addition, that the entries in F also are subject to perturbations. Since we have no physical model of the controller, we assume that the entries F_{ij} in F can vary, independently of each other (and independently of the perturbations in $[A, B]$), in the intervals $[F^c_{ij} - \rho|F^c_{ij}|, F^c_{ij} + \rho|F^c_{ij}|]$, where ρ is the uncertainty level, and F^c_{ij} are the "nominal" values as computed in the previous experiment.

As in the previous cases, we linearized the dependence of $\widehat{A}$ on the perturbations, thus arriving at a box model of perturbations in the matrix of the closed-loop system. Then we solved the conservative approximation (31) of Problem 1 associated with system (14) and the supply matrix (15). Since we were interested solely in the stability of the closed-loop system under perturbations and did not care of any kind of performance, we looked for the common dissipativity certificate Z in a pretty wide "matrix interval" $\mathcal{I} = \{Z : 10^{-7}\mathbf{Z} \preceq Z \preceq \mathbf{Z}\}$, which in the situation of Example 4 basically means that we do not impose restrictions on Z except for being positive definite.

The results of our experiment are as follows. The solution of (31) yields a level of perturbations $\widehat{\rho} = 0.041$ and a positive definite matrix Z, which is a common Lyapunov stability certificate for all perturbed instances of the matrix $\widehat{A}$ of the closed-loop system when the level of perturbations is $\widehat{\rho}$. Thus, we can be sure that the closed-loop system remains stable whatever are 4.1% perturbations of the physical parameters of our mechanical system and 4.1% perturbations of the coefficient in the feedback matrix, even when these perturbations are dynamical. A natural question is, How conservative is this conclusion? Note that, a priori, there is no reason to be too optimistic in this respect, since the existence of a common Lyapunov stability certificate, as a sufficient condition for stability, may be quite conservative already by itself, and we are dealing with conservative approximation of this condition. However, the experiment demonstrates that we are lucky: simulating about 1,000 random perturbations of the closed-loop system at different uncertainty levels, it turns out that at the uncertainty level $1.6\widehat{\rho} = 0.065$ there already exist perturbations which make the closed-loop system unstable. Thus, the closed-loop system definitely survives perturbations not exceeding 4.1% and can be crushed by 6.5% perturbations.

6. Conclusions. We have developed techniques for specifying the magnitudes of dynamic perturbations in the parameters of a linear system which preserve a desired property of the system (such as positive-realness, nonexpansiveness, etc.). The standard sufficient condition for this is the solvability of an associated *infinite* system $\mathcal{S}$ of linear matrix inequalities. The latter condition, however, is usually NP-hard to verify, so that one is forced to look for *efficiently verifiable* sufficient conditions for $\mathcal{S}$ to be solvable. We propose such a condition and demonstrate that in many cases it is *provably tight, within an absolute constant factor*, $\frac{\pi}{2}$ in most cases (for details, see

Propositions 4.2, 4.3, 4.4, 4.5). This "guaranteed tightness" is a specific (and, to the best of our knowledge, unique) feature of the paper.

Recently, it turned out that the matrix cube theorem, which underlies all our developments, can be extended to the complex case and even with a model of uncertainty richer than the interval one. These extensions could then imply corresponding extensions of the results we have presented here.

REFERENCES

[1] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust semidefinite programming*, in Handbook on Semidefinite Programming, H. Wolkowicz, R. Saigal, and L. Vandenberghe, eds., Kluwer Academic, Norwell, MA, 2000, pp. 139–162.

[2] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, MPS SIAM Ser. Optim. 2, SIAM, Philadelphia, 2001.

[3] A. Ben-Tal and A. Nemirovski, *On tractable approximations of uncertain linear matrix inequalities affected by interval uncertainty*, SIAM J. Optim., 12 (2002), pp. 811–833.

[4] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.

[5] L. El Ghaoui, H. Lebret, *Robust solutions to least-squares problems with uncertain data*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 1035–1064.

[6] L. El Ghaoui, F. Oustry, H. Lebret, *Robust solutions to uncertain semidefinite programs*, SIAM J. Optim., 9 (1998), pp. 33–52.

[7] A. Nemirovski, *Several NP-hard problems arising in robust stability analysis*, Math. Control Signal Systems, 6 (1993), pp. 99–105.

[8] C. Scherer and S. Weiland, *Linear Matrix Inequalities in Control*, Dutch Institute of Systems and Control graduate course lecture notes, version 2.0, http://www.ocp.tudelft.nl/sr/personal/Scherer/lmi.pdf (April 1999).

[9] J. C. Willems, *Dissipative dynamical systems, part* I*: General theory*, Arch. Rational Mech. Anal., 45 (1971), pp. 321–351.

[10] J. C. Willems, *Dissipative dynamical systems, part* II*: Linear systems with quadratic supply rates*, Arch. Ration. Mech. Anal., 45 (1971), pp. 352–393.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 6, pp. 1696–1721

MULTIDIMENSIONAL BACKWARD STOCHASTIC RICCATI EQUATIONS AND APPLICATIONS*

MICHAEL KOHLMANN† AND SHANJIAN TANG‡

Abstract. Backward stochastic Riccati differential equations (BSRDEs for short) arise from the solution of general linear quadratic optimal stochastic control problems with random coefficients. The existence and uniqueness question of the global adapted solutions has been open since Bismut's pioneering research publication in 1978 [*Séminaire de Probabilités* XII, Lecture Notes in Math. 649, C. Dellacherie, P. A. Meyer, and M. Weil, eds., Springer–Verlag, Berlin, 1978, pp. 180–264]. One distinguishing difficulty lies in the quadratic nonlinearity of the drift term in the second unknown component. In a previous article [*Stochastic Process. Appl.*, 97 (2002), pp. 255–288], the authors solved the one-dimensional case driven by Brownian motions. In this paper the multidimensional case driven by Brownian motions is studied. A closeness property for solutions of BSRDEs with respect to their coefficients is stated and is proved for general BSRDEs, which is used to obtain the existence of a global adapted solution to some BSRDEs. The global existence and uniqueness results are obtained for two classes of BSRDEs, whose generators contain a quadratic term of L (the second unknown component). More specifically, the two classes of BSRDEs are (for the regular case $N > 0$)

$$\begin{cases} dK = -[A^*K + KA + Q - LD(N + D^*KD)^{-1}D^*L]\,dt + L\,dw, \\ K(T) = M, \end{cases}$$

under the condition $d = 1$, and (for the singular case)

$$\begin{cases} dK = -[A^*K + KA + C^*KC + Q + C^*L + LC \\ \qquad -(KB + C^*KD + LD)(D^*KD)^{-1}(KB + C^*KD + LD)^*]\,dt + L\,dw, \\ K(T) = M, \end{cases}$$

under the condition $d = 1$ and $m = n$. The arguments given in this paper are completely new, and they consist of some simple techniques of algebraic matrix transformations and direct applications of the closeness property mentioned above. We make full use of the special structure (the nonnegativity of the quadratic term, for example) of the underlying Riccati equation. Applications in optimal stochastic control are exposed.

Key words. backward stochastic Riccati equation, stochastic linear quadratic control problem, algebraic transformation, Feynman–Kac formula

AMS subject classifications. 90A09, 90A46, 93E20, 60G48

PII. S0363012900378760

1. Introduction.

Let d and d_0 be two nonnegative integers with $d_0 \le d$. Let $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t, 0 \le t \le T\})$ be a fixed complete probability space on which is defined a standard d-dimensional $\mathcal{F}_t$-adapted Brownian motion $w(t) \equiv (w_1(t), \dots, w_d(t))^*$. Assume that $\{\mathcal{F}_t, 0 \le t \le T\}$ is the completion, by the totality $\mathcal{N}$ of all null sets of $\mathcal{F}$, of the natural filtration $\{\mathcal{F}_t^w, 0 \le t \le T\}$ generated by w. Denote by $\{\mathcal{G}_t, 0 \le t \le T\}$ the P-augmented natural filtration generated by the $(d - d_0)$-dimensional Brownian motion $(w_{d_0+1}, \dots, w_d)$. Assume that all the coefficients A, B, C_i, and D_i

*Received by the editors September 29, 2000; accepted for publication (in revised form) June 26, 2002; published electronically February 4, 2003. The authors gratefully acknowledge the support by the Center of Finance and Econometrics, University of Konstanz.

http://www.siam.org/journals/sicon/41-6/37876.html

†Department of Mathematics and Statistics, University of Konstanz, D-78457, Konstanz, Germany (michael.kohlmann@uni-konstanz.de).

‡Laboratory of Mathematics for Nonlinear Sciences and Department of Mathematics, Fudan University, Shanghai 200433, China (sjtangk@online.sh.cn). This author is supported by a Research Fellowship from the Alexander von Humboldt Foundation under the project "A Stochastic Control-Theoretic Treatise on Continuous-Time Investment and Derivative Pricing" and by the National Natural Science Foundation of China under grant 79790130.

are $\{\mathcal{F}_t, , 0 \le t \le T\}$-progressively measurable bounded matrix-valued processes, defined on $\Omega \times [0,T]$, of dimensions $n \times n$, $n \times m$, $n \times n$, and $n \times m$, respectively, with $i = 1, \dots, d$. Also assume that M is an $\mathcal{F}_T$-measurable nonnegative bounded $n \times n$ random matrix, and Q and N are $\{\mathcal{F}_t, 0 \le t \le T\}$-progressively measurable, bounded, nonnegative, and uniformly positive $n \times n$ and $m \times m$ matrix processes, respectively.

Consider the following backward stochastic Riccati differential equation (BSRDE for short):

(1)
$$\begin{cases} dK = -\left[A^*K + KA + \sum_{i=1}^{d} C_i^* K C_i + Q + \sum_{i=1}^{d} (C_i^* L_i + L_i C_i)\right. \\ \qquad -\left(KB + \sum_{i=1}^{d} C_i^* K D_i + \sum_{i=1}^{d} L_i D_i\right)\left(N + \sum_{i=1}^{d} D_i^* K D_i\right)^{-1} \\ \qquad \left.\times \left(KB + \sum_{i=1}^{d} C_i^* K D_i + \sum_{i=1}^{d} L_i D_i\right)^*\right] dt + \sum_{i=1}^{d} L_i\, dw_i, \qquad 0 \le t < T, \\ K(T) = M. \end{cases}$$

It will be called the BSRDE $(A, B; C_i, D_i, i = 1, \dots, d; Q, N, M)$ in the following for convenience of indicating the concerned coefficients. When the coefficients A, B, C_i, D_i, Q, N, and M are all deterministic, then $L_1 = \cdots = L_d = 0$ and BSRDE (1) is reduced to the following nonlinear matrix ordinary differential equation:

(2)
$$\begin{cases} dK = -\left[A^*K + KA + \sum_{i=1}^{d} C_i^* K C_i + Q - \left(KB + \sum_{i=1}^{d} C_i^* K D_i\right)\right. \\ \qquad \left.\times \left(N + \sum_{i=1}^{d} D_i^* K D_i\right)^{-1} \left(KB + \sum_{i=1}^{d} C_i^* K D_i\right)^*\right] dt, \\ \qquad 0 \le t < T, \\ K(T) = M, \end{cases}$$

which was solved by Wonham [31] by applying Bellman's principle of quasi linearization and a monotone convergence approach. Bismut [2, 3] initially studied the case of random coefficients, but he could solve only some special simple cases. He always assumed that the randomness of the coefficients comes only from a smaller filtration $\{\mathcal{G}_t\}$, which leads to $L_1 = \cdots = L_{d_0} = 0$. He further assumed in his paper [2] that

$$C_{d_0+1} = \cdots = C_d = 0, \qquad D_{d_0+1} = \cdots = D_d = 0, \tag{3}$$

under which BSRDE (1) becomes the following one:

(4)
$$\begin{cases} dK = -\left[A^*K + KA + \sum_{i=1}^{d_0} C_i^* K C_i + Q\right. \\ \qquad -\left(KB + \sum_{i=1}^{d_0} C_i^* K D_i\right)\left(N + \sum_{i=1}^{d_0} D_i^* K D_i\right)^{-1} \left.\left(KB + \sum_{i=1}^{d_0} C_i^* K D_i\right)^*\right] dt \\ \qquad + \sum_{i=d_0+1}^{d} L_i\, dw_i, \qquad 0 \le t < T, \\ K(T) = M, \end{cases}$$

and the generator does not involve L at all. In his work [3] he assumed that

$$D_{d_0+1} = \cdots = D_d = 0, \tag{5}$$

under which BSRDE (1) becomes the following one:

(6)
$$\begin{cases} dK = -\left[A^*K + KA + \sum_{i=1}^{d} C_i^* K C_i + Q + \sum_{i=d_0+1}^{d} (C_i^* L_i + L_i C_i) \right. \\ \qquad - \left. \left(KB + \sum_{i=1}^{d_0} C_i^* K D_i\right)\left(N + \sum_{i=1}^{d_0} D_i^* K D_i\right)^{-1}\left(KB + \sum_{i=1}^{d_0} C_i^* K D_i\right)^*\right] dt \\ \qquad + \sum_{i=d_0+1}^{d} L_i\, dw_i, \qquad 0 \le t < T, \\ K(T) = M, \end{cases}$$

and the generator depends on the second unknown variable $(L_{d_0+1}, \ldots, L_d)^*$ only in a linear way. His method consists of constructing an appropriate contraction mapping. Later, Peng [20] gave a nice treatment on the proof of existence and uniqueness for BSRDE (6) by using Bellman's principle of quasi linearization and a method of monotone convergence—a generalization of Wonham's approach to the random situation.

As early as 1978, Bismut [3] commented on page 220, "Nous ne pourrons pas démontrer l'existence de solution pour l'équation (2.49) dans le cas général." In English, it reads, "We could not prove the existence of solution for equation (2.49) for the general case." On page 238, he further pointed out that the essential difficulty for solution of the general BSRDE (2.49) lies in the fact that the integrand of the martingale term appears in the generator in a quadratic way. Note that Bismut [3] referred to the more general case: the concerned system is allowed to have jumps and the associated BSRDE is driven by a Martingale with possible jumps. BSRDE (1) is only a particular case of BSRDE (2.49) in [3]. However, BSRDE (1) possesses the difficult nature described by Bismut [3].

Two decades later in 1998, Peng [21] included the existence and uniqueness question for BSRDE (1) in his list of open problems on backward stochastic differential equations (BSDEs for short), which will be called the Bismut–Peng problem hereafter to be distinguished from Bismut's original problem (the latter is more general). Subsequently, there appear related discussions concerning the problem, for which the reader is referred to Chen, Li, and Zhou [4] and Chen and Yong [5].

Recently, the authors [14] solved the one-dimensional case of the Bismut–Peng problem with an approximation approach.

In this paper, we are concerned with the multidimensional case. We prove the global existence and uniqueness result for BSRDE (1) for the following class of multidimensional case:

$$d = 1, \quad B = C = 0.$$

That is, we solve the following BSRDE:

$$\begin{cases} dK = -\left[A^*K + KA + Q - LD(N + D^*KD)^{-1}D^*L\right] dt + L\, dw, \\ \qquad\qquad 0 \le t < T, \\ K(T) = M. \end{cases} \tag{7}$$

This BSRDE is special but typical, for the generator contains a quadratic term of L. This result is stated as Theorem 2.3.

Consider then the case where the control weight matrix N reduces to zero. Kohlmann and Zhou [16] discussed such a case under the following three assumptions: (a) all the coefficients involved are deterministic; (b) $C_1 = \cdots = C_d = 0, D_1 = \cdots = D_d = I_{m\times m}$, and $M = I$; (c) $A + A^* \geq BB^*$. Their arguments are based on applying a result of Chen, Li, and Zhou [4]. The authors [13] considered a general framework along those of an analogue of Bismut [3] and Peng [20], which has the following features: (a) the coefficients A, B, C, D, N, Q, M are allowed to be random, but are only $\{\mathcal{G}_t, 0 \leq t \leq T\}$-progressively measurable processes or $\mathcal{G}_T$-measurable random variable; (b) the assumptions in [16] are dispensed with or generalized; (c) the condition (5) is assumed to be satisfied. In [13], we generalized Bismut's previous result on existence and uniqueness of a solution of BSRDE (6) to the singular case under the following additional two assumptions:

$$M \geq \varepsilon I_{n\times n},\ \sum_{i=1}^{d} D_i^* D_i(t) \geq \varepsilon I_{m\times m} \text{ for some deterministic constant } \varepsilon > 0. \tag{8}$$

Later, the authors [14] proved the existence and uniqueness result for the one-dimensional singular case $N = 0$ under the assumption (8), but for a more general framework: the coefficients A, B, C, D, N, Q, M are allowed to be $\{\mathcal{F}_t, 0 \leq t \leq T\}$-progressively measurable processes or an $\mathcal{F}_T$-measurable random variable, and all the coefficients D_i, $i = 1, \ldots, d$, may be nonzero matrices.

In this paper we also obtain the global existence and uniqueness for the following multidimensional singular case:

$$\begin{gathered} d = 1, \quad m = n, \quad N = 0, \quad D^*D \geq \varepsilon I_{m\times m}, \\ M \geq \varepsilon I_{n\times n} \text{ for some deterministic constant } \varepsilon > 0. \end{gathered}$$

That is, we solve the following BSRDE:

$$\begin{cases} dK = -\,[A^*K + KA + C^*KC + Q + C^*L + LC \\ \qquad\quad -\,(KB + C^*KD + LD)(D^*KD)^{-1}(KB + C^*KD + LD)^*]\,dt + L\,dw, \\ \qquad\quad 0 \leq t < T, \\ K(T) = M. \end{cases} \tag{9}$$

This result is stated as Theorem 2.2.

BSRDE (1) arises from the solution of the optimal control problem

$$\inf_{u\in\mathcal{L}^2_{\mathcal{F}}(0,T;R^m)} J(u; 0, x), \tag{10}$$

where, for $t \in [0, T]$ and $x \in R^n$,

$$\begin{aligned} &J(u; t, x) \\ &:= E^{\mathcal{F}_t}\left[\int_t^T [\langle Nu, u\rangle + \langle QX^{t,x;u}, X^{t,x;u}\rangle]\,ds + \langle MX^{t,x;u}(T), X^{t,x;u}(T)\rangle\right] \end{aligned} \tag{11}$$

and $X^{t,x;u}(\cdot)$ solves the following stochastic differential equation:

$$\begin{cases} dX = (AX + Bu)\,ds + \sum_{i=1}^{d}(C_iX + D_iu)\,dw_i, \qquad t \leq s \leq T, \\ X(t) = x. \end{cases} \tag{12}$$

The following heuristic connection is well known: if BSRDE (1) has a solution (K, L), the solution for the above linear quadratic optimal control problem (LQ problem for short) has the following closed form (also called the feedback form):

$$u(t) = -\left(N + \sum_{i=1}^{d} D_i^* K D_i\right)^{-1} \left[B^* K + \sum_{i=1}^{d} D_i^* K C_i + \sum_{i=1}^{d} D_i^* L_i\right] X(t), \tag{13}$$

and the associated value function V has the following quadratic form:

$$V(t,x) := \operatorname*{essinf}_{u \in \mathcal{L}^2_{\mathcal{F}}(t,T;R^m)} J(u;t,x) = \langle K(t)x, x\rangle, \qquad 0 \le t \le T, x \in R^n. \tag{14}$$

In this way, on the one hand, solution of the above LQ problem is reduced to solving BSRDE (1). On the other hand, formula (14) actually provides a representation—of Feynman–Kac type—for the solution of BSRDE (1). The reader will see that this kind of representation plays an important role in the proofs given here for Theorems 2.1, 2.2, and 2.3.

The arguments given in this paper are completely new. They result from two observations. The first one is that in the following simple case,

$$\begin{aligned} &A = B = C = 0, \quad d = 1, m = n, \\ &D \text{ is nonsingular, and } D \text{ and } N \text{ are constant matrices,} \end{aligned} \tag{15}$$

the difficult quadratic term of L can be removed by doing some simple algebraic transformation, and the resulting BSRDE is globally solvable in view of the result of Bismut [3] and Peng [20]. As a consequence, the above simple case is globally solved. However, this case is too restricted. Then comes the second observation: by using some other tricks and by applying Theorem 2.1, some more general cases can be solved. Specifically, the following restrictions,

$$A = 0, \ m = n, \text{ and } D \text{ is nonsingular,} \tag{16}$$

are all removed, and the restrictive condition that

$$D \text{ and } N \text{ are constant matrices} \tag{17}$$

is improved. For the singular case, we have only the two restrictions $d = 1$ and $n = m$ remaining. Theorem 2.1 provides a way to obtain the solvability of more general BSRDEs from that of simple ones. We hope that the Bismut–Peng problem will be completely solved in the near future by using the above-mentioned methodology.

It is worth noting that backward stochastic differential equations were originally formulated in a linear form by Bismut [1] and then well studied in a Lipschitz nonlinear form by Pardoux and Peng [19].

It might be helpful to the reader to give the following remarks.

Yong and Zhou [32] give a good account of the stochastic LQ theory for the case of deterministic coefficients, with emphasis on the indefinite feature of the quadratic cost functional. Several recent papers are also devoted to the stochastic LQ problem, among which we cite Chen and Zhou [8] and Chen and Yong [5, 6, 7] for the reader's convenience. The papers [8, 6, 7] are mainly devoted to the study of the associated BSRDEs (or simply of the deterministic Riccati differential equations). Their existence and uniqueness results on BSRDEs could not cover ours. All of their global existence and uniqueness results are concerned with the case of deterministic coefficients, which exclude the quadratic growth in L of BSRDE—the main interesting and

difficult feature of this paper. The local existence and uniqueness results in [6] require either the condition $D_1 = D_2 = \cdots = D_d \equiv 0$—which also excludes the quadratic growth in L of BSRDE—or an additional regularity of the coefficients (that is, the conditions on the Malliavin derivatives of the coefficients), which is unnecessary in this paper even for the global existence and uniqueness assertions. Moreover, our approach is different from theirs.

The results of this paper were briefly reviewed in [15], which was presented at the Workshop on Mathematical Finance, held in Konstanz, Germany, on October 5–7, 2000.

When this second version was prepared in January and February, 2002, one year had passed since the submission of the original manuscript, and the Bismut–Peng problem has been closed in [28] by the second author by developing a different approach. A presentation more general than [28], incorporating the singular case, is available in Tang [29], which was reported at the International Conference on Mathematical Finance, held on May 10–13, 2001, at Fudan University, Shanghai, China. However, a complete solution of the Bismut–Peng problem in the spirit of either this paper or the authors' previous paper [14] is still interesting.

The rest of the paper is organized as follows. Section 2 contains a list of notation, two preliminary propositions, and the statement of the main results, which consist of Theorems 2.1–2.3. The proofs of these three theorems are given in sections 3–5, respectively. Finally, in section 6, application of Theorems 2.2 and 2.3 is given to the regular and singular stochastic LQ problems, both with and without constraint.

2. Preliminaries and the main results. Throughout this paper, we make the following assumptions. Let d and d_0 be two nonnegative integers with $d_0 \leq d$. Let $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t, 0 \leq t \leq T\})$ be a fixed complete probability space on which is defined a standard d-dimensional $\mathcal{F}_t$-adapted Brownian motion $w(t) \equiv (w_1(t), \ldots, w_d(t))^*$. Assume that $\{\mathcal{F}_t, 0 \leq t \leq T\}$ is the completion, by the totality $\mathcal{N}$ of all null sets of $\mathcal{F}$, of the natural filtration $\{\mathcal{F}_t^w, 0 \leq t \leq T\}$ generated by w. Denote by $\{\mathcal{G}_t, 0 \leq t \leq T\}$ the P-augmented natural filtration generated by the $(d - d_0)$-dimensional Brownian motion $(w_{d_0+1}, \ldots, w_d)$. Assume that all the coefficients A, B, C_i, and D_i are $\{\mathcal{F}_t, , 0 \leq t \leq T\}$-progressively measurable bounded matrix-valued processes, defined on $\Omega \times [0, T]$, of dimensions $n \times n$, $n \times m$, $n \times n$, and $n \times m$, respectively, with $i = 1, \ldots, d$. Also assume that M is an $\mathcal{F}_T$-measurable nonnegative bounded $n \times n$ random matrix, and Q and N are $\{\mathcal{F}_t, 0 \leq t \leq T\}$-progressively measurable, bounded, and nonnegative matrix processes of dimensions $n \times n$ and $m \times m$, respectively.

Notation. Throughout this paper, the following additional notation will be used:

M^*:	the transpose of any vector or matrix M;
$\lvert M\rvert$:	equals $\sqrt{\sum_{ij} m_{ij}^2}$ for any vector or matrix $M = (m_{ij})$;
$\langle M_1, M_2\rangle$:	the inner product of the two vectors M_1 and M_2;
R^n:	the n-dimensional Euclidean space;
R_+:	the set of all nonnegative real numbers;
$\mathcal{S}^n$:	the Euclidean space of all $n \times n$ symmetric matrices;
$\mathcal{S}_+^n$:	the set of all $n \times n$ nonnegative definite matrices;
$C([0, T]; H)$:	the Banach space of H-valued continuous functions on $[0, T]$, endowed with the maximum norm for a given Hilbert space H;
$\mathcal{L}_{\mathcal{F}}^2(0, T; H)$:	the Banach space of H-valued $\{\mathcal{F}_t, 0 \leq t \leq T\}$-adapted square -integrable stochastic processes f on $[0, T]$, endowed with the norm $(E\int_0^T \lvert f(t)\rvert^2\, dt)^{1/2}$ for a given Euclidean space H;

$\mathcal{L}^\infty_{\mathcal{F}}(0,T;H)$: the Banach space of H-valued, $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted, essentially bounded stochastic processes f on $[0,T]$, endowed with the norm $\operatorname{ess\,sup}_{t,\omega} |f(t)|$ for a given Euclidean space H;

$L^2(\Omega,\mathcal{F},P;H)$: the Banach space of H-valued norm-square-integrable random variables on the probability space $(\Omega,\mathcal{F},P)$ for a given Banach space H;

and $L^\infty(\Omega,\mathcal{F},P;C([0,T];\mathcal{S}^n))$ is the Banach space of $C([0,T];\mathcal{S}^n)$-valued, essentially maximum-norm-bounded random variables f on the probability space $(\Omega,\mathcal{F},P)$, endowed with the norm $\operatorname{ess\,sup}_{\omega\in\Omega} \max_{0\le t\le T} |f(t,\omega)|$. The definitions of $\mathcal{L}^\infty_{\mathcal{F}}(0,T;\mathcal{S}^n_+)$ and $L^\infty(\Omega,\mathcal{F},P;C([0,T];\mathcal{S}^n_+))$ are obvious.

PROPOSITION 2.1. *Assume that the coefficients A, B, C_i, and D_i are $\{\mathcal{G}_t, 0 \le t \le T\}$-progressively measurable bounded matrix-valued processes, defined on $\Omega \times [0,T]$, of dimensions $n \times n$, $n \times m$, $n \times n$, $n \times m$, respectively, with $i = 1, \dots, d$. Assume that M is a $\mathcal{G}_T$-measurable, nonnegative, and bounded $n \times n$ random matrix. Assume that Q and N are $\mathcal{G}_t$-progressively measurable, bounded, nonnegative and uniformly positive, $n \times n$ and $m \times m$ matrix processes, respectively. Then, BSRDE* (6) *has a unique $\{\mathcal{G}_t, 0 \le t \le T\}$-adapted global solution (K, L) with*

$$K \in \mathcal{L}^\infty_{\mathcal{G}}(0,T;\mathcal{S}^n_+) \cap L^\infty(\Omega,\mathcal{G}_T,P;C([0,T];\mathcal{S}^n_+)), \qquad L \in \mathcal{L}^2_{\mathcal{G}}(0,T;\mathcal{S}^n).$$

Proposition 2.1 and its proof can be found in Bismut [3] and Peng [20].

For $t \in [0,T]$ and $x \in R^n$, the following stochastic differential equation,

$$\begin{cases} dX = (AX + Bu)\,ds + \displaystyle\sum_{i=1}^{d} (C_i X + D_i u)\,dw_i, & t \le s \le T, \\ X(t) = x \end{cases} \tag{18}$$

has a unique solution (see Gal'chuk [10]) denoted by $X^{t,x;u}(\cdot)$. Define

$$J(u;t,x) := E^{\mathcal{F}_t}\left[\int_t^T [\langle Nu, u\rangle + \langle QX^{t,x;u}, X^{t,x;u}\rangle]\,ds + \langle MX^{t,x;u}(T), X^{t,x;u}(T)\rangle\right]. \tag{19}$$

Consider the optimal control problem

$$\inf_{u\in\mathcal{L}^2_{\mathcal{F}}(0,T;R^m)} J(u;0,x). \tag{20}$$

PROPOSITION 2.2. *Let (K, L) be an $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted solution of BSRDE* (1) *with*

$$K \in \mathcal{L}^\infty_{\mathcal{F}}(0,T;\mathcal{S}^n_+) \cap L^\infty(\Omega,\mathcal{F}_T,P;C([0,T];\mathcal{S}^n_+)), \qquad L \in \mathcal{L}^2_{\mathcal{F}}(0,T;\mathcal{S}^n),$$

and $N(t) + \sum_{i=1}^d D_i^* K D_i(t)$ being uniformly positive. Then, the value function $V(t,x)$ defined by

$$V(t,x) := \operatorname*{essinf}_{u\in\mathcal{L}^2_{\mathcal{F}}(t,T;R^m)} J(u;t,x) \qquad \forall (t,x) \in [0,T] \times R^n$$

has the following quadratic form:

$$V(t,x) = \langle K(t)x, x\rangle.$$

The detailed proof can be found in [13].

The main results of this paper consist of the following three theorems.

THEOREM 2.1. *We make the following six assumptions:*

(i) *For all* $\gamma \geq 0$, *the coefficients* $A^\gamma, B^\gamma, C_i^\gamma, D_i^\gamma, Q^\gamma$, *and* N^γ *are* $\{\mathcal{F}_t, 0 \leq t \leq T\}$-*progressively measurable matrix-valued processes, defined on* $\Omega \times [0,T]$, *of dimensions* $n \times n$, $n \times m$, $n \times n$, $n \times m$, $n \times n$, *and* $m \times m$, *respectively.*

(ii) M^γ *is an* $\mathcal{F}_T$-*measurable and nonnegative* $n \times n$ *random matrix.*

(iii) Q^γ *is a.s. a.e. nonnegative.*

(iv) *There are two deterministic positive constants* ε_1 *and* ε_2, *which are independent of the parameter* γ, *such that*

$$|A^\gamma(t)|, |B^\gamma(t)|, |C_i^\gamma(t)|, |D_i^\gamma(t)|, |Q^\gamma(t)|, |N^\gamma(t)|, |M^\gamma| \leq \varepsilon_1$$

and

$$N^\gamma \geq \varepsilon_2 I_{m\times m}.$$

(v) *As* $\gamma \to 0$, $A^\gamma(t), B^\gamma(t), C_i^\gamma(t), D_i^\gamma(t), Q^\gamma(t)$, *and* $N^\gamma(t)$ *converge uniformly in* (t,ω) *to* $A^0(t), B^0(t), C_i^0(t), D_i^0(t), Q^0(t)$, *and* $N^0(t)$, *respectively. Moreover,* M^γ *converges uniformly in* ω *to* M^0 *as* $\gamma \to 0$.

(vi) *For all* $\gamma > 0$ *the BSRDE* $(A^\gamma, B^\gamma; C_i^\gamma, D_i^\gamma, i = 1, \ldots, d; Q^\gamma, N^\gamma, M^\gamma)$ *has a unique* $\{\mathcal{F}_t, 0 \leq t \leq T\}$-*adapted global solution* (K^γ, L^γ) *with*

$$K^\gamma \in \mathcal{L}^\infty_{\mathcal{F}}(0,T;\mathcal{S}^n_+) \cap L^\infty(\Omega, \mathcal{F}_T, P; C([0,T];\mathcal{S}^n_+)), \qquad L^\gamma \in \mathcal{L}^2_{\mathcal{F}}(0,T;\mathcal{S}^n).$$

Then, there is a pair of processes (K,L) *with*

$$K \in \mathcal{L}^\infty_{\mathcal{F}}(0,T;\mathcal{S}^n_+) \cap L^\infty(\Omega, \mathcal{F}_T, P; C([0,T];\mathcal{S}^n_+)), \qquad L \in (\mathcal{L}^2_{\mathcal{F}}(0,T;\mathcal{S}^n))^d,$$

such that

$$\begin{aligned} &\lim_{\gamma\to 0} K^\gamma = K \quad \text{strongly in } \mathcal{L}^\infty_{\mathcal{F}}(0,T;\mathcal{S}^n_+) \cap L^\infty(\Omega, \mathcal{F}_T, P; C([0,T];\mathcal{S}^n_+)),\\ &\lim_{\gamma\to 0} L^\gamma = L \quad \text{strongly in } (\mathcal{L}^2_{\mathcal{F}}(0,T;\mathcal{S}^n))^d. \end{aligned} \tag{21}$$

Moreover, (K,L) *is a unique* $\{\mathcal{F}_t, 0 \leq t \leq T\}$-*adapted global solution of the BSRDE* $(A^0, B^0, C^0, D^0, Q^0, N^0, M^0)$.

If the above assumption of uniform convergence of $(A^\gamma, C^\gamma, Q^\gamma, M^\gamma)$ *is replaced with the following one:*

$$\lim_{\gamma\to 0} \operatorname*{esssup}_{\omega\in\Omega} \int_0^T (|A^\gamma - A^0| + |C^\gamma - C^0|^2 + |Q^\gamma - Q^0|)\, ds + |M^\gamma - M^0| \to 0, \tag{22}$$

then the above assertions still hold.

Remark 2.1. When the assumption of uniform positivity on the control weight matrix N is relaxed to nonnegativity, Theorem 2.1 still holds with the additional assumption that there is a deterministic positive constant ε_3, being independent of γ, such that

$$\sum_{i=1}^d (D_i^\gamma)^* D_i^\gamma \geq \varepsilon_3 I_{m\times m}, \quad M^\gamma \geq \varepsilon_3 I_{n\times n} \quad \forall \gamma > 0.$$

THEOREM 2.2 (the singular case). *Assume that $N \equiv 0$, $d = 1$, $n = m$, and $Q(t) \geq 0$. Also assume that there is a deterministic positive constant ε_3 such that*

$$M \geq \varepsilon_3 I_{n\times n} \tag{23}$$

and

$$D^*D \geq \varepsilon_3 I_{m\times m}. \tag{24}$$

Then, BSRDE (9) *has a unique $\{\mathcal{F}_t, 0 \leq t \leq T\}$-adapted global solution (K, L) with*

$$K \in \mathcal{L}^\infty_{\mathcal{F}}(0,T;\mathcal{S}^n_+) \cap L^\infty(\Omega,\mathcal{F}_T,P;C([0,T];\mathcal{S}^n_+)), \qquad L \in \mathcal{L}^2_{\mathcal{F}}(0,T;\mathcal{S}^n),$$

and $K(t,\omega)$ being uniformly positive with respect to (t,ω).

THEOREM 2.3 (the regular case). *Assume that $d = 1$, $M \geq 0$, $Q(t) \geq 0$, and $N(t) \geq \varepsilon_3 I_{m\times m}$ for some positive constant ε_3. Further assume that $B = C = 0$, and D and N satisfy the following:*

$$\begin{aligned} &\lim_{h\to 0+} \operatorname*{esssup}_{\omega\in\Omega} \max_{t_1,t_2\in[0,T];\,|t_1-t_2|\leq h} |D(t_1) - D(t_2)| = 0,\\ &\lim_{h\to 0+} \operatorname*{esssup}_{\omega\in\Omega} \max_{t_1,t_2\in[0,T];\,|t_1-t_2|\leq h} |N(t_1) - N(t_2)| = 0. \end{aligned} \tag{25}$$

Then, BSRDE (7) *has a unique $\{\mathcal{F}_t, 0 \leq t \leq T\}$-adapted global solution (K, L) with*

$$K \in \mathcal{L}^\infty_{\mathcal{F}}(0,T;\mathcal{S}^n_+) \cap L^\infty(\Omega,\mathcal{F}_T,P;C([0,T];\mathcal{S}^n_+)), \qquad L \in \mathcal{L}^2_{\mathcal{F}}(0,T;\mathcal{S}^n).$$

The proofs of the above three theorems are given in sections 3, 4, and 5, respectively.

3. The proof of Theorem 2.1. For all $(t, K, L) \in [0,T] \times \mathcal{S}^n_+ \times (\mathcal{S}^n)^d$, write

$$\begin{aligned} F^\gamma(t,K,L) := &-[KB^\gamma(t) + \sum_{i=1}^d C_i^\gamma(t)^* K D_i^\gamma(t) + \sum_{i=1}^d L_i D_i^\gamma(t)]\\ &\times [N^\gamma(t) + \sum_{i=1}^d D_i^\gamma(t)^* K D_i^\gamma(t)]^{-1}\\ &\times \left[KB^\gamma(t) + \sum_{i=1}^d C_i^\gamma(t)^* K D_i^\gamma(t) + \sum_{i=1}^d L_i D_i^\gamma(t)\right]^*. \end{aligned} \tag{26}$$

The generator of the BSRDE $(A^\gamma, B^\gamma; C_i^\gamma, D_i^\gamma, i = 1,\dots,d; Q^\gamma, N^\gamma, M^\gamma)$ is

$$\begin{aligned} G^\gamma(t,K,L) := &(A^\gamma)^*K + KA^\gamma + \sum_{i=1}^d (C_i^\gamma)^* K C_i^\gamma + Q^\gamma\\ &+ \sum_{i=1}^d ((C_i^\gamma)^* L_i + L_i C_i^\gamma) + F^\gamma(t,K,L). \end{aligned} \tag{27}$$

We have the following a priori estimates.

LEMMA 3.1. *Let the set of coefficients $(A^\gamma, B^\gamma; C_i^\gamma, D_i^\gamma, i = 1,\dots,d; Q^\gamma, N^\gamma, M^\gamma)$ satisfy the assumptions made in Theorem* 2.1. *Let (K^γ, L^γ) be a global $\{\mathcal{F}_t, 0 \leq t \leq T\}$-adapted solution to the BSRDE $(A^\gamma, B^\gamma; C_i^\gamma, D_i^\gamma, i = 1,\dots,d; Q^\gamma, N^\gamma, M^\gamma)$ with*

$$K^\gamma \in \mathcal{L}^\infty_{\mathcal{F}}(0,T;\mathcal{S}^n_+) \cap L^\infty(\Omega,\mathcal{F}_T,P;C([0,T];\mathcal{S}^n_+)) \quad \text{and} \quad L^\gamma \in (\mathcal{L}^2_{\mathcal{F}}(0,T;\mathcal{S}^n))^d.$$

Then, there is a deterministic positive constant ε_0, which is independent of γ, such that for all $\gamma > 0$, the following estimates hold:

$$0 \le K^\gamma(t) \le \varepsilon_0 I_{n\times n}, \quad E^{\mathcal{F}_t}\left(\int_t^T |L^\gamma|^2\,ds\right)^p \le \varepsilon_0 \quad \forall p \ge 1. \tag{28}$$

Proof of Lemma 3.1. Note that (K^γ, L^γ) satisfies the BSRDE:

$$\begin{cases} dK^\gamma = -\Big[(A^\gamma)^* K^\gamma + K^\gamma A^\gamma + \sum_{i=1}^d (C_i^\gamma)^* K^\gamma C_i^\gamma + Q^\gamma + \sum_{i=1}^d ((C_i^\gamma)^* L_i^\gamma + L_i^\gamma C_i^\gamma) \\ \qquad\qquad + F^\gamma(t, K^\gamma, L^\gamma)\Big]\,dt + \sum_{i=1}^d L_i^\gamma\,dw_i, \qquad 0 \le t < T, \\ K^\gamma(T) = M^\gamma. \end{cases} \tag{29}$$

Using Itô's formula, we get

$$\begin{cases} d|K^\gamma|^2 = -\Big[4\,\mathrm{tr}\,\big[(K^\gamma)^2 A^\gamma\big] + \sum_{i=1}^d 2\,\mathrm{tr}\,[K^\gamma (C_i^\gamma)^* K^\gamma C_i^\gamma] + 2\,\mathrm{tr}\,(K^\gamma Q^\gamma) \\ \qquad\qquad + \sum_{i=1}^d 4\,\mathrm{tr}\,(K^\gamma L_i^\gamma C_i^\gamma) + 2\,\mathrm{tr}\,[K^\gamma F^\gamma(t, K^\gamma, L^\gamma)] - |L^\gamma|^2\Big]\,dt \\ \qquad\qquad + \sum_{i=1}^d 2\,\mathrm{tr}\,(K^\gamma L_i^\gamma)\,dw_i, \qquad 0 \le t < T, \\ |K^\gamma|^2(T) = |M^\gamma|^2. \end{cases} \tag{30}$$

We observe that since

$$F^\gamma(t, K^\gamma, L^\gamma) \le 0, \qquad K^\gamma \ge 0,$$

we have

$$2\,\mathrm{tr}\,[K^\gamma F^\gamma(t, K^\gamma, L^\gamma)] = 2\,\mathrm{tr}\,\Big[(K^\gamma)^{\frac{1}{2}} F^\gamma(t, K^\gamma, L^\gamma)\,(K^\gamma)^{\frac{1}{2}}\Big] \le 0. \tag{31}$$

Hence,

$$\begin{aligned} |K^\gamma|^2(t) + \int_t^T |L^\gamma|^2\,ds \le |M^\gamma|^2 + \int_t^T \Big[4\,\mathrm{tr}\,\big[(K^\gamma)^2 A^\gamma\big] + \sum_{i=1}^d 2\,\mathrm{tr}\,[K^\gamma (C_i^\gamma)^* K^\gamma C_i^\gamma] \\ + 2\,\mathrm{tr}\,(K^\gamma Q^\gamma) + \sum_{i=1}^d 4\,\mathrm{tr}\,(K^\gamma L_i^\gamma C_i^\gamma)\Big]\,ds \\ - \int_t^T \sum_{i=1}^d 2\,\mathrm{tr}\,(K^\gamma L_i^\gamma)\,dw_i, \qquad 0 \le t < T. \end{aligned} \tag{32}$$

Using the elementary inequality

$$2ab \le a^2 + b^2$$

and taking the expectation on both sides with respect to $\mathcal{F}_r$ for $r \geq t$, we obtain that

(33)
$$E^{\mathcal{F}_r}|K^\gamma|^2(t) + \frac{1}{2}E^{\mathcal{F}_r}\int_t^T |L^\gamma|^2\,ds \leq \varepsilon_4 + \varepsilon_4 \int_t^T E^{\mathcal{F}_r}|K^\gamma|^2(s)\,ds, \qquad 0 \leq r \leq t < T.$$

Using Gronwall's inequality, we derive from the last inequality the first one of the estimates (28). In return, we derive from the second last inequality that

$$\int_t^T |L^\gamma|^2\,ds \leq \varepsilon_5 + \varepsilon_5 \int_0^T |L^\gamma|\,ds - \int_t^T \sum_{i=1}^d 2\,\mathrm{tr}\ (K^\gamma L_i^\gamma)\ dw_i. \tag{34}$$

Therefore,

$$\begin{aligned} &E^{\mathcal{F}_t}\left(\int_t^T |L^\gamma|^2\,ds\right)^p \\ &\leq 3^p\left[\varepsilon_5^p + \varepsilon_5^p E^{\mathcal{F}_t}\left(\int_t^T |L^\gamma|\,ds\right)^p + E^{\mathcal{F}_t}\left|\int_t^T \sum_{i=1}^d 2\mathrm{tr}(K^\gamma L_i^\gamma)\,dw_i\right|^p\right]. \end{aligned} \tag{35}$$

We have from the Burkhölder–Davis–Gundy inequality the following:

$$E^{\mathcal{F}_t}\left|\int_t^T \sum_{i=1}^d 2\,\mathrm{tr}\ (K^\gamma L_I^\gamma)\ dw_i\right|^p \leq 2^p E^{\mathcal{F}_t}\left|\int_t^T |K^\gamma|^2|L^\gamma|^2\,ds\right|^{p/2},$$

while from the Cauchy–Schwarz inequality, we have

$$E^{\mathcal{F}_t}\left(\int_t^T |L^\gamma|\,ds\right)^p \leq T^{p/2}E^{\mathcal{F}_t}\left(\int_t^T |L^\gamma|^2\,ds\right)^{p/2}.$$

Finally, we get

$$E^{\mathcal{F}_t}\left(\int_t^T |L^\gamma|^2\,ds\right)^p \leq 3^p\varepsilon_5^p + [3^pT^{p/2}\varepsilon_5^p + 6^p n^{p/2}\varepsilon_0^p]E^{\mathcal{F}_t}\left(\int_t^T |L^\gamma|^2\,ds\right)^{p/2}, \tag{36}$$

which implies the last estimate of the lemma.

Now, consider the optimal control problem,

$$\text{Problem } \mathcal{P}_\gamma\text{:} \qquad \inf_{u\in\mathcal{L}^2_{\mathcal{F}}(0,T;R^m)} J^\gamma(u;0,x), \tag{37}$$

where for $t \in [0,T]$ and $x \in R^n$,

(38)
$$J^\gamma(u;t,x) := E^{\mathcal{F}_t}\left\{\int_t^T [\langle N^\gamma u, u\rangle + \langle Q^\gamma X_\gamma^{t,x;u}, X_\gamma^{t,x;u}\rangle]\,ds + \langle M^\gamma X_\gamma^{t,x;u}(T), X_\gamma^{t,x;u}(T)\rangle\right\}$$

and $X_\gamma^{t,x;u}(\cdot)$ solves the following stochastic differential equation:

$$\begin{cases} dX = (A^\gamma X + B^\gamma u)\,ds + \sum_{i=1}^d (C_i^\gamma X + D_i^\gamma u)\,dw_i, & t \leq s \leq T, \\ X(t) = x. \end{cases} \tag{39}$$

The associated value function V^γ is defined as

$$V^\gamma(t,x) := \operatorname*{essinf}_{u\in\mathcal{L}^2_{\mathcal{F}}(t,T;R^m)} J^\gamma(u;t,x), \qquad (t,x)\in[0,T]\times R^n. \tag{40}$$

Then, from Proposition 2.2, we have

$$\langle K^\gamma(t)x,x\rangle = V^\gamma(t,x) \qquad \forall (t,x)\in[0,T]\times R^n.$$

From Lemma 3.1, we deduce

$$V^\gamma(t,x)\le \varepsilon_0|x|^2 \qquad \forall (t,x)\in[0,T]\times R^n.$$

So, the optimal control $\widehat{u}^\gamma$ for problem $\mathcal{P}_\gamma$ satisfies

$$\varepsilon_2 E^{\mathcal{F}_t}\int_t^T |\widehat{u}^\gamma|^2\,ds = E^{\mathcal{F}_t}\int_t^T \langle N^\gamma\widehat{u}^\gamma,\widehat{u}^\gamma\rangle\,ds \le \varepsilon_0|x|^2.$$

Set

$$\mathcal{U}^x_{ad}(t,T) := \left\{ u\in\mathcal{L}^2_{\mathcal{F}}(t,T;R^m) : \varepsilon_2 E^{\mathcal{F}_t}\int_t^T |u|^2\,ds \le \varepsilon_0|x|^2 \right\} \qquad \forall x\in R^n. \tag{41}$$

Then, we have

$$V^\gamma(t,x) := \operatorname*{essinf}_{u\in\mathcal{U}^x_{ad}(t,T)} J^\gamma(u;t,x). \tag{42}$$

Define

$$\begin{aligned} &K^{\gamma\tau} := K^\gamma - K^\tau, \quad L_i^{\gamma\tau} := L_i^\gamma - L_i^\tau, \quad X^{t,x;u}_{\gamma\tau} := X^{t,x;u}_\gamma - X^{t,x;u}_\tau,\\ &A^{\gamma\tau} := A^\gamma - A^\tau, \quad B^{\gamma\tau} := B^\gamma - B^\tau, \quad C_i^{\gamma\tau} := C_i^\gamma - C_i^\tau,\\ &D^{\gamma\tau} := D^\gamma - D^\tau, \quad Q^{\gamma\tau} := Q^\gamma - Q^\tau, \quad N^{\gamma\tau} := N^\gamma - N^\tau,\\ &M^{\gamma\tau} := M^\gamma - M^\tau. \end{aligned} \tag{43}$$

LEMMA 3.2. *Let the assumptions of Theorem* 2.1 *be satisfied. Then, there are three deterministic positive constants* ε_6, ε_7, *and* ε_8, *which are independent of the parameters* γ *and* τ, *such that the following three estimates hold:*

(i) *For each* $(t,x)\in[0,T]\times R^n$, *a.s.*

$$E^{\mathcal{F}_t}\max_{t\le s\le T}|X^{t,x;u}_\gamma(s)|^2 \le \varepsilon_6|x|^2 + \varepsilon_6 E^{\mathcal{F}_t}\int_t^T |u(s)|^2\,ds. \tag{44}$$

(ii) *For each* $(t,x)\in[0,T]\times R^n$, *a.s.*

$$\begin{aligned} E^{\mathcal{F}_t}\max_{t\le s\le T}|X^{t,x;u}_{\gamma\tau}(s)|^2 \le{}& \varepsilon_7 E^{\mathcal{F}_t}\int_t^T (|A^{\gamma\tau}|+|C^{\gamma\tau}|^2)|X^{t,x;u}_\gamma(s)|^2\,ds\\ &+ \varepsilon_7 E^{\mathcal{F}_t}\int_t^T (|B^{\gamma\tau}|+|D^{\gamma\tau}|^2)|u(s)|^2\,ds. \end{aligned} \tag{45}$$

(iii) *For each* $(t,x)\in[0,T]\times R^n$, *a.s.*

$$\begin{aligned} &|J^\gamma(u;t,x) - J^\tau(u;t,x)|\\ &\le \varepsilon_8 E^{\mathcal{F}_t}[|M^{\gamma\tau}||X^{t,x;u}_\gamma(T)|^2 + |X^{t,x;u}_{\gamma\tau}(T)|(|X^{t,x;u}_\gamma(T)| + |X^{t,x;u}_\tau(T)|)]\\ &\quad + \varepsilon_8 E^{\mathcal{F}_t}\int_t^T |X^{t,x;u}_{\gamma\tau}(s)|[|X^{t,x;u}_\gamma(s)| + |X^{t,x;u}_\tau(s)|]\,ds\\ &\quad + \varepsilon_8 E^{\mathcal{F}_t}\int_t^T |Q^{\gamma\tau}||X^{t,x;u}_\gamma(s)|^2\,ds + \varepsilon_8 E^{\mathcal{F}_t}\int_t^T |N^{\gamma\tau}||u(s)|^2\,ds. \end{aligned} \tag{46}$$

Proof of Lemma 3.2. Note that $X_{\gamma\tau}^{t,x;u}$ satisfies the following stochastic differential equation:

$$\begin{cases} dX_{\gamma\tau} = (A^\tau X_{\gamma\tau} + A^{\gamma\tau} X_\gamma + B^{\gamma\tau} u)\, ds + \sum_{i=1}^{d} (C_i^\tau X_{\gamma\tau} + C_i^{\gamma\tau} X_\gamma + D_i^{\gamma\tau} u)\, dw_i, \\ X_{\gamma\tau}(t) = 0. \end{cases}$$

So, in view of the assumptions of Theorem 2.1, the first two estimates are actually a consequence of the continuous dependence upon the parameters of the solution of a stochastic differential equation, and the proof is standard. The last estimate results from an immediate application of the mean-value formula for a differential function.

LEMMA 3.3. *Let the assumptions of Theorem* 2.1 *be satisfied. Then, we have the following three inequalities:*

(i) *For each* $(t,x) \in [0,T] \times R^n$, *for all* $u \in \mathcal{U}_{ad}^x(t,T)$,

$$E^{\mathcal{F}_t} \max_{t \le s \le T} |X_\gamma^{t,x;u}(s)|^2 \le \varepsilon_6 (1 + \varepsilon_2^{-1} \varepsilon_0) |x|^2. \tag{47}$$

(ii) *For each* $(t,x) \in [0,T] \times R^n$, *for all* $u \in \mathcal{U}_{ad}^x(t,T)$,

$$\begin{aligned} E^{\mathcal{F}_t} \max_{t \le s \le T} |X_{\gamma\tau}^{t,x;u}(s)|^2 \le{}& \varepsilon_7 \varepsilon_6 (1 + \varepsilon_2^{-1} \varepsilon_0) |x|^2 \operatorname*{esssup}_\omega \int_0^T (|A^{\gamma\tau}| + |C^{\gamma\tau}|^2)\, ds \\ &+ \varepsilon_7 \varepsilon_2^{-1} \varepsilon_0 |x|^2 \operatorname*{esssup}_{s,\omega} [|B^{\gamma\tau}(s)| + |D^{\gamma\tau}(s)|^2]. \end{aligned} \tag{48}$$

(iii) *For each* $(t,x) \in [0,T] \times R^n$, *for all* $u \in \mathcal{U}_{ad}^x(t,T)$,

$$\begin{aligned} &|J^\gamma(u;t,x) - J^\tau(u;t,x)| \\ &\le \varepsilon_8 \operatorname*{esssup}_\omega |M^{\gamma\tau}|\ E^{\mathcal{F}_t} |X_\gamma^{t,x;u}(T)|^2 \\ &\quad + \varepsilon_8 \left[E^{\mathcal{F}_t} |X_{\gamma\tau}^{t,x;u}(T)|^2\right]^{1/2} \left[E^{\mathcal{F}_t} (2|X_\gamma^{t,x;u}(T)|^2 + 2|X_\tau^{t,x;u}(T)|^2)\right]^{1/2} \\ &\quad + \varepsilon_8 T \left[E^{\mathcal{F}_t} \sup_{t \le s \le T} |X_{\gamma\tau}^{t,x;u}(s)|^2\right]^{1/2} \left[E^{\mathcal{F}_t} \sup_{t \le s \le T} [2|X_\gamma^{t,x;u}(s)|^2 + 2|X_\tau^{t,x;u}(s)|^2]\right]^{1/2} \\ &\quad + \varepsilon_8 \operatorname*{esssup}_\omega \int_0^T |Q^{\gamma\tau}|\, ds\ E^{\mathcal{F}_t} \sup_{t \le s \le T} |X_\gamma^{t,x;u}(s)|^2 + \varepsilon_8 \varepsilon_2^{-1} \varepsilon_0 |x|^2 \operatorname*{esssup}_{s,\omega} |N^{\gamma\tau}(s)|. \end{aligned} \tag{49}$$

Proof of Lemma 3.3. Since $u \in \mathcal{U}_{ad}^x(t,T)$, we have

$$E^{\mathcal{F}_t} \int_t^T |u|^2\, ds \le \varepsilon_2^{-1} \varepsilon_0 |x|^2. \tag{50}$$

Putting (50) into the first estimate of Lemma 3.2, we get the first inequality of Lemma 3.3. Putting (50) and the first inequality of Lemma 3.3 into the second estimate of Lemma 3.2, we get the second one. The last one is derived from (50) and applying the Cauchy–Schwarz inequality in the third estimate of Lemma 3.2.

Now we are in a position to prove Theorem 2.1.

Combining the first and the last inequalities of Lemma 3.3, we conclude that for each $(t,x)\in[0,T]\times R^n$, for all $u\in\mathcal{U}_{ad}^x(t,T)$,

$$
\begin{aligned}
&|J^\gamma(u;t,x)-J^\tau(u;t,x)|\\
&\le \varepsilon_8\varepsilon_6(1+\varepsilon_2^{-1}\varepsilon_0)|x|^2\operatorname*{esssup}_\omega|M^{\gamma\tau}|\\
&\quad+2|x|\varepsilon_8(T+1)\sqrt{\varepsilon_6(1+\varepsilon_2^{-1}\varepsilon_0)}\left[E^{\mathcal{F}_t}\sup_{t\le s\le T}|X_{\gamma\tau}^{t,x;u}(s)|^2\right]^{1/2}\\
&\quad+\varepsilon_8\varepsilon_6(1+\varepsilon_2^{-1}\varepsilon_0)|x|^2\operatorname*{esssup}_\omega\int_0^T|Q^{\gamma\tau}|\,ds+\varepsilon_8\varepsilon_2^{-1}\varepsilon_0|x|^2\operatorname*{esssup}_{s,\omega}|N^{\gamma\tau}(s)|.
\end{aligned}
\tag{51}
$$

Noting the second inequality of Lemma 3.3, we have

$$
\begin{aligned}
&|J^\gamma(u;t,x)-J^\tau(u;t,x)|\\
&\le \varepsilon_8\varepsilon_6(1+\varepsilon_2^{-1}\varepsilon_0)|x|^2\operatorname*{esssup}_\omega|M^{\gamma\tau}|+2|x|\varepsilon_8(T+1)\sqrt{\varepsilon_6(1+\varepsilon_2^{-1}\varepsilon_0)}\\
&\quad\times\left[\varepsilon_7\varepsilon_6(1+\varepsilon_2^{-1}\varepsilon_0)|x|^2\operatorname*{esssup}_\omega\int_0^T(|A^{\gamma\tau}|+|C^{\gamma\tau}|^2)\,ds\right.\\
&\quad\left.+\varepsilon_7\varepsilon_2^{-1}\varepsilon_0|x|^2\operatorname*{esssup}_{s,\omega}[|B^{\gamma\tau}(s)|+|D^{\gamma\tau}(s)|^2]\right]^{1/2}\\
&\quad+\varepsilon_8\varepsilon_6(1+\varepsilon_2^{-1}\varepsilon_0)|x|^2\operatorname*{esssup}_\omega\int_0^T|Q^{\gamma\tau}|\,ds+\varepsilon_8\varepsilon_2^{-1}\varepsilon_0|x|^2\operatorname*{esssup}_{s,\omega}|N^{\gamma\tau}(s)|
\end{aligned}
\tag{52}
$$

for each $(t,x)\in[0,T]\times R^n$, for all $u\in\mathcal{U}_{ad}^x(t,T)$. Therefore, we have

$$
\begin{aligned}
&|V^\gamma(t,x)-V^\tau(t,x)|\\
&\le \varepsilon_8\varepsilon_6(1+\varepsilon_2^{-1}\varepsilon_0)|x|^2\operatorname*{esssup}_\omega|M^{\gamma\tau}|+2|x|\varepsilon_8(T+1)\sqrt{\varepsilon_6(1+\varepsilon_2^{-1}\varepsilon_0)}\\
&\quad\times\left[\varepsilon_7\varepsilon_6(1+\varepsilon_2^{-1}\varepsilon_0)|x|^2\operatorname*{esssup}_\omega\int_0^T(|A^{\gamma\tau}|+|C^{\gamma\tau}|^2)\,ds\right.\\
&\quad\left.+\varepsilon_7\varepsilon_2^{-1}\varepsilon_0|x|^2\operatorname*{esssup}_{s,\omega}[|B^{\gamma\tau}(s)|+|D^{\gamma\tau}(s)|^2]\right]^{1/2}\\
&\quad+\varepsilon_8\varepsilon_6(1+\varepsilon_2^{-1}\varepsilon_0)|x|^2\operatorname*{esssup}_\omega\int_0^T|Q^{\gamma\tau}|\,ds+\varepsilon_8\varepsilon_2^{-1}\varepsilon_0|x|^2\operatorname*{esssup}_{s,\omega}|N^{\gamma\tau}(s)|
\end{aligned}
\tag{53}
$$

for each $(t,x)\in[0,T]\times R^n$.

In view of the assumptions of Theorem 2.1, (53) implies that for each $(t,x)\in[0,T]\times R^n$, $V^\gamma(t,x)$ converges to $V^0(t,x)$ as $\gamma\to0$. Moreover, this convergence is uniform in (t,ω). Hence, K^γ converges to some K^0 in the set

$$
\mathcal{L}_{\mathcal{F}}^\infty(0,T;\mathcal{S}_+^n)\cap L^\infty(\Omega,\mathcal{F}_T,P;C([0,T];\mathcal{S}_+^n)).
$$

In the following, we show the strong convergence of L^γ. Note that $(K^{\gamma\tau},L^{\gamma\tau})$ satisfies the BSDE

$$
\begin{cases}
dK^{\gamma\tau}(t)=-\left[G^\gamma(t,K^\gamma,L^\gamma)-G^\tau(t,K^\tau,L^\tau)\right]\,dt+\displaystyle\sum_{i=1}^d L_i^{\gamma\tau}\,dw_i,\\
K^{\gamma\tau}(T)=M^{\gamma\tau}.
\end{cases}
\tag{54}
$$

Using Itô's formula, we have

$$\begin{aligned} &E|K^{\gamma\tau}(t)|^2 + E\int_t^T |L^{\gamma\tau}(s)|^2\,ds \\ &= E|M^{\gamma\tau}|^2 + E\int_t^T \operatorname{tr}\,\{K^{\gamma\tau}[G^\gamma(s,K^\gamma,L^\gamma) - G^\tau(t,K^\tau,L^\tau)]\}\,ds. \end{aligned} \tag{55}$$

Since

$$|G^\gamma(s,K^\gamma,L^\gamma) - G^\tau(t,K^\tau,L^\tau)| \le \varepsilon(1+|L^\gamma|^2+|L^\tau|^2) \tag{56}$$

for some deterministic constant ε, which is independent of γ and τ, we have

$$E\int_t^T |L^{\gamma\tau}(s)|^2\,ds \le E|M^{\gamma\tau}|^2 + \varepsilon \operatorname*{esssup}_{s,\omega} |K^{\gamma\tau}(s)| E\int_t^T (1+|L^\gamma|^2+|L^\tau|^2)\,ds. \tag{57}$$

From the second a priori estimate of Lemma 3.1, we conclude that L^γ converges to some L^0 strongly in $\mathcal{L}^2_{\mathcal{F}}(0,T;\mathcal{S}^n)$. By passing to the limit in the BSRDE $(A^\gamma, B^\gamma; C_i^\gamma, D_i^\gamma, i=1,\dots,d; Q^\gamma, N^\gamma, M^\gamma)$, we show that (K^0,L^0) is an $\{\mathcal{F}_t, 0\le t\le T\}$-adapted global solution of the BSRDE $(A^0,B^0;C_i^0,D_i^0,i=1,\dots,d;Q^0,N^0,M^0)$.

4. The proof of Theorem 2.2. This section gives the proof of Theorem 2.2. The main idea is to do the matrix inverse transformation

$$\widetilde{K} := K^{-1}, \tag{58}$$

which turns out to satisfy a Riccati equation whose generator depends on the martingale term only in a linear way.

First, note that D has an inverse. We can rewrite the BSRDE (9) as

$$\begin{cases} dK = -[-\widetilde{A}^*K - K\widetilde{A} + Q - K\widetilde{B}K^{-1}\widetilde{B}^*K - LK^{-1}L \\ \qquad\quad + K\widetilde{B}K^{-1}L + LK^{-1}\widetilde{B}^*K]\,dt + L\,dw, \\ K(T) = M, \end{cases} \tag{59}$$

where

$$\widetilde{A} := -A + BD^{-1}C, \qquad \widetilde{B} := -BD^{-1}.$$

We have the following rules for the first-order and the second-order differentials of the inverse of a positive matrix as a matrix-valued function defined on the set of all $n\times n$ positive matrices:

$$\begin{aligned} d\left(K^{-1}\right) &= -K^{-1}(dK)K^{-1}, \\ d^2\left(K^{-1}\right) &= d\left(d\left(K^{-1}\right)\right) \\ &= -d\left(K^{-1}(dK)K^{-1}\right) \\ &= -d(K^{-1})(dK)K^{-1} - K^{-1}(dK)d(K^{-1}) \\ &= 2K^{-1}(dK)K^{-1}(dK)K^{-1}. \end{aligned} \tag{60}$$

Here, $d(K^{-1})$ is defined as the matrix whose (i,j)-component is the first-order differential of the (i,j)-component of K^{-1}, and $d^2(K^{-1})$ is defined as the matrix whose

(i,j)-component is the second-order differential of the (i,j)-component of K^{-1}. Using Itô's formula, we can write the equation for the inverse $\widetilde{K}$ of K:

$$\begin{cases} d\widetilde{K} = -[\widetilde{K}\widetilde{A}^* + \widetilde{A}\widetilde{K} - \widetilde{K}Q\widetilde{K} + \widetilde{B}\widetilde{K}\widetilde{B}^* + \widetilde{B}\widetilde{L} + \widetilde{L}\widetilde{B}^*]\,dt + \widetilde{L}\,dw, \\ \widetilde{K}(T) = M^{-1}, \end{cases} \tag{61}$$

where

$$\widetilde{L} := -K^{-1}LK^{-1}.$$

In what follows, we shall show how to construct an $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted global solution for BSRDE (9), starting from the preceding BSRDE $(\widetilde{A}, Q^{1/2}; \widetilde{B}, 0; 0, I_{m\times m}, M^{-1})$, that is, BSRDE (61).

From Proposition 2.1, the above BSRDE $(\widetilde{A}, Q^{1/2}; \widetilde{B}, 0; 0, I_{m\times m}, M^{-1})$ has a unique $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted global solution $(\widetilde{K}, \widetilde{L})$ with

$$\widetilde{K} \in \mathcal{L}^\infty_{\mathcal{F}}(0,T;\mathcal{S}^n_+) \cap L^\infty(\Omega, \mathcal{F}_T, P; C([0,T];\mathcal{S}^n_+)) \quad \text{and} \quad \widetilde{L} \in \mathcal{L}^2_{\mathcal{F}}(0,T;\mathcal{S}^n).$$

Proposition 2.1 does not assert that $\widetilde{K}$ is uniformly positive. However, from the fact that $\widetilde{K}(T) = M^{-1} \ge \varepsilon_1^{-1} I_{n\times n}$, we derive that $\widetilde{K}$ is uniformly positive. Define

$$K := \widetilde{K}^{-1} \quad \text{and} \quad L := -\widetilde{K}^{-1}\widetilde{L}\widetilde{K}^{-1}. \tag{62}$$

Then,

$$K \in \mathcal{L}^\infty_{\mathcal{F}}(0,T;\mathcal{S}^n_+) \cap L^\infty(\Omega, \mathcal{F}_T, P; C([0,T];\mathcal{S}^n_+)) \quad \text{and} \quad L \in \mathcal{L}^2_{\mathcal{F}}(0,T;\mathcal{S}^n). \tag{63}$$

Moreover, $K(t)$ is uniformly positive in (t,ω).

Again using Itô's formula, in view of BSRDE (61), we show that (K,L) is a global $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted solution of BSRDE (9).

The uniqueness is derived from Proposition 2.2. In fact, assume that $(\widehat{K}, \widehat{L})$ is another global $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted solution of BSRDE (9) with

$$\widehat{K} \in \mathcal{L}^\infty_{\mathcal{F}}(0,T;\mathcal{S}^n_+) \cap L^\infty(\Omega, \mathcal{F}_T, P; C([0,T];\mathcal{S}^n_+)) \quad \text{and} \quad \widehat{L} \in \mathcal{L}^2_{\mathcal{F}}(0,T;\mathcal{S}^n). \tag{64}$$

Then, from Proposition 2.2, we see that

$$\langle K(t)x, x\rangle = V(t,x) = \langle \widehat{K}(t)x, x\rangle, \quad \text{a.s.,} \quad \forall (t,x) \in [0,T] \times R^n.$$

Therefore we have $K(t) = \widehat{K}(t)$ almost surely for all $t \in [0,T]$. Set

$$\delta K := K - \widehat{K}, \quad \delta L_i := L_i - \widehat{L}_i, \quad \delta G := G(t,K,L) - G(t,\widehat{K},\widehat{L}).$$

On the one hand, we have $\delta K = 0$. On the other hand, the pair $(\delta K, \delta L)$ satisfies the following BSDE:

$$\begin{cases} d\delta K(t) = -\delta G\,dt + \displaystyle\sum_{i=1}^d \delta L_i(t)\,dw_i(t), \qquad 0 \le t < T, \\ \delta K(T) = 0. \end{cases} \tag{65}$$

Using Itô's formula, we have

$$E\int_t^T |\delta L(s)|^2\,ds \le E|\delta K(T)|^2 + \varepsilon \operatorname*{esssup}_{s,\omega} |\delta K(s)| E\int_t^T (1 + |L|^2 + |\widehat{L}|^2)\,ds = 0. \tag{66}$$

Here, ε is a positive constant. Hence, $\delta L = L - \widehat{L} = 0$.

The proof is complete.

5. The proof of Theorem 2.3. For the regular case, the situation is a little complex: we easily see that the above matrix inverse transformation on the first unknown matrix variable cannot eliminate the quadratic term of the second unknown variable. However, we can still solve some classes of BSRDEs through doing other appropriate matrix transformations.

PROPOSITION 5.1. *Assume that* $Q \geq A^*(D^{-1})^*ND^{-1} + (D^{-1})^*ND^{-1}A, \quad m = n$, *and* D *and* N *are nonsingular constant matrices. Then, Theorem* 2.3 *holds.*

Proof of Proposition 5.1. Write

$$\widehat{N} := (D^{-1})^*ND^{-1}. \tag{67}$$

Then BSRDE (7) is equivalent to the BSRDE $(A, 0; 0, I_{n\times n}; Q, \widehat{N}, M)$, i.e.,

$$\begin{cases} dK = -[A^*K + KA + Q - L(\widehat{N} + K)^{-1}L]\,dt + L\,dw, \\ \qquad\qquad 0 \leq t < T, \\ K(T) = M. \end{cases} \tag{68}$$

Define

$$\widehat{Q} := Q - A^*\widehat{N} - \widehat{N}A, \qquad \widehat{M} := \widehat{N} + M. \tag{69}$$

It is easy to see that $\widehat{Q}$ is nonnegative and $\widehat{M}$ is uniformly positive.

We also see that if (K, L) is a global $\{\mathcal{F}_t, 0 \leq t \leq T\}$-adapted solution of the BSRDE $(A, 0; 0, I_{n\times n}; Q, \widehat{N}, M)$ such that

$$K \in \mathcal{L}^\infty_{\mathcal{F}}(0, T; \mathcal{S}^n_+) \cap L^\infty(\Omega, \mathcal{F}_T, P; C([0, T]; \mathcal{S}^n_+)) \quad \text{and} \quad L \in \mathcal{L}^2_{\mathcal{F}}(0, T; \mathcal{S}^n), \tag{70}$$

then the pair $(\widehat{K}, \widehat{L})$, defined by $\widehat{K} := \widehat{N} + K$ and $\widehat{L} := L$, is a global $\{\mathcal{F}_t, 0 \leq t \leq T\}$-adapted solution of the BSRDE $(A, 0; 0, I_{n\times n}; \widehat{Q}, 0, \widehat{M})$, i.e.,

$$\begin{cases} d\widehat{K} = -[A^*\widehat{K} + \widehat{K}A + \widehat{Q} - \widehat{L}\widehat{K}^{-1}\widehat{L}]\,dt + \widehat{L}\,dw, \qquad 0 \leq t < T, \\ \widehat{K}(T) = \widehat{M}. \end{cases} \tag{71}$$

Moreover, $\widehat{K} \in \mathcal{L}^\infty_{\mathcal{F}}(0, T; \mathcal{S}^n_+) \cap L^\infty(\Omega, \mathcal{F}_T, P; C([0, T]; \mathcal{S}^n_+))$, $\widehat{L} \in \mathcal{L}^2_{\mathcal{F}}(0, T; \mathcal{S}^n)$, and $\widehat{K}$ is uniformly positive.

Inversely, if $(\widehat{K}, \widehat{L})$ is a global $\{\mathcal{F}_t, 0 \leq t \leq T\}$-adapted solution of BSRDE (71) such that $\widehat{K} \in \mathcal{L}^\infty_{\mathcal{F}}(0, T; \mathcal{S}^n_+) \cap L^\infty(\Omega, \mathcal{F}_T, P; C([0, T]; \mathcal{S}^n_+))$, $\widehat{L} \in \mathcal{L}^2_{\mathcal{F}}(0, T; \mathcal{S}^n)$, and $\widehat{K}$ is uniformly positive, then the pair (K, L), defined by $K := \widehat{K} - \widehat{N}$ and $L := \widehat{L}$, is a global $\{\mathcal{F}_t, 0 \leq t \leq T\}$-adapted solution of the BSRDE $(A, 0; 0, I_{n\times n}; Q, \widehat{N}, M)$, satisfying the following:

$$K \in \mathcal{L}^\infty_{\mathcal{F}}(0, T; \mathcal{S}^n) \cap L^\infty(\Omega, \mathcal{F}_T, P; C([0, T]; \mathcal{S}^n)) \quad \text{and} \quad L \in \mathcal{L}^2_{\mathcal{F}}(0, T; \mathcal{S}^n). \tag{72}$$

At the moment, it is not clear that K is nonnegative. However, it is clear that $\widehat{N} + K$ is uniformly positive. Making use of the nonnegativity of Q, M, and $\widehat{N}$, we can deduce that K is nonnegative. In fact, according to our previous paper [13], since $(\widehat{K}, \widehat{L})$ is a global $\{\mathcal{F}_t, 0 \leq t \leq T\}$-adapted solution of BSRDE (71) such that $\widehat{K} \in \mathcal{L}^\infty_{\mathcal{F}}(0, T; \mathcal{S}^n_+) \cap L^\infty(\Omega, \mathcal{F}_T, P; C([0, T]; \mathcal{S}^n_+))$, $\widehat{L} \in \mathcal{L}^2_{\mathcal{F}}(0, T; \mathcal{S}^n)$, and $\widehat{K}$ is uniformly positive, the following closed system (it is a homogeneous matrix-valued stochastic differential equation with possibly unbounded coefficients!),

$$\begin{cases} dU(s) = A(s)U(s)\,ds - [\widehat{K}(s)]^{-1}\widehat{L}(s)U(s)\,dw(s), \qquad s \in (t, T] \\ \qquad\quad = A(s)U(s)\,ds - [\widehat{N}(s) + K(s)]^{-1}L(s)U(s)\,dw(s), \qquad s \in (t, T], \\ U(t) = I_{n\times n}, \end{cases} \tag{73}$$

has a solution $\{U(s,t), t \le s \le T\}$ for each $t \in [0,T]$, satisfying the following:

$$E \max_{t \le s \le T} |U(s,t)|^2 < \infty, \qquad E \int_t^T |L(s)U(s,t)|^2 \, ds < \infty. \tag{74}$$

Therefore, we can use Itô's formula to derive from BSRDE (68) the representation

$$K(t) = U(t,t)^* K(t) U(t,t) = E^{\mathcal{F}_t} \Bigg\{ U(T,t)^* M U(T,t) + \int_t^T U(s,t)^* [Q + L(s)(\widehat{N}+K)^{-1} \widehat{N} (\widehat{N}+K)^{-1} L(s)] U(s,t) \, ds \Bigg\} \qquad \forall\, t \in [0,T]. \tag{75}$$

This formula implies that $K(t)$ is nonnegative for each $t \in [0,T]$.

While from Theorem 2.2, we see that BSRDE (71) has a unique global $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted solution $(\widehat{K}, \widehat{L})$ such that $\widehat{K} \in \mathcal{L}^\infty_{\mathcal{F}}(0,T;\mathcal{S}^n_+) \cap L^\infty(\Omega, \mathcal{F}_T, P; C([0,T];\mathcal{S}^n_+))$, $\widehat{L} \in \mathcal{L}^2_{\mathcal{F}}(0,T;\mathcal{S}^n)$, and $\widehat{K}$ is uniformly positive. Therefore $(\widehat{K} - \widehat{N}, \widehat{L})$ is a global $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted solution of BSRDE (7), satisfying the following:

$$\widehat{K} - \widehat{N} \in \mathcal{L}^\infty_{\mathcal{F}}(0,T;\mathcal{S}^n_+) \cap L^\infty(\Omega, \mathcal{F}_T, P; C([0,T];\mathcal{S}^n_+)) \quad \text{and} \quad \widehat{L} \in \mathcal{L}^2_{\mathcal{F}}(0,T;\mathcal{S}^n). \tag{76}$$

The proof is then complete.

PROPOSITION 5.2. *Assume that $A = 0$ and D and N are constant matrices. Then, Theorem* 2.3 *holds.*

Proof of Proposition 5.2. First assume $m = n$. Consider the following approximating BSRDEs:

$$\begin{cases} dK = -[Q - LD_\gamma (N + D_\gamma^* K D_\gamma)^{-1} D_\gamma^* L] \, dt + L \, dw, \\ K(T) = M, \end{cases} \tag{77}$$

where

$$D_\gamma := D + \gamma I_{m \times m}, \gamma > 0.$$

As γ is sufficiently small, D_γ is nonsingular. From Proposition 5.1, we see that BSRDE (77) has a unique global $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted solution (K_γ, L_γ) for sufficiently small $\gamma > 0$, such that

$$K_\gamma \in \mathcal{L}^\infty_{\mathcal{F}}(0,T;\mathcal{S}^n_+) \cap L^\infty(\Omega, \mathcal{F}_T, P; C([0,T];\mathcal{S}^n_+)) \quad \text{and} \quad L_\gamma \in \mathcal{L}^2_{\mathcal{F}}(0,T;\mathcal{S}^n). \tag{78}$$

From Theorem 2.1, we see that as γ tends to zero, K_γ uniformly converges to some $K \in \mathcal{L}^\infty_{\mathcal{F}}(0,T;\mathcal{S}^n_+) \cap L^\infty(\Omega, \mathcal{F}_T, P; C([0,T];\mathcal{S}^n_+))$ and L_γ strongly converges to some $L \in \mathcal{L}^2_{\mathcal{F}}(0,T;\mathcal{S}^n)$, and that (K,L) is an $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted solution of the BSRDE (7) with $A = 0$.

Consider the case $n > m$. Then consider the $n \times n$ matrices $\widetilde{D}$, whose first m columns are D and whose last $(n-m)$ columns are zero column vectors, and $\widetilde{N}$, which is defined as

$$\widetilde{N} := \begin{pmatrix} R & 0 \\ 0 & I \end{pmatrix}.$$

BSRDE (7) with $A = 0$ is rewritten as

$$\begin{cases} dK = -[Q - L\widetilde{D}(\widetilde{N} + \widetilde{D}^* K \widetilde{D})^{-1} \widetilde{D}^* L]\, dt + L\, dw, \\ K(T) = M. \end{cases}$$

From the preceding result, we obtain the desired existence result.

Consider the case $n < m$. Then, there is an $m \times m$ orthogonal transformation matrix T such that

$$D = [\widehat{D}, 0]T, \qquad \widehat{D} \in R^{n \times n} \text{ and is nonsingular.}$$

Write

$$\widetilde{N} := (T^{-1})^* N T^{-1} := \begin{pmatrix} \widehat{N}_{11} & \widehat{N}_{12} \\ \widehat{N}_{12}^* & \widehat{N}_{22} \end{pmatrix} > 0.$$

Then, $\widehat{N}_{11} > 0$. BSRDE (7), when $A = 0$, is rewritten as

$$\begin{cases} dK = -[Q - L\widehat{D}(\widetilde{N}_{11} + \widehat{D}^* K \widehat{D})^{-1} \widehat{D}^* L]\, dt + L\, dw, \\ K(T) = M. \end{cases}$$

From the preceding result, we obtain the desired existence result.

PROPOSITION 5.3. *Assume that $A = 0$ and D and N are piecewisely constant $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted bounded matrix processes. Then, Theorem* 2.3 *holds.*

Proof of Proposition 5.3. Since D and N are piecewisely constant $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted bounded matrix processes, there is a finite partition,

$$0 =: t_0 < t_1 < \cdots < t_J := T,$$

such that on each interval $[t_i, t_{i+1}] \subset [0, T]$, D and N are constant $\mathcal{F}_{t_i}$-measurable bounded random matrices. From Proposition 5.2, the BSRDE

$$\begin{cases} dK = -[Q - LD(N + D^* K D)^{-1} D^* L]\, dt + L\, dw, \\ \qquad\qquad t_{J-1} \le t < T, \\ K(T) = M \end{cases} \tag{79}$$

has a unique $\{\mathcal{F}_t, t_{J-1} \le t \le T\}$-adapted solution (K^J, L^J) with

$$K^J \in \mathcal{L}^\infty_{\mathcal{F}}(t_{J-1}, T; \mathcal{S}^n_+) \cap L^\infty(\Omega, \mathcal{F}_T, P; C([t_{J-1}, T]; \mathcal{S}^n_+)), \qquad L^J \in \mathcal{L}^2_{\mathcal{F}}(t_{J-1}, T; \mathcal{S}^n).$$

Assume that for some $i = 2, \ldots, J$, the BSRDE

$$\begin{cases} dK = -[Q - LD(N + D^* K D)^{-1} D^* L]\, dt + L\, dw, \\ \qquad\qquad t_{i-1} \le t < t_i, \\ K(t_i) = K^{i+1}(t_i) \end{cases} \tag{80}$$

has a unique $\{\mathcal{F}_t, t_{i-1} \le t \le t_i\}$-adapted solution (K^i, L^i) with

$$K^i \in \mathcal{L}^\infty_{\mathcal{F}}(t_{i-1}, t_i; \mathcal{S}^n_+) \cap L^\infty(\Omega, \mathcal{F}_{t_i}, P; C([t_{i-1}, t_i]; \mathcal{S}^n_+)), \qquad L^i \in \mathcal{L}^2_{\mathcal{F}}(t_{i-1}, t_i; \mathcal{S}^n).$$

Note that when $i = J$, we use the convention $K^{J+1}(t_J) := M$. Then, we conclude from Proposition 5.2 that the BSRDE

$$\begin{cases} dK = -[Q - LD(N + D^* K D)^{-1} D^* L]\, dt + L\, dw, \\ \qquad\qquad t_{i-2} \le t < t_{i-1}, \\ K(t_{i-1}) = K^i(t_{i-1}) \end{cases} \tag{81}$$

has a unique $\{\mathcal{F}_t, t_{i-2} \le t \le t_{i-1}\}$-adapted solution (K^{i-1}, L^{i-1}) with

$$K^{i-1} \in \mathcal{L}^{\infty}_{\mathcal{F}}(t_{i-2}, t_{i-1}; \mathcal{S}^n_+) \cap L^{\infty}(\Omega, \mathcal{F}_{t_{i-1}}, P; C([t_{i-2}, t_{i-1}]; \mathcal{S}^n_+)),$$
$$L^{i-1} \in \mathcal{L}^{2}_{\mathcal{F}}(t_{i-2}, t_{i-1}; \mathcal{S}^n).$$

In this backward inductive way, we may define J pairs of processes $\{(K^i, L^i)\}_{i=1}^J$. Define on the whole time interval $[0, T]$ the pair of $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted processes (K, L) as follows:

$$K(t) := \sum_{i=1}^{J} K^i(t)\chi_{[t_{i-1}, t_i)}(t), \qquad L(t) := \sum_{i=1}^{J} L^i(t)\chi_{[t_{i-1}, t_i)}(t).$$

We see that (K, L) solves BSRDE (7). We then obtain the desired existence result.

PROPOSITION 5.4. *Assume that $A = 0$. Then, Theorem 2.3 holds.*

Proof of Proposition 5.4. For an arbitrary positive integer k, consider the 2^k-partition of the time interval. Define

$$D^k(t) = D\left(\frac{i-1}{2^k}T\right) \qquad \forall t \in \left[\frac{i-1}{2^k}T, \frac{i}{2^k}T\right),\ i = 1, 2, \ldots, 2^k,$$

and

$$N^k(t) = N\left(\frac{i-1}{2^k}T\right) \qquad \forall t \in \left[\frac{i-1}{2^k}T, \frac{i}{2^k}T\right),\ i = 1, 2, \ldots, 2^k.$$

For each k, D^k and N^k are piecewisely constant, $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted, bounded matrix processes. Further, in view of (25), $D^k(t)$ and $N^k(t)$ converge respectively to D and N, uniformly in (t, ω). That is, we have

$$\lim_{k\to\infty} \operatorname*{esssup}_{\omega\in\Omega} \max_{t\in[0,T]} |D^k(t) - D(t)| = 0, \qquad \lim_{k\to\infty} \operatorname*{esssup}_{\omega\in\Omega} \max_{t\in[0,T]} |N^k(t) - N(t)| = 0.$$

From Proposition 5.3, we see that the BSRDE $(0, 0, 0, D^k; Q, N^k; M)$ has a global $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted solution (K^k, L^k), and then from Theorem 2.1, we see that Theorem 2.3 holds.

Proof of Theorem 2.3. The case $A = 0$ is solved by Proposition 5.4. For the case $A \neq 0$, consider the following transformation:

$$\widetilde{K} := \Phi^* K \Phi, \qquad \widetilde{L} := \Phi^* L \Phi,$$

where Φ is the solution of the differential matrix equation

$$\begin{cases} \dfrac{d\Phi}{dt}(t) = A(t)\Phi(t), \quad t \in (0, T], \\ \Phi(0) = I_{n\times n}. \end{cases}$$

Using Itô's formula, we get the BSDE for $(\widetilde{K}, \widetilde{L})$:

$$\begin{cases} d\widetilde{K}(t) = -[\widetilde{Q} - \widetilde{L}\widetilde{D}(N + \widetilde{D}^*\widetilde{K}\widetilde{D})^{-1}\widetilde{D}\widetilde{L}]\,dt + \widetilde{L}\,dw(t), \qquad t \in (0, T], \\ \widetilde{K}(T) = \widetilde{M}, \end{cases}$$

where $\widetilde{Q} := \Phi^* Q \Phi$, $\widetilde{M} := \Phi(T)^* M \Phi(T)$, $\widetilde{D} := \Phi^{-1} D$. Note that the trajectories of $\widetilde{D}$ are still uniformly continuous like D. From Proposition 5.4, we see that the BSRDE $(0, 0, 0, \widetilde{D}; \widetilde{Q}, N, \widetilde{M})$ has a global adapted solution $(\widetilde{K}, \widetilde{L})$, and thus the pair

$$((\Phi^*)^{-1}\widetilde{K}\Phi^{-1}, (\Phi^*)^{-1}\widetilde{L}\Phi^{-1})$$

solves the original BSRDE $(A, 0, 0, D; Q, N, M)$.

The uniqueness can be proved in the same way as in the proof of Theorem 2.2.

6. Application to stochastic LQ problems.

6.1. The unconstrained stochastic LQ problem.

Assume that

$$\xi \in L^2(\Omega, \mathcal{F}_T, P; R^n), \qquad q, f, g_i \in \mathcal{L}^2_{\mathcal{F}}(0, T; R^n). \tag{82}$$

Consider the following optimal control problem (denoted by $\mathcal{P}_0$):

$$\min_{u \in \mathcal{L}^2_{\mathcal{F}}(0,T;R^m)} J(u; 0, x), \tag{83}$$

with

$$\begin{aligned} J(u; t, x) = {} & E^{\mathcal{F}_t} \langle M(X^{t,x;u}(T) - \xi), X^{t,x;u}(T) - \xi \rangle \\ & + E^{\mathcal{F}_t} \int_t^T [\langle Q(X^{t,x;u} - q), X^{t,x;u} - q \rangle + \langle Nu, u \rangle] \, ds \end{aligned} \tag{84}$$

and $X^{t,x;u}$ being the solution of the stochastic differential equation

$$\begin{cases} dX = (AX + Bu + f) \, ds + \sum_{i=1}^d (C_i X + D_i u + g_i) \, dw_i, & t < s \le T, \\ X(t) = x, \qquad u \in \mathcal{L}^2_{\mathcal{F}}(t, T; R^m). \end{cases} \tag{85}$$

The value function V is defined as

$$V(t, x) := \operatorname*{essinf}_{u \in \mathcal{L}^2_{\mathcal{F}}(t,T;R^m)} J(u; t, x), \qquad (t, x) \in [0, T] \times R^n. \tag{86}$$

Define $\Gamma : [0, T] \times \mathcal{S}^n_+ \times R^{n \times d} \to R^{m \times n}$ by

$$\Gamma(\cdot, S, L) = -\left(N + \sum_{i=1}^d D_i^* S D_i \right)^{-1} \left(B^* S + \sum_{i=1}^d D_i^* S C_i + \sum_{i=1}^d D_i^* L_i \right) \tag{87}$$

and

$$\widehat{A} := A + B\Gamma(\cdot, K, L), \quad \widehat{C}_i := C_i + D_i \Gamma(\cdot, K, L), \quad i = 1, \dots, d. \tag{88}$$

We have the following theorem.

THEOREM 6.1. *Suppose that the assumptions of Theorem* 2.2 *or Theorem* 2.3 *are satisfied. Let* (K, L) *be the unique* $\{\mathcal{F}_t, 0 \le t \le T\}$*-adapted solution of BSRDE* (1) *such that* $K \in \mathcal{L}^\infty_{\mathcal{F}}(0, T; \mathcal{S}^n_+) \cap L^\infty(\Omega, \mathcal{F}_T, P; C([0, T]; \mathcal{S}^n_+))$ *and* $L \in \mathcal{L}^2_{\mathcal{F}}(0, T; \mathcal{S}^n)$*. Let* (ψ, ϕ) *be the* $\{\mathcal{F}_t, 0 \le t \le T\}$*-adapted solution of the BSDE*

$$\begin{cases} d\psi(t) = -\left[\widehat{A}^* \psi + \sum_{i=1}^d \widehat{C}_i^* (\phi_i - K g_i) - K f - \sum_{i=1}^d L_i g_i + Q q \right] dt + \sum_{i=1}^d \phi_i \, dw_i, \\ \psi(T) = M\xi, \quad \phi := (\phi_1, \dots, \phi_d) \end{cases} \tag{89}$$

such that $(\psi, \phi) \in (\mathcal{L}^2_{\mathcal{F}}(0, T; R^n))^{(n+1)}$.

Then, the optimal control $\widehat{u}$ for the nonhomogeneous stochastic LQ problem $\mathcal{P}_0$ exists uniquely and has the following feedback law:

$$\begin{aligned}\widehat{u} = -\left(N+\sum_{i=1}^{d} D_i^* K D_i\right)^{-1}\Bigg[\left(B^*K+\sum_{i=1}^{d} D_i^* K C_i+\sum_{i=1}^{d} D_i^* L_i\right)\widehat{X}\\ -B^*\psi+\sum_{i=1}^{d} D_i^*(Kg_i-\phi_i)\Bigg].\end{aligned}\tag{90}$$

The value function $V(t,x),(t,x)\in[0,T]\times R^n$, has the following explicit formula:

$$V(t,x)=\langle K(t)x,x\rangle-2\langle\psi(t),x\rangle+V^0(t),\qquad (t,x)\in[0,T]\times R^n,\tag{91}$$

with

$$\begin{aligned}V^0(t) :=\ & E^{\mathcal{F}_t}\langle M\xi,\xi\rangle+E^{\mathcal{F}_t}\int_t^T\langle Qq,q\rangle\,ds-2E^{\mathcal{F}_t}\int_t^T\langle\psi,f\rangle\,ds\\ &+E^{\mathcal{F}_t}\int_t^T\sum_{i=1}^{d}[\langle Kg_i,g_i\rangle-2\langle\phi_i,g_i\rangle]\,ds\\ &-E^{\mathcal{F}_t}\int_t^T\left\langle\left(N+\sum_{i=1}^{d} D_i^* K D_i\right)u^0,u^0\right\rangle ds\end{aligned}\tag{92}$$

and

$$u^0:=\left(N+\sum_{i=1}^{d} D_i^* K D_i\right)^{-1}\left[B^*\psi+\sum_{i=1}^{d} D_i^*(\phi_i-Kg_i)\right],\qquad t\le s\le T.\tag{93}$$

The reader is referred to our previous paper [13] for a detailed proof.

6.2. The constrained stochastic LQ problem. Fix $x_T\in R^n$. Define

$$U_{\rm ad}(t,x):=\{u\in\mathcal{L}^2_{\mathcal{F}}(t,T;R^m):EX^{t,x;u}(T)=x_T\}\quad\forall(t,x)\in[0,T]\times R^n,\tag{94}$$

where $X^{t,x;u}$ is the solution of stochastic differential equation (85). Then, consider the following constrained LQ problem (denoted by $\mathcal{P}_c^{t,x}$):

$$\inf_{u\in U_{\rm ad}(0,x)} J(u;0,x),\tag{95}$$

where the cost functional $J(u;t,x)$ is defined by (84). Note that the set of admissible controls $U_{\rm ad}(t,x)$ contains the terminal expected constraint.

Let $\Psi(\cdot,t)$ be the unique solution of the stochastic differential equation:

$$\begin{cases}dY_s=A(s)Y_s\,ds+\sum_{i=1}^{d} C_i(s)Y_s\,dw_i(s),\qquad t\le s\le T,\\ Y_t=I_{n\times n}.\end{cases}\tag{96}$$

To guarantee that $U_{\rm ad}(t,x)$ is not empty, assume that the matrix

$$\Delta(t):=E\int_t^T E^{\mathcal{F}_s}\Psi(T,s)B(s)B^*(s)E^{\mathcal{F}_s}\Psi^*(T,s)\,ds\tag{97}$$

is nonsingular. Then, for all $x \in R^n$, the following control,

$$u(s) := B^*(s)E^{\mathcal{F}_s}\Psi^*(T,s)\Delta(t)^{-1}\left[x_T - E\int_t^T \Psi(T,s)f(s)\,ds\right], \qquad s \in (t,T], \tag{98}$$

belongs to $U_{\mathrm{ad}}(t,x)$.

We have the following existence result.

THEOREM 6.2. *Let the assumptions of Theorem* 2.2 *or Theorem* 2.3 *be satisfied. Assume that* $U_{\mathrm{ad}}(0,x)$ *is not empty. Then, the problem* $\mathcal{P}_c^{0,x}$ *has a unique optimal control.*

Proof of Theorem 6.2. The main idea is to choose a sequence $\{u^k; k = 1,2,\ldots\}$ such that

$$u^k \in U_{\mathrm{ad}}(0,x), \qquad \lim_{k\to\infty} J(u^k;0,x) = \inf_{u\in U_{\mathrm{ad}}(0,x)} J(u;0,x).$$

Define $X^k := x^{0,x;u^k}$. Note that $U_{\mathrm{ad}}(0,x)$ is close and convex. Therefore, $\frac{1}{2}(u^k+u^l) \in U_{\mathrm{ad}}(0,x)$. The LQ structure implies the following equality:

$$\begin{aligned} &2E\left\langle M\left(\frac{X^k(T)-X^l(T)}{2}\right), \frac{X^k(T)-X^l(T)}{2}\right\rangle \\ &+2E\int_0^T\left[\left\langle Q\left(\frac{X^k-X^l}{2}\right), \frac{X^k-X^l}{2}\right\rangle + \left\langle N\left(\frac{u^k-u^l}{2}\right), \frac{u^k-u^l}{2}\right\rangle\right] ds \\ &= J(u^k;0,x) + J(u^l;0,x) - 2J\left(\frac{u^k+u^l}{2};0,x\right). \end{aligned} \tag{99}$$

Then, we have for a positive constant ε,

$$E\int_0^T |u^k-u^l|^2\,ds \le \varepsilon\left[J(u^k;0,x) + J(u^l;0,x) - 2J\left(\frac{u^k+u^l}{2};0,x\right)\right], \tag{100}$$

which implies that $E\int_0^T |u^k-u^l|^2\,ds \to 0$ as $k,l\to\infty$. (100) is obvious for the regular case. For the singular case, we deduce it using the estimate

$$E\int_0^T |u^k-u^l|^2\,ds \le \varepsilon E|X^k(T)-X^l(T)|^2 \quad \text{for a positive constant } \varepsilon,$$

which is derived from [13, Lemma 2.2].

Hence the sequence $\{u^k, k = 1,2,\ldots\}$ is a Cauchy sequence, and it has a limit $u \in U_{\mathrm{ad}}(0,x)$. Then u is an optimal control.

It remains to show the uniqueness. Let u^1 and u^2 be both optimal. Then similar to that above, we have

$$E\int_0^T |u^1-u^2|^2\,ds \le \varepsilon\left[J(u^1;0,x) + J(u^2;0,x) - 2J\left(\frac{u^1+u^2}{2};0,x\right)\right] \le 0. \tag{101}$$

So, $u^1 = u^2$. The proof is complete.

Due to the limitation of space, we will, in what follows, just sketch how to solve the unique optimal control of Theorem 6.2 in terms of the solution of the associated BSRDE.

Using the stochastic maximum principle (see Peng [22], and Tang and Li [30], for example), we have the following. Let $\widetilde{u}$ be the optimal control, and $\widetilde{X} := X^{0,x;\widetilde{u}}$. Then, there is some $\lambda \in R^n$, and a pair of processes $(\widetilde{p}, \widetilde{q})$, such that

(102)
$$\begin{cases} d\widetilde{p} = -\left[A^*\widetilde{p} + Q(\widetilde{X} - q) + \sum_{i=1}^{d} C_i^*\widetilde{q}_i\right] ds + \sum_{i=1}^{d} \widetilde{q}_i\, dw_i, \qquad 0 < s \le T, \\ \widetilde{p}(T) = M(\widetilde{X}(T) - \xi) - \lambda \end{cases}$$

and

$$B^*\widetilde{p} + \sum_{i=1}^{d} D_i^*\widetilde{q}_i + N\widetilde{u} = 0. \tag{103}$$

Using Itô's formula and equality (103), we get the equation for $\widetilde{\psi} := K\widetilde{X} - \widetilde{p}$:

(104)
$$\begin{cases} d\widetilde{\psi}(t) = -\left[\widehat{A}^*\widetilde{\psi} + \sum_{i=1}^{d} \widehat{C}_i^*(\widetilde{\phi}_i - Kg_i) - Kf - \sum_{i=1}^{d} L_i g_i + Qq\right] dt + \sum_{i=1}^{d} \widetilde{\phi}_i\, dw_i, \\ \widetilde{\psi}(T) = M\xi + \lambda \end{cases}$$

where (K, L) is the unique $\{\mathcal{F}_t, 0 \le t \le T\}$-adapted solution of BSRDE (1), and the explicit formula of the optimal control,

$$\begin{aligned} \widetilde{u} = -\left(N + \sum_{i=1}^{d} D_i^* K D_i\right)^{-1} \Bigg[\left(B^*K + \sum_{i=1}^{d} D_i^* K C_i + \sum_{i=1}^{d} D_i^* L_i\right)\widehat{X} \\ -B^*\widetilde{\psi} + \sum_{i=1}^{d} D_i^*(Kg_i - \widetilde{\phi}_i)\Bigg], \end{aligned} \tag{105}$$

where the Lagrange multiple λ is determined such that the terminal constraint $E\widetilde{X}(T) = x_T$ is satisfied.

6.3. A comment on application of the LQ theory in mathematical finance. One-dimensional singular LQ problems arise from mathematical finance. The mean-variance hedging problem and the dynamic version of Markowitz's mean-variance portfolio selection problem are one-dimensional singular LQ problems.

The mean-variance hedging problem was extensively studied, among others, by Duffie and Richardson [9], Schweizer [25, 26, 27], Hipp [12], Monat and Stricker [18], Pham, Rheinländer, and Schweizer [23], Gourieroux, Laurent, and Pham [11], and Laurent and Pham [17]. Most of these works use a projection argument. Recently, Kohlmann and Zhou [16] used a natural LQ theory approach to solve the case of deterministic market conditions. Kohlmann and Tang [13, 14] used a natural LQ theory approach to solve the case of stochastic market conditions, and the optimal hedging portfolio and the variance-optimal martingale measure are characterized in terms of the solution of the associated BSRDE.

The study on continuous time mean-variance portfolio selection problem is dated back to Richardson [24]. The reader is referred to Zhou and Li [33] for recent developments on this problem.

Acknowledgment. The second author would like to thank the hospitality of Department of Mathematics and Statistics and the Center of Finance and Econometrics, Universität Konstanz, Germany.

REFERENCES

[1] J. M. Bismut, *Conjugate convex functions in optimal stochastic control*, J. Math. Anal. Appl., 44 (1973), pp. 384–404.
[2] J.-M. Bismut, *Linear quadratic optimal stochastic control with random coefficients*, SIAM J. Control Optim., 14 (1976), pp. 419–444.
[3] J. M. Bismut, *Controle des systems lineares quadratiques: applications de l'integrale stochastique*, in Séminaire de Probabilités XII, Lecture Notes in Math. 649, C. Dellacherie, P. A. Meyer, and M. Weil, eds., Springer–Verlag, Berlin, 1978, pp. 180–264.
[4] S. Chen, X. Li, and X. Y. Zhou, *Stochastic linear quadratic regulators with indefinite control weight costs*, SIAM J. Control Optim., 36 (1998), pp. 1685–1702.
[5] S. Chen and J. Yong, *Stochastic linear quadratic optimal control problems*, Appl. Math. Optim., 43 (2001), pp. 21–45.
[6] S. Chen and J. Yong, *Stochastic linear quadratic optimal control problems with random coefficients*, Chinese Ann. Math. Ser. B, 21 (2000), pp. 323–338.
[7] S. Chen and J. Yong, *Solvability of a stochastic linear quadratic optimal control problem*, in Applied Probability, R. Chan, Y.-K. Kwok, D. Yao, and Q. Zhang, eds., AMS, Providence, RI, 2002, pp. 35–43.
[8] S. Chen and X. Y. Zhou, *Stochastic linear quadratic regulators with indefinite control weight costs.* II, SIAM J. Control Optim., 39 (2000), pp. 1065–1081.
[9] D. Duffie and H. R. Richardson, *Mean-variance hedging in continuous time*, Ann. Appl. Probab., 1 (1991), pp. 1–15.
[10] L. I. Gal'chuk, *Existence and uniqueness of a solution for stochastic equations with respect to semimartingales*, Theory Probab. Appl., 23 (1978), pp. 751–763.
[11] C. Gourieroux, J. P. Laurent, and H. Pham, *Mean-variance hedging and numéraire*, Math. Finance, 8 (1998), pp. 179–200.
[12] C. Hipp, *Hedging general claims*, in Proceedings of the Third AFIR Colloquium, Vol. 2, Rome, 1993, International Actuarial Association, pp. 603–613.
[13] M. Kohlmann and S. Tang, *Minimization of risk and LQ theory*, SIAM J. Control Optim., submitted.
[14] M. Kohlmann and S. Tang, *Global adapted solution of one-dimensional backward stochastic Riccati equations, with application to the mean-variance hedging*, Stochastic Process. Appl., 97 (2002), pp. 255–288.
[15] M. Kohlmann and S. Tang, *New developments in backward stochastic Riccati equations and their applications*, in Mathematical Finance, Trends Math., M. Kohlmann and S. Tang, eds., Birkhäuser, Boston, 2001, pp. 194–214.
[16] M. Kohlmann and X. Y. Zhou, *Relationship between backward stochastic differential equations and stochastic controls: A linear-quadratic approach*, SIAM J. Control Optim., 38 (2000), pp. 1392–1407.
[17] J. P. Laurent and H. Pham, *Dynamic programming and mean-variance hedging*, Finance Stoch., 3 (1999), pp. 83–110.
[18] P. Monat and C. Stricker, *Föllmer-Schweizer decomposition and mean-variance hedging of general claims*, Ann. Probab., 23 (1995), pp. 605–628.
[19] E. Pardoux and S. Peng, *Adapted solution of backward stochastic equation*, Systems Control Lett., 14 (1990), pp. 55–61.
[20] S. Peng, *Stochastic Hamilton–Jacobi–Bellman equations*, SIAM J. Control Optim., 30 (1992), pp. 284–304.
[21] S. Peng, *Open problems on backward stochastic differential equations*, in Control of Distributed Parameter and Stochastic Systems, (Hangzhou, 1998), Kluwer Academic Publishers, Norwell, MA, 1999, pp. 265–273.
[22] S. Peng, *A general stochastic maximum principle for optimal control problems*, SIAM J. Control Optim., 28 (1990), pp. 966–979.
[23] H. Pham, T. Rheinländer, and M. Schweizer, *Mean-variance hedging for continuous processes: New proofs and examples*, Finance Stoch., 2 (1998), pp. 173–198.
[24] H. Richardson, *A minimum result in continuous trading portfolio optimization*, Management Sci., 35 (1989), pp. 1045–1055.
[25] M. Schweizer, *Mean-variance hedging for general claims*, Ann. Appl. Probab., 2 (1992),

pp. 171–179.
[26] M. Schweizer, *Approaching random variables by stochastic integrals*, Ann. Probab., 22 (1994), pp. 1536–1575.
[27] M. Schweizer, *Approximation pricing and the variance-optimal martingale measure*, Ann. Probab., 24 (1996), pp. 206–236.
[28] S. Tang, *General linear quadratic optimal stochastic control problems with random coefficients: Linear stochastic Hamilton systems and backward stochastic Riccati equations*, SIAM J. Control Optim., to appear.
[29] S. Tang, *Financial mean-variance problems and stochastic LQ problems: Linear stochastic Hamilton systems and backward stochastic Riccati equations*, in Recent Developments in Mathematical Finance, J. Yong, ed., World Scientific, River Edge, NJ, 2002, pp. 190–203.
[30] S. Tang and X. Li, *Necessary conditions for optimal control of stochastic systems with random jumps*, SIAM J. Control Optim., 32 (1994), pp. 1447–1475.
[31] W. M. Wonham, *On a matrix Riccati equation of stochastic control*, SIAM J. Control, 6 (1968), pp. 681–697.
[32] J. Yong and X. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer–Verlag, Berlin, New York, 1999.
[33] X. Zhou and D. Li, *Continuous time mean-variance portfolio selection: A stochastic LQ framework*, Appl. Math. Optim., 42 (2000), pp. 19–33.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 6, pp. 1722–1732

ADAPTIVE LOW-GAIN INTEGRAL CONTROL OF MULTIVARIABLE WELL-POSED LINEAR SYSTEMS*

HARTMUT LOGEMANN† AND STUART TOWNLEY‡

To Ruth F. Curtain, on the occasion of her 60th birthday

Abstract. The principle of low-gain integral control for finite-dimensional systems is well known. More recently, low-gain integral control results have been obtained for classes of infinite-dimensional systems. In this paper we show that integral control with a simple and natural adaptation of the integrator gain achieves tracking of constant reference signals for every exponentially stable, multivariable, well-posed, infinite-dimensional, linear system whose steady-state gain matrix has its spectrum in the open right-half plane. Our results considerably extend, improve, and simplify previous work by the authors [*SIAM J. Control Optim.* 35 (1997), pp. 78–116].

Key words. adaptive control, adaptive tracking, integral control, well-posed infinite-dimensional systems

AMS subject classifications. 93C20, 93C25, 93C40, 93D15, 93D21

PII. S0363012901396680

1. Introduction.

There has been much interest over the last twenty-five years in low-gain integral control. Indeed, the following principle has become well established (see Davison [2], Lunze [7], and Morari [10]): closing the loop around an asymptotically stable, finite-dimensional, continuous-time plant, with transfer-function matrix $\mathbf{G}(s)$, compensated by an integrator $(k/s)I$ (see Figure 1), will result in an asymptotically stable closed-loop system which achieves asymptotic tracking of arbitrary constant reference signals, provided that the gain parameter $k > 0$ is sufficiently small and the eigenvalues of the steady-state gain matrix $\mathbf{G}(0)$ have positive real parts, i.e.,

$$\text{spectrum}(\mathbf{G}(0)) \subset \{s \in \mathbb{C} \mid \operatorname{Re} s > 0\}. \tag{1.1}$$

This principle has been extended to various classes of infinite-dimensional systems; see Logemann and Townley [6] and the references therein. The generalization in [6] applies to so-called regular well-posed systems. We remark that the class of well-posed linear systems is the largest class of infinite-dimensional systems for which a well-developed state-space and frequency-domain theory exists; see Curtain and Weiss [3], Salamon [15], Staffans and Weiss [19], and Weiss [20], to mention just a few references. Well-posed systems are rather general in the sense that they capture most distributed parameter systems and all time-delay systems (retarded and neutral) which are of interest in applications. A well-posed system is called regular if the average of its step-response over $[0, t]$ converges as $t \to 0$ (equivalently, if its transfer functions $\mathbf{G}(s)$ converges as $s \to \infty$ on the positive real axis). Whilst the authors believe that any physically motivated well-posed linear system is regular, for a given well-posed system, regularity can be difficult to check.

*Received by the editors October 18, 2001; accepted for publication (in revised form) July 8, 2002; published electronically February 4, 2003. This work was supported in part by UK EPSRC grant GR/L78086.

http://www.siam.org/journals/sicon/41-6/39668.html

†Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, United Kingdom (hl@maths.bath.ac.uk).

‡School of Mathematical Sciences, University of Exeter, Exeter EX4 4QE, United Kingdom (townley@maths.ex.ac.uk).

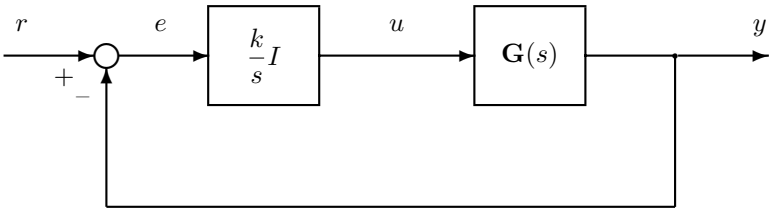


FIG. 1. *Low-gain control system.*

One of the main issues in the design of low-gain integral controllers is the tuning of the integrator gain k. There have been two basic approaches to the tuning problem—either steady-state data from the plant is used off-line to determine suitable ranges for the gain k (see, for example, [2], Logemann, Ryan, and Townley [5], or [7]) or else simple on-line adaptive tuning of k is used (see, for example, Miller and Davison [8, 9] in the finite-dimensional case and [6] in the infinite-dimensional case). Of particular relevance here is a result in [6] which shows that the adaptive integral controller

$$\dot{u}(t) = \gamma^{-p}(t)(r - y(t))\,, \qquad \dot{\gamma}(t) = \|r - y(t)\|^2 \tag{1.2}$$

achieves asymptotic tracking of arbitrary constant reference signals r, provided that the following three assumptions are satisfied:

(i) the plant is an exponentially stable, regular, well-posed, infinite-dimensional system;
(ii) the steady-state gain matrix $\mathbf{G}(0)$ is symmetric and positive definite;
(iii) the parameter p in (1.2) satisfies $p \in (0, 1/2)$.

We note that earlier work by Cook [1] shows that in the finite-dimensional single-input single-output case, assumption (iii) can be relaxed to $p \in (0, 1]$. Of course, the symmetry assumption in (ii) is restrictive and highly nonrobust, essentially limiting the applications of the above result to single-input single-output systems. The main result of this note (Theorem 3.1) shows that assumption (ii) can be replaced by the considerably weaker assumption (1.1), that the regularity assumption in (i) can be dropped, and that (iii) can replaced by $p \in (0, 1]$. Furthermore, in comparing the results we present here to those in [6], the proofs are dramatically simplified and more importantly give a clearer insight into the structure of the resulting closed-loop system. We emphasize that our main result is new even in the finite-dimensional case.

Notation. $\mathbb{R}_+ := [0, \infty)$; for $\alpha \in \mathbb{R}$, set $\mathbb{C}_\alpha := \{s \in \mathbb{C} \mid \operatorname{Re} s > \alpha\}$; let Z be a real or complex Banach space; for $\alpha \in \mathbb{R}$, we define the exponentially weighted L^p-space $L^p_\alpha(\mathbb{R}_+, Z) := \{f \in L^p_{\mathrm{loc}}(\mathbb{R}_+, Z) \mid f(\cdot)\exp(-\alpha\,\cdot) \in L^p(\mathbb{R}_+, Z)\}$ and endow it with the norm $\|f\|_{p,\alpha} := \|e^{-\alpha\,\cdot} f(\cdot)\|_{L^p}$; let $H^2(\mathbb{C}_\alpha, Z)$ denote the Hardy–Lebesgue space of square-integrable holomorphic functions defined on $\mathbb{C}_\alpha$ with values in Z; $H^\infty(\mathbb{C}_\alpha, Z)$ denotes the space of bounded holomorphic functions defined on $\mathbb{C}_\alpha$ with values in Z; $\mathcal{B}(Z_1, Z_2)$ denotes the space of bounded linear operators from a Banach space Z_1 to a Banach space Z_2; we write $\mathcal{B}(Z)$ for $\mathcal{B}(Z, Z)$; let $A : \operatorname{dom}(A) \subset Z \to Z$ be a linear operator, where $\operatorname{dom}(A)$ denotes the domain of A; the resolvent set of A and the spectrum of A is denoted by $\varrho(A)$and $\sigma(A)$, respectively; the Laplace transform is denoted by $\mathfrak{L}$.

2. Preliminaries on well-posed sytems. There are a number of equivalent definitions of well-posed systems; see [3, 14, 15, 16, 17, 18, 19, 20, 21]. We will be

brief in the following and refer the reader to the above references for more details. Throughout this section, we shall be considering a well-posed system Σ with state-space X, input space $\mathbb{R}^m$, and output space $\mathbb{R}^m$, generating operators (A, B, C), input-output operator G, and transfer function $\mathbf{G}$. Here X is a real Hilbert space with norm denoted by $\|\cdot\|$, A is the generator of a strongly continuous semigroup $\mathbf{T} = (\mathbf{T}_t)_{t\geq 0}$ on X, $B \in \mathcal{B}(\mathbb{R}^m, X_{-1})$, and $C \in \mathcal{B}(X_1, \mathbb{R}^m)$, where X_1 denotes the space $\mathrm{dom}(A)$ endowed with the norm $\|x\|_1 := \|x\| + \|Ax\|$ (the graph norm of A), whilst X_{-1} denotes the completion of X with respect to the norm $\|x\|_{-1} = \|(\lambda I - A)^{-1}x\|$, where $\lambda \in \varrho(A)$ (different choices of λ lead to equivalent norms). Clearly, $X_1 \subset X \subset X_{-1}$ and the canonical injections are bounded and dense. The semigroup $\mathbf{T}$ restricts to a strongly continuous semigroup on X_1 and extends to a strongly continuous semigroup on X_{-1} with the exponential growth constant being the same on all three spaces; the generator of the restriction (extension) of $\mathbf{T}$ is a restriction (extension) of A; we shall use the same symbol $\mathbf{T}$ (respectively, A) for the original semigroup (respectively, generator) and the associated restrictions and extensions: with this convention, we may write $A \in \mathcal{B}(X, X_{-1})$ (considered as a generator on X_{-1}, the domain of A is X). Moreover, the operator B is an *admissible control operator* for $\mathbf{T}$, i.e., for each $t \in \mathbb{R}_+$ there exists $\alpha_t \geq 0$ such that

$$\left\| \int_0^t \mathbf{T}_{t-\tau} B u(\tau) d\tau \right\| \leq \alpha_t \|u\|_{L^2([0,t],\mathbb{R}^m)} \quad \forall\, u \in L^2([0,t], \mathbb{R}^m)\,.$$

The operator C is an *admissible observation operator* for $\mathbf{T}$, i.e., for each $t \in \mathbb{R}_+$ there exists $\beta_t \geq 0$ such that

$$\left(\int_0^t \|C\mathbf{T}_\tau x\|^2 d\tau \right)^{1/2} \leq \beta_t \|x\| \quad \forall\, x \in X_1\,.$$

The control operator B is said to be *bounded* if it is so as a map from the input space $\mathbb{R}^m$ to the state-space X; otherwise, it is said to be *unbounded*; the observation operator C is said to be *bounded* if it can be extended continuously to X; otherwise, C is said to be *unbounded*.

The so-called Λ-*extension* C_Λ of C is defined by

$$C_\Lambda x = \lim_{s\to\infty,\, s\in\mathbb{R}} Cs(sI - A)^{-1}x\,,$$

with $\mathrm{dom}(C_\Lambda)$ consisting of all $x \in X$ for which the above limit exists. For every $x \in X$, $\mathbf{T}_t x \in \mathrm{dom}(C_\Lambda)$ for a.a. $t \in \mathbb{R}_+$, and, if $\omega > \omega(\mathbf{T})$, then $C_\Lambda \mathbf{T} x \in L^2_\omega(\mathbb{R}_+, \mathbb{R}^m)$, where

$$\omega(\mathbf{T}) := \lim_{t\to\infty} \frac{1}{t} \ln \|\mathbf{T}_t\|$$

denotes the exponential growth constant of $\mathbf{T}$. The transfer function $\mathbf{G}$ satisfies

$$\frac{1}{s-\lambda}\left(\mathbf{G}(s) - \mathbf{G}(\lambda)\right) = -C(sI - A)^{-1}(\lambda I - A)^{-1}B \quad \forall\, s, \lambda \in \mathbb{C}_{\omega(\mathbf{T})},\ s \neq \lambda\,, \tag{2.1}$$

and $\mathbf{G} \in H^\infty(\mathbb{C}_\omega, \mathbb{R}^{m\times m})$ for every $\omega > \omega(\mathbf{T})$. Moreover, the input-output operator $G : L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m) \to L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m)$ is continuous and shift-invariant; for every $\omega > \omega(\mathbf{T})$, $G \in \mathcal{B}(L^2_\omega(\mathbb{R}_+, \mathbb{R}^m))$ and

$$(\mathfrak{L}(Gu))(s) = \mathbf{G}(s)(\mathfrak{L}(u))(s) \quad \forall\, s \in \mathbb{C}_\omega,\ \forall\, u \in L^2_\omega(\mathbb{R}_+, \mathbb{R}^m)\,.$$

In the following, let $\lambda \in \mathbb{C}_{\omega(\mathbf{T})}$ be fixed but arbitrary. For $x^0 \in X$ and $u \in L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m)$, let x and y denote the state and output functions of Σ, respectively, corresponding to the initial condition $x(0) = x^0 \in X$ and the input function u. Then $x(t) = \mathbf{T}_t x^0 + \int_0^t \mathbf{T}_{t-\tau} B u(\tau) d\tau$ for all $t \in \mathbb{R}_+$, $x(t) - (\lambda I - A)^{-1} B u(t) \in \mathrm{dom}(C_\Lambda)$ for a.a. $t \in \mathbb{R}_+$,[1] and

$$\dot{x}(t) = Ax(t) + Bu(t)\,, \quad x(0) = x^0\,, \quad \text{for a.a. } t \in \mathbb{R}_+\,, \tag{2.2a}$$

$$y(t) = C_\Lambda \big(x(t) - (\lambda I - A)^{-1} B u(t)\big) + \mathbf{G}(\lambda) u(t)\,, \quad \text{for a.a. } t \geq 0\,. \tag{2.2b}$$

Of course, the differential equation (2.2a) has to be interpreted in X_{-1}. Note that the output equation (2.2b) yields the following formula for the input-output operator G:

$$(Gu)(t) = C_\Lambda \left[\int_0^t \mathbf{T}_{t-\tau} B u(\tau) d\tau \; - (\lambda I - A)^{-1} B u(t) \right] + \mathbf{G}(\lambda) u(t) \tag{2.3}$$
$$\forall\, u \in L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m), \text{ for a.a. } t \in \mathbb{R}_+.$$

In the following, we identify Σ and (2.2) and refer to (2.2) as a well-posed system. We say that the well-posed system (2.2) is *exponentially stable* if $\omega(\mathbf{T}) < 0$. If the well-posed system (2.2) is *regular*, i.e., the following limit

$$\lim_{s \to \infty,\, s \in \mathbb{R}} \mathbf{G}(s) = D$$

exists, then $x(t) \in \mathrm{dom}(C_\Lambda)$ for a.a. $t \in \mathbb{R}_+$, the output equation (2.2b) simplifies to

$$y(t) = C_\Lambda x(t) + Du(t)\,, \quad \text{for a.a. } t \geq 0,$$

and

$$(Gu)(t) = C_\Lambda \int_0^t \mathbf{T}_{t-\tau} B u(\tau) d\tau + Du(t) \quad \forall\, u \in L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m)\,, \text{ for a.a. } t \in \mathbb{R}_+\,.$$

Moreover, in the regular case, we have that $(sI - A)^{-1} B \mathbb{R}^m \subset \mathrm{dom}(C_\Lambda)$ for all $s \in \varrho(A)$ and

$$\mathbf{G}(s) = C_\Lambda (sI - A)^{-1} B + D \quad \forall\, s \in \mathbb{C}_{\omega(\mathbf{T})}\,.$$

The matrix $D \in \mathbb{R}^{m \times m}$ is called the *feedthrough matrix* of (2.2). We mention that if the control operator B or the observation operator C is bounded, then (2.2) is regular.

3. Main result. We consider adaptive low-gain integral control of an exponentially stable well-posed system of the form (2.2). By exponential stability, we may assume w.l.o.g. that $\lambda = 0$ in (2.2b), and hence the plant equations are given by

$$\dot{x} = Ax + Bu\,, \quad x(0) = x^0 \in X\,, \tag{3.1a}$$

$$y = C_\Lambda (x + A^{-1} B u) + \mathbf{G}(0) u\,. \tag{3.1b}$$

Let $r \in \mathbb{R}^m$ be a given reference vector and consider the following simple adaptive low-gain integral controller:

$$\dot{u} = \gamma^{-p} (r - y)\,, \quad u(0) = u^0 \in \mathbb{R}^m\,, \tag{3.2a}$$

$$\dot{\gamma} = \|r - y\|^2\,, \quad \gamma(0) = \gamma^0 > 0\,, \tag{3.2b}$$

[1] It was stated in [21] (without proof) that, for arbitrary $u \in L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m)$, $x(t) - (\lambda I - A)^{-1} B u(t) \in \mathrm{dom}(C_\Lambda)$ for a.a. $t \in \mathbb{R}_+$ and the output formula (2.2b) holds. The proof can be found in [19].

where $0 < p \leq 1$. The closed-loop system is then given by

$$\dot{x} = Ax + Bu\,, \quad x(0) = x^0 \in X\,, \tag{3.3a}$$

$$\dot{u} = \gamma^{-p}\left(r - C_\Lambda(x + A^{-1}Bu) - \mathbf{G}(0)u\right)\,, \quad u(0) = u^0 \in \mathbb{R}^m\,, \tag{3.3b}$$

$$\dot{\gamma} = \|r - C_\Lambda(x + A^{-1}Bu) - \mathbf{G}(0)u\|^2\,, \quad \gamma(0) = \gamma^0 > 0\,. \tag{3.3c}$$

Let $T > 0$. A continuous function $(x, u, \gamma) : [0, T) \to X \times \mathbb{R}^m \times \mathbb{R}$ is called a *solution* of (3.3) if $(x(0), u(0), \gamma(0)) = (x^0, u^0, \gamma^0)$, (x, u, γ) is absolutely continuous on $[0, t]$ as a $(X_{-1} \times \mathbb{R}^m \times \mathbb{R})$-valued function for every $t \in (0, T)$ and the differential equations in (3.3) are satisfied almost everywhere on $[0, T)$.

Our main result, Theorem 3.1, shows that the controller (3.2) achieves tracking of constant reference signals for all exponentially stable well-posed systems (3.1) whose steady-state gain matrix $\mathbf{G}(0)$ satisfies $\sigma(\mathbf{G}(0)) \subset \mathbb{C}_0$.

THEOREM 3.1. *Assume that the well-posed system* (3.1) *is exponentially stable with* $\sigma(\mathbf{G}(0)) \subset \mathbb{C}_0$ *and that* $0 < p \leq 1$ *in* (3.2). *Let* $r \in \mathbb{R}^m$ *be given and define* $u^r := [\mathbf{G}(0)]^{-1}r$. *Then, for all* $(x^0, u^0, \gamma^0) \in X \times \mathbb{R}^m \times (0, \infty)$, *there exists a unique solution* $(x, u, \gamma) : \mathbb{R}_+ \to X \times \mathbb{R}^m \times (0, \infty)$ *of the closed-loop system* (3.3) *and the following statements hold:*

(1) $\lim_{t\to\infty} \gamma(t) = \gamma^\infty < \infty$ *;*
(2) $u - u^r \in L^2(\mathbb{R}_+, \mathbb{R}^m)$ *and* $\lim_{t\to\infty} u(t) = u^r$ *;*
(3) $x + A^{-1}Bu^r \in L^2(\mathbb{R}_+, X)$ *and* $\lim_{t\to\infty} \|x(t) + A^{-1}Bu^r\| = 0$ *;*
(4) $e := r - y \in L^2(\mathbb{R}_+, \mathbb{R}^m)$ *and* e *admits a decomposition of the form* $e = e_1 + e_2$, *where* $e_1 \in C(\mathbb{R}_+, \mathbb{R}^m)$, $e_2 \in L^2_\omega(\mathbb{R}_+, \mathbb{R}^m)$ *for every* $\omega > \omega(\mathbf{T})$, *and* $\lim_{t\to\infty} e_1(t) = 0$ *; moreover, if there exists* $t_0 \geq 0$ *such that* $\mathbf{T}_{t_0}(Ax^0 + Bu^0) \in X$, *then* $e \in C([t_0, \infty), \mathbb{R}^m)$ *and* $\lim_{t\to\infty} e(t) = 0$.

Proof. As in [6] it can be proved that there exists a unique maximally defined solution $(x, u, \gamma) : [0, T) \to X \times \mathbb{R}^m \times (0, \infty)$ to the closed-loop system (3.3), where $T = \infty$ if γ is bounded on $[0, T)$. To analyze the stability of the closed-loop system (3.3), we use a change of coordinates. Define

$$z(t) := x(t) + A^{-1}Bu(t)\,, \qquad v(t) := u(t) - u^r\,; \tag{3.4}$$

it follows from (3.1) and (3.2) that

$$\dot{z}(t) = Az(t) + \gamma^{-p}(t)A^{-1}Be(t)\,, \quad \text{for a.a. } t \in [0, T)\,, \tag{3.5a}$$

$$\dot{v}(t) = \gamma^{-p}(t)e(t)\,, \quad \text{for a.a. } t \in [0, T)\,, \tag{3.5b}$$

where

$$e(t) := r - y(t) = -\left(C_\Lambda z(t) + \mathbf{G}(0)v(t)\right)\,, \quad \text{for a.a. } t \in [0, T)\,. \tag{3.6}$$

Of course, the derivative on the left-hand side of (3.5a) has to be interpreted in X_{-1}. There are two advantages to viewing the closed-loop system in the coordinates (3.4): the unbounded B in (3.1a) is replaced by a bounded $A^{-1}B$ and it turns out that the stability of $-\mathbf{G}(0)$ is easier to exploit in the (z, v)-coordinates than in the (x, u)-coordinates.

To proceed, we use the stability of $-\mathbf{G}(0)$ to obtain the existence of $Q = Q^T \in \mathbb{R}^{m\times m}$ with $Q > 0$ such that

$$Q\mathbf{G}(0) + \mathbf{G}(0)^T Q = I. \tag{3.7}$$

Furthermore, using the exponential stability of $\mathbf{T}$ and the admissibility of C, a standard argument (see the appendix) shows that there exists $P = P^* \in \mathcal{B}(X)$ with $P \geq 0$ and such that

$$\langle Ax_1, Px_2\rangle + \langle Px_1, Ax_2\rangle = -\langle x_1, x_2\rangle - \langle Cx_1, Cx_2\rangle \quad \forall\, x_1, x_2 \in X_1\,. \tag{3.8}$$

Our aim is to show that γ is bounded on $[0,T)$, from which we will deduce that $T = \infty$ and that statements (1)–(4) hold. To this end, define a function $V : [0,T) \to \mathbb{R}_+$ by

$$V(t) = \langle z(t), Pz(t)\rangle + \langle v(t), Qv(t)\rangle \quad \forall\, t \in [0,T)\,,$$

where the first inner product is taken in X and the second inner product is the standard inner product in $\mathbb{R}^m$. Since, due to lack of regularity of z as a X-valued function, V is in general not differentiable, we adopt an approximation argument. Define

$$z_n(t) := \mathbf{T}_t z_n^0 + A^{-1}\int_0^t \mathbf{T}_{t-\tau} B\gamma^{-p}(\tau)e(\tau)d\tau \quad \forall\, t \in [0,T)\,,$$

where $z_n^0 \in X_1$ is such that $z_n^0 \to z(0) = x^0 + A^{-1}Bu^0$ as $n \to \infty$. Clearly, $z_n(t) \in X_1$ for all $t \in [0,T)$. It follows from the admissibility of B and well-known results on abstract Cauchy problems (see [11, p. 109]) that z_n is absolutely continuous as a X-valued function and

$$\dot{z}_n(t) = Az_n(t) + \gamma^{-p}(t)A^{-1}Be(t)\,, \quad \text{for a.a. } t \in [0,T)\,.$$

Therefore, the function

$$V_n : [0,T) \to \mathbb{R}_+\,, \quad t \mapsto \langle z_n(t), Pz_n(t)\rangle + \langle v(t), Qv(t)\rangle$$

is absolutely continuous. Invoking (3.5)–(3.8), we compute the derivative of V_n to be

$$\dot{V}_n = -\|z_n\|^2 - \|Cz_n\|^2 + 2\gamma^{-p}\langle z_n, PA^{-1}Be\rangle - \gamma^{-p}\|v\|^2 - 2\gamma^{-p}\langle v, QC_\Lambda z\rangle. \tag{3.9}$$

Integrating (3.9) from s to t, where $0 \leq s \leq t < T$, we obtain

$$\begin{aligned} V_n(t) - V_n(s) = -\int_s^t \big(&\|z_n(\tau)\|^2 + \|Cz_n(\tau)\|^2 + \gamma^{-p}\|v(\tau)\|^2 \\ &- 2\gamma^{-p}(\tau)\langle z_n(\tau), PA^{-1}Be(\tau)\rangle + 2\gamma^{-p}(\tau)\langle v(\tau), QC_\Lambda z(\tau)\rangle\big)\, d\tau. \end{aligned} \tag{3.10}$$

It follows from (3.5a) that $z(t) := \mathbf{T}_t z(0) + A^{-1}\int_0^t \mathbf{T}_{t-\tau}B\gamma^{-p}(\tau)e(\tau)d\tau$, showing that $z(t) - z_n(t) = \mathbf{T}_t(z(0) - z_n^0)$ for all $t \in [0,T)$ and hence for all $t \in [0,T)$

$$\lim_{n\to\infty}\|z(t) - z_n(t)\| = 0, \quad \lim_{n\to\infty}\|z - z_n\|_{L^2(0,t)} = 0, \quad \lim_{n\to\infty}\|C_\Lambda z - C_\Lambda z_n\|_{L^2(0,t)} = 0,$$

where the last limit follows from the admissibility of C. Consequently, letting $n \to \infty$ in (3.10), we may conclude that

$$\begin{aligned} V(t) - V(s) = -\int_s^t \big(&\|z(\tau)\|^2 + \|C_\Lambda z(\tau)\|^2 + \gamma^{-p}\|v(\tau)\|^2 \\ &- 2\gamma^{-p}(\tau)\langle z(\tau), PA^{-1}Be(\tau)\rangle + 2\gamma^{-p}(\tau)\langle v(\tau), QC_\Lambda z(\tau)\rangle\big)\, d\tau. \end{aligned} \tag{3.11}$$

Denoting the integrand on the right-hand side of (3.11) by $f(\tau)$ and using (3.6), routine estimates give

$$\begin{aligned} f &\geq \|z\|^2 + \|C_\Lambda z\|^2 + c_1\gamma^{-p}\|\mathbf{G}(0)v\|^2 \\ &\quad - c_2\gamma^{-p}\left(\|z\|\|C_\Lambda z\| + \|z\|\|\mathbf{G}(0)v\| + \|C_\Lambda z\|\|\mathbf{G}(0)v\|\right) \\ &\geq \|z\|^2 + \|C_\Lambda z\|^2 + c_1\gamma^{-p}\|\mathbf{G}(0)v\|^2 - c_2\gamma^{-p}\left(\|z\|^2 + \|C_\Lambda z\|^2\right) \\ &\quad - c_2\gamma^{-p}\left(\lambda\|z\|^2 + \|\mathbf{G}(0)v\|^2/\lambda\right) - c_2\gamma^{-p}\left(\lambda\|C_\Lambda z\|^2 + \|\mathbf{G}(0)v\|^2/\lambda\right), \end{aligned} \tag{3.12}$$

where $c_1, c_2 > 0$ are suitable constants and $\lambda > 0$ is arbitrary. In the following we choose

$$\lambda = 4c_2/c_1\,. \tag{3.13}$$

We show that γ is bounded on $[0,T)$. If

$$\gamma^{-p}(\tau) > 1/[2c_2(1+\lambda)] \quad \forall\, t \in [0,T)\,, \tag{3.14}$$

then there is nothing to prove. So assume that (3.14) does not hold. Then, by monotonicity of γ, there exists $t_0 \in [0,T)$ such that

$$\gamma^{-p}(\tau) \leq 1/[2c_2(1+\lambda)] \quad \forall\, t \in [t_0,T)\,.$$

Combining this with (3.12) and (3.13) yields

$$\begin{aligned} f(\tau) &\geq \left(\|z(\tau)\|^2 + \|C_\Lambda z(\tau)\|^2 + c_1\gamma^{-p}(\tau)\|\mathbf{G}(0)v(\tau)\|^2\right)/2 \\ &\geq c_3\gamma^{-p}(\tau)\left(\|C_\Lambda z(\tau)\|^2 + \|\mathbf{G}(0)v(\tau)\|^2\right) \qquad \forall\, t \in [t_0,T)\,, \end{aligned}$$

where $c_3 > 0$ is a suitable constant. Noting that

$$\|r-y\|^2 = \|e\|^2 = \|C_\Lambda z + \mathbf{G}(0)v\|^2 \leq 2\left(\|C_\Lambda z\|^2 + \|\mathbf{G}(0)v\|^2\right),$$

we see that there exists a constant $c_4 > 0$ such that

$$f(\tau) \geq c_4\gamma^{-p}(\tau)\|r-y(\tau)\|^2 \quad \forall\, \tau \in [t_0,T)\,.$$

Therefore, by (3.11),

$$V(t) - V(t_0) = -\int_{t_0}^{t} f(\tau)d\tau \leq -c_4\int_{t_0}^{t}\gamma^{-p}(\tau)\|r-y(\tau)\|^2 d\tau \quad \forall\, t \in [t_0,T)\,. \tag{3.15}$$

But $\dot\gamma = \|r-y\|^2$, and hence, by (3.15),

$$\int_{\gamma(t_0)}^{\gamma(t)} w^{-p}dw = \int_{t_0}^{t}\gamma^{-p}(\tau)\dot\gamma(\tau)d\tau \leq V(t_0)/c_4 \quad \forall\, t \in [t_0,T)\,.$$

Since $0 < p \leq 1$, this inequality implies that γ is bounded on $[0,T)$, showing that $T = \infty$ and also establishing statement (1).

To prove statements (2) and (3) note that

$$e \in L^2(\mathbb{R}_+, \mathbb{R}^m)\,, \tag{3.16}$$

which follows immediately from the boundedness of γ and the fact that $\dot{\gamma} = \|e\|^2$. Since $(A, A^{-1}B, C)$ are the generators of an exponentially stable regular system (with zero feedthrough), it follows from (3.5a) and (3.16) that

$$\lim_{t\to\infty} \|z(t)\| = 0\,, \quad z \in L^2(\mathbb{R}_+, X)\,, \quad C_\Lambda z \in L^2(\mathbb{R}_+, \mathbb{R}^m). \tag{3.17}$$

Using that $C_\Lambda z \in L^2(\mathbb{R}_+, \mathbb{R}^m)$, it follows from (3.6), (3.16), and the invertibility of $\mathbf{G}(0)$ that $v \in L^2(\mathbb{R}_+, \mathbb{R}^m)$. But by (3.5b) and (3.16) we also have that $\dot{v} \in L^2(\mathbb{R}_+, \mathbb{R}^m)$, showing that $\lim_{t\to\infty} v(t) = 0$ and completing the proof of statement (2). Since $x + A^{-1}Bu^r = z + A^{-1}B(u^r - u)$, statement (3) follows from statement (2) and (3.17).

To prove statement (4), we first note that we have already shown that $e \in L^2(\mathbb{R}_+, \mathbb{R}^m)$, and so

$$\dot{u} \in L^2(\mathbb{R}_+, \mathbb{R}^m)\,. \tag{3.18}$$

We recall that G denotes the input-output operator of (3.1). Define a shift-invariant operator $H : L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m) \to L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m)$ by setting

$$(Hw)(t) := \int_0^t \left((Gw)(\tau) - \mathbf{G}(0)w(\tau)\right) d\tau \quad \forall\, w \in L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m) \quad \forall\, t \in \mathbb{R}_+\,.$$

The transfer function $\mathbf{H}$ of H is given by $\mathbf{H}(s) = (\mathbf{G}(s) - \mathbf{G}(0))/s$. Clearly, for every $\omega > \omega(\mathbf{T})$

$$\mathbf{H} \in H^2(\mathbb{C}_\omega, \mathbb{C}^{m\times m}) \cap H^\infty(\mathbb{C}_\omega, \mathbb{C}^{m\times m})\,,$$

showing in particular that H is bounded, i.e., $H \in \mathcal{B}(L^2(\mathbb{R}_+, \mathbb{R}^m))$. Using that G commutes with the integration operator (by shift-invariance), a routine calculation gives

$$Gu = H\dot{u} + \mathbf{G}(0)u + G(u^0\theta) - \mathbf{G}(0)u^0\,,$$

where θ denotes the unit-step function. Invoking the output formula $y = C_\Lambda \mathbf{T}x^0 + Gu$, we may write $e = r - y = e_1 + e_2$, where

$$e_1 := \mathbf{G}(0)(u^r - u) - H\dot{u}\,, \quad e_2 := \mathbf{G}(0)u^0 - G(u^0\theta) - C_\Lambda \mathbf{T}x^0\,.$$

Clearly, e_1 is continuous. Using (3.18) and the boundedness of H and G, it follows that $H\dot{u}$ and $(d/dt)(H\dot{u})$ are in $L^2(\mathbb{R}_+, \mathbb{R}^m)$ and hence $\lim_{t\to\infty}(H\dot{u})(t) = 0$. Combining this with statement (2) shows that $\lim_{t\to\infty} e_1(t) = 0$. Let $\omega > \omega(\mathbf{T})$. To prove that $e_2 \in L^2_\omega(\mathbb{R}_+, \mathbb{R}^m)$, it is sufficient to show that $g := \mathbf{G}(0)u^0 - G(u^0\theta) \in L^2_\omega(\mathbb{R}_+, \mathbb{R}^m)$. But $(\mathfrak{L}g)(s) = -\mathbf{H}(s)u^0$, and so $\mathfrak{L}g \in H^2(\mathbb{C}_\omega, \mathbb{C}^m)$, which in turn implies (by a well-known result of Paley and Wiener, see [12, p. 405]) that $g \in L^2_\omega(\mathbb{R}_+, \mathbb{R}^m)$. Finally, assume that there exists $t_0 \geq 0$ such that $\mathbf{T}_{t_0}(Ax^0 + Bu^0) \in X$. Taking the Laplace transform of e_2 gives

$$(\mathfrak{L}e_2)(s) = \left(\mathbf{G}(0) - \mathbf{G}(s)\right) u^0/s - C(sI - A)^{-1}x^0 \quad \forall\, s \in \mathbb{C}_\omega\,.$$

Invoking (2.1) leads to

$$\begin{aligned}(\mathfrak{L}e_2)(s) &= -C(sI - A)^{-1}A^{-1}Bu^0 - C(sI - A)^{-1}x^0 \\ &= -C(sI - A)^{-1}A^{-1}(Ax^0 + Bu^0) \quad \forall\, s \in \mathbb{C}_\omega,\end{aligned}$$

implying that $e_2(t) = -C_\Lambda \mathbf{T}_t A^{-1}(Ax^0 + Bu^0)$ for a.a. $t \in \mathbb{R}_+$. Hence, since $\mathbf{T}_{t_0}(Ax^0 + Bu^0) \in X$,

$$e_2(t) = -C\mathbf{T}_{t-t_0} A^{-1} \mathbf{T}_{t_0}(Ax^0 + Bu^0)\,, \quad \text{for a.a. } t \geq t_0\,,$$

showing that e_2, and hence e, is continuous on $[t_0, \infty)$ and $\lim_{t\to\infty} e_2(t) = 0 = \lim_{t\to\infty} e(t)$. □

Remark 3.2. (1) Statement (4) in Theorem 3.1 shows that the tracking error e becomes small in the sense that $e = e_1 + e_2$, where $e_1(t)$ converges to 0 as $t \to \infty$, $e_1 \in L^2(\mathbb{R}_+, \mathbb{R}^m)$, and $e_2 \in L^2_\omega(\mathbb{R}_+, \mathbb{R}^m)$ for $\omega > \omega(\mathbf{T})$. This implies, in particular, "tracking in measure," i.e., for all $\varepsilon > 0$ we have that

$$\lim_{\tau\to\infty} \mu_L\left(\{t \geq \tau \,|\, \|e(t)\| \geq \varepsilon\}\right) = 0\,,$$

where μ_L denotes the Lebesgue measure on $\mathbb{R}_+$. The last part of statement (4) shows that "asymptotic tracking" (i.e., $\lim_{t\to\infty} e(t) = 0$) is guaranteed, provided that $\mathbf{T}_{t_0}(Ax^0 + Bu^0) \in X$ for some $t_0 \geq 0$ (which, for example, is the case if $\mathbf{T}$ is holomorphic).

(2) Under the conditions of Theorem 3.1, it is easy to see that if $\mathbf{T}_{t_0} x^0 \in X_1$ for some $t_0 \geq 0$ and the convolution kernel of G is a finite (matrix-valued) Borel measure, then $\lim_{t\to\infty} e(t) = 0$.

(3) Combining the above change of coordinates technique with the approach in [4], it can be shown that the adaptation law (3.2b) can be generalized to $\dot\gamma(t) = \|r - y(t)\|^q$ for arbitrary $q \geq 1$.

(4) Theorem 3.1 remains true if we replace the finite-dimensional input space $\mathbb{R}^m$ by an arbitrary real Hilbert space; the proof carries over word for word to this more general situation.

(5) Suppose that the parameter p in (3.2a) satisfies $p \in (0, 1]$. Then, by Theorem 3.1, for the adaptive low-gain integral controller (3.2) to achieve its objective, it is sufficient that the following two assumptions are satisfied:

(i) the semigroup generated by A is exponentially stable;

(ii) $\sigma(\mathbf{G}(0)) \subset \mathbb{C}_0$.

It can be shown that if the well-posed system (2.2) is low-gain integral stabilizable, i.e., there exists $k^* \in (0, \infty]$ such that (2.2) is exponentially stabilized by the integrator $\dot u = -ky$ for all $k \in (0, k^*)$, then $\sigma(A) \cap \mathbb{C}_0 = \emptyset$ and $\sigma(\mathbf{G}(0)) \subset \overline{\mathbb{C}}_0$. Note that these necessary conditions for low-gain integral stabilizability are only "slightly weaker" than the sufficient conditions (i) and (ii) for adaptive low-gain integral control. Simple counterexamples (see the appendix) show that low-gain integral stabilizability does not imply either of the above conditions (i) or (ii).

(6) If, in (3.2a), $p > 1$, then in general the adaptive controller fails, that is, the conclusions of Theorem 3.1 are not valid. To see this, consider the case of an exponentially stable regular single-input single-output system with $C = 0$ and feedthrough $D = 1$. Then $\mathbf{G}(s) \equiv \mathbf{G}(0) = D = 1$. Suppose that $r = 0$, so that $y(t) = u(t)$. Let $p = 1 + \varepsilon$, where $\varepsilon > 0$. Then (3.2) becomes

$$\begin{aligned} \dot u &= -\gamma^{-(1+\varepsilon)} u\,, \quad u(0) \in \mathbb{R}\,, \\ \dot\gamma &= u^2\,, \qquad\qquad \gamma(0) > 0\,, \end{aligned}$$

from which it follows that $u\dot u = -\gamma^{-(1+\varepsilon)}\dot\gamma$. Integration from 0 to t yields

$$u^2(t) = u^2(0) + \frac{2}{\varepsilon}\left(\gamma^{-\varepsilon}(t) - \gamma^{-\varepsilon}(0)\right).$$

If $u(0) > \sqrt{2\gamma^{-\varepsilon}(0)/\varepsilon}$, then $\lim_{t\to\infty} u(t) > 0$, and so the tracking of $r = 0$ is not achieved.

4. Appendix.

Existence of a self-adjoint positive semidefinite solution to the Lyapunov equation (3.8). It follows from the exponential stability of $\mathbf{T}$ and the admissibility of C that the bilinear form $F : X \times X \to \mathbb{R}$ defined by

$$F(x_1, x_2) := \int_0^\infty \langle \mathbf{T}_t x_1, \mathbf{T}_t x_2 \rangle dt + \int_0^\infty \langle C_\Lambda \mathbf{T}_t x_1, C_\Lambda \mathbf{T}_t x_2 \rangle dt$$

is bounded. Consequently, there exists $P \in \mathcal{B}(X)$ such that $F(x_1, x_2) = \langle x_1, Px_2 \rangle$ for all $x_1, x_2 \in X$ (see [13, Theorem 12.8, p. 296]). It is clear that $P = P^* \geq 0$. Moreover,

$$\begin{aligned}\langle Ax_1, Px_2 \rangle + \langle Ax_2, Px_1 \rangle &= \int_0^\infty \frac{d}{dt} \langle \mathbf{T}_t x_1, \mathbf{T}_t x_2 \rangle + \int_0^\infty \frac{d}{dt} \langle C\mathbf{T}_t x_1, C\mathbf{T}_t x_2 \rangle \\ &= -\langle x_1, x_2 \rangle - \langle Cx_1, Cx_2 \rangle \quad \forall\, x_1, x_2 \in \mathrm{dom}(A^2)\,.\end{aligned}$$

Since $\mathrm{dom}(A^2)$ is dense in X_1, $A \in \mathcal{B}(X_1, X)$, and $C \in \mathcal{B}(X_1, \mathbb{R}^m)$ the above identity extends to all of X_1, showing that (3.8) holds.

Counterexamples showing that low-gain integral stabilizability does not imply conditions (i) or (ii) in part (5) of Remark 3.2. Consider the finite-dimensional system (with zero feedthrough) given by

$$A = \begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad C = (1/2, 1).$$

This system is integral stabilizable (with $k^* = \infty$), but $0 \in \sigma(A)$, showing that condition (i) in part (5) of Remark 3.2 does not hold.

The finite-dimensional system given by

$$A = \begin{pmatrix} -2 & 0 \\ 0 & -2 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad D = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

is integral stabilizable (with $k^* = \infty$). However, the transfer function is given by

$$\mathbf{G}(s) = C(sI - A)^{-1}B + D = \begin{pmatrix} 0 & -\dfrac{s+1}{s+2} \\ \dfrac{s+1}{s+2} & 0 \end{pmatrix},$$

and thus $\sigma(\mathbf{G}(0)) = \{\pm i/2\}$, showing that condition (ii) in part (5) of Remark 3.2 does not hold.

REFERENCES

[1] P.A. Cook, *Controllers with universal tracking properties*, in Proceedings of the International IMA Conference on Control: Modelling, Computation, Information, Manchester, 1992.

[2] E.J. Davison, *Multivariable tuning regulators: The feedforward and robust control of a general servomechanism problem*, IEEE Trans. Automat. Control, 21 (1976), pp. 35–47.

[3] R.F. Curtain and G. Weiss, *Well-posedness of triples of operators in the sense of linear systems theory*, in Control and Estimation of Distributed Parameter System, F. Kappel, K. Kunisch, and W. Schappacher, eds., Birkhäuser–Verlag, Basel, 1989, pp. 41–59.

[4] A. Ilchmann and H. Logemann, *High-gain adaptive stabilization of multivariable systems revisited*, Systems Control Lett., 18 (1992), pp. 355–364.
[5] H. Logemann, E.P. Ryan, and S. Townley, *Integral control of linear systems with actuator nonlinearities: Lower bounds for the maximal regulating gain*, IEEE Trans. Automat. Control, 44 (1999), pp. 1315–1319.
[6] H. Logemann and S. Townley, *Low-gain control of uncertain regular linear systems*, SIAM J. Control Opt., 35 (1997), pp. 78–116.
[7] J. Lunze, *Robust Multivariable Feedback Control*, Prentice–Hall, London, 1988.
[8] D.E. Miller and E.J. Davison, *An adaptive tracking problem with a control input constraint*, Automatica J. IFAC, 29 (1993), pp. 877–887.
[9] D.E. Miller and E.J. Davison, *The self-tuning robust servomechanism problem*, IEEE Trans. Automat. Control, 34 (1989), pp. 511–523.
[10] M. Morari, *Robust stability of systems with integral control*, IEEE Trans. Automat. Control, 30 (1985), pp. 574–577.
[11] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer–Verlag, New York, 1983.
[12] W. Rudin, *Real and Complex Analysis*, 2nd ed., Tata McGraw–Hill, New Delhi, 1974.
[13] W. Rudin, *Functional Analysis*, Tata McGraw–Hill, New Delhi, 1974.
[14] D. Salamon, *Realization theory in Hilbert space*, Math. Systems Theory, 21 (1989), pp. 147–164.
[15] D. Salamon, *Infinite-dimensional linear systems with unbounded control and observation: A functional analytic approach*, Trans. Amer. Math. Soc., 300 (1987), pp. 383–431.
[16] O.J. Staffans, *Well-Posed Linear Systems*, manuscript, 2001; also available online from http://www.abo.fi/~staffans/.
[17] O.J. Staffans *J-energy preserving well-posed linear systems*, Int. J. Appl. Math. Comput. Sci., 11 (2001), pp. 1361–1378.
[18] O.J. Staffans, *Quadratic optimal control of stable well-posed linear systems*, Trans. Amer. Math. Soc., 349 (1997), pp. 3679–3715.
[19] O.J. Staffans and G. Weiss, *Transfer functions of regular linear systems* II. *The system operator and the Lax-Phillips semigroup*, Trans. Amer. Math. Soc., 354 (2002), pp. 3229–3262.
[20] G. Weiss, *Transfer functions of regular linear systems* I. *Characterization of regularity*, Trans. Amer. Math. Soc., 342 (1994), pp. 827–854.
[21] G. Weiss, *The representation of regular linear systems on Hilbert spaces*, in Control and Estimation of Distributed Parameter System, F. Kappel, K. Kunisch, and W. Schappacher, eds., Birkhäuser–Verlag, Basel, 1989, pp. 401–416.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 6, pp. 1733–1747

OPTIMAL SHAPE CONTROL PROBLEM FOR THE NAVIER–STOKES EQUATIONS*

BUI AN TON†

Abstract. Optimal control techniques are used to find approximate solutions of an inverse problem for a plane nonstationary flow. The shape of the region, the inflow, and the outflow are determined from partial measurements of the flow in a fixed subdomain.

Key words. shape control, optimal design, open loop, multicontrols, inverse problem, Navier–Stokes

AMS subject classifications. 35K55, 49K20, 49N20, 49N50, 76D05

PII. S0363012901391287

1. Introduction.

In this paper, we shall apply optimal control techniques to find an approximate solution to an inverse problem for nonstationary plane Navier–Stokes equations. One wishes to determine the optimal shape and the boundary velocity controls from partial measurements of the velocity in a subregion, with the flow having a minimum drag.

Pioneering work on optimal shape designs for the Navier–Stokes equations was done by Pironneau [13], where a minimum drag profile submerged in a homogeneous, steady, viscous fluid was obtained using optimal control techniques. In [4], Gunzburger and Kim studied a two-dimensional channel flow of an incompressible, stationary, viscous flow to determine the shape of a bump on a part of the boundary that minimizes the viscous drag. Optimal shape control problems associated with the Navier–Stokes equations may have wide applications to aerodynamic and hydrodynamic problems. An application of optimal shape theory in fluid mechanics to the design of riblets as a drag reduction device can be found in Armugan and Pironneau [1].

Optimal control techniques have been used to approximate solutions of inverse problems. The approach has been developed by Chavent [3], James and Sepulveda [6], Lenhart, Protopopescu, and Yong [8], [9], [10].

In all the cited works, a single control problem was considered as a single cost function was involved, even in the case of several controls. It is known that for multicontrols problems, open and closed loops are two different notions. In [15], we have established the existence of an open loop for a general class of evolution inclusions.

In this paper, we shall apply the multicontrol open loop technique to find an approximate solution to an inverse problem for a nonstationary two-dimensional channel flow of an incompressible, viscous fluid. One wants to find the shape of a bump on a part of the boundary, the inflow and the outflow velocity, from partial measurements of the velocity in a fixed subdomain. The work of Gunzburger and Kim [4] on the optimal shape for the stationary Navier–Stokes equations is extended to the time-dependent case.

*Received by the editors June 22, 2001; accepted for publication (in revised form) July 22, 2002; published electronically February 4, 2003.
http://www.siam.org/journals/sicon/41-6/39128.html
†Department of Mathematics, University of British Columbia, Vancouver, BC, V6T 1Z2, Canada (bui@math.ubc.ca).

2. Setting of the problem. We consider the two-dimensional incompressible flow of a viscous fluid passing through a channel having a finite depth. Let $\mathbf{v}^2, \mathbf{v}^3$ be the velocities at the inflow Γ_1 and at the outflow Γ_2 of the channel, respectively, with

$$\Gamma_1 = \{(\psi,\eta):\ \psi = -2; \eta \in [-2,0]\}, \qquad \Gamma_2 = \{(\psi,\eta):\ \psi = 2,\ \eta \in [-2,0]\},$$
$$\Gamma = \Gamma_1 \cup \Gamma_2.$$

Along the top and the bottom sides Γ_3, $\Gamma(u_1)$ of the channel, the velocity vanishes with

$$\Gamma_3 = \{(\psi,\eta) : |\ \psi\ | \leq 2,\ \eta = 0\},$$

and

$$\Gamma(u^1) = \{(\psi,\ \eta):\ \eta = u^1(\psi),\ |\ \psi\ | \leq 2,\ \eta \in [-2,-1],\ u^1 \in U_1\}$$

represents the bump, which is to be determined.

The domain bounded by the curves $\Gamma, \Gamma(u^1)$ is $Q(u^1)$.

We denote by $\mathcal{U}_1$ the following compact subset of $L^2(I)$:

$$\begin{aligned}\mathcal{U}_1 = \{u^1:\ &\|u^1\|_{H^2(I)} \leq C,\ -2 \leq u^1(\psi) \leq -1,\ \forall \psi \in I\\ &I = [-2,2],\ u^1(\pm 1) = -2,\ u^1_\psi(\pm 1) = 0,\ u^1 = -2\\ &\text{for } 1 \leq |\ \psi\ | \leq 2\}.\end{aligned} \tag{2.1}$$

Thus, u^1 is in $C^1(I)$ and there is no excessive oscillation on the bump $\Gamma(u^1)$.

For the inflow and outflow $\mathbf{v}^2, \mathbf{v}^3$, we shall assume that $\mathbf{v}^j$ belongs to the set $\mathcal{U}_j$ with

$$\mathcal{U}_j = \Big\{\mathbf{u}^j:\ \mathbf{u}^j = (u^j_1, u^j_2),\ \|\mathbf{u}^j\|_{H^2(\Gamma_{j-1})} \leq C,$$
$$\int_{\Gamma_j} \mathbf{u}^j.\mathbf{n} d\sigma = 0,\ \text{support } \mathbf{u}^j \subset \Gamma_j\Big\}. \tag{2.2}$$

Let $\mathcal{H}(Q(u^1))$ be the $L^2(Q(u^1))$-closure of the set

$$\{\mathbf{y}:\ \mathbf{y} = (y_1, y_2),\ \mathbf{y} \in C_0^\infty(Q(u^1)),\ \text{div}(\mathbf{y}) = 0 \text{ in } Q(u^1)\},$$

and $\mathcal{H}^k(Q(u^1))$ is the space

$$\{\mathbf{y}:\ \mathbf{y} \in H^k(Q(u^1)),\ \text{div}(\mathbf{y}) = 0 \text{ in } Q(u^1),\ \mathbf{y} = 0 \text{ on } \Gamma_3 \cup \Gamma(u^1)\}.$$

Set

$$\mathcal{H}^k_0(Q(u^1)) = \{\mathbf{y}:\ \mathbf{y} \in H^k_0(Q(u^1)) \cap \mathcal{H}(Q(u^1))\}.$$

We shall write $\mathbf{v}$ for $(\mathbf{v}^2, \mathbf{v}^3)$. We consider, for each $\{u^1, \mathbf{v}\} \in U$, the nonstationary Navier–Stokes equations:

$$\begin{aligned}\mathbf{y}' - \nu\Delta\mathbf{y} + (\mathbf{y}.\nabla)\mathbf{y} + \nabla p &= \mathbf{f} \text{ in } Q(u^1) \times (0,T),\\ \nabla.\mathbf{y} &= 0 \text{ in } Q(u^1) \times (0,T),\\ \mathbf{y}(.,0) &= \mathbf{y}_0 \text{ in } Q(u^1),\\ \mathbf{y} = \mathbf{v}^j \text{ on } \Gamma_{j-1} \times (0,T),\ \mathbf{y} &= 0 \text{ on } \Gamma(u^1) \cup \Gamma_3,\ j = 2,3.\end{aligned} \tag{2.3}$$

DEFINITION 2.1. *The vector function* $\mathbf{y}$ *is said to be a weak solution of* (2.3) *with controls* $\{u^1, \mathbf{v}\}$ *if*

- $\{\mathbf{y},\, \mathbf{y}'\} \in L^2(0,T;\mathcal{H}^1(Q(u^1)) \cap L^\infty(0,T;\mathcal{H}(Q(u^1))) \times L^2(0,T;\mathcal{H}(Q(u^1))^*)$,
- $\mathbf{y}$ *satisfies* (2.3).

Let

$$\Omega \subset Q(u^1)\ \ \forall u^1 \in \mathcal{U}_1;\ \ \mathbf{h},\mathbf{k} \in L^2(0,T;L^2(\Omega)) \times L^2(0,T;L^2(\omega)).$$

The vector functions $\mathbf{h},\mathbf{k}$ represent the experimental measurements of the velocity of the fluid in the subregion Ω and the measurement of the flow at

$$\omega = \{(\psi,\eta):\ \psi = 3/2,\ \eta \in [-1,0]\}.$$

We shall associate with (2.3) the cost functionals

$$J_1(\mathbf{y};u^1;\mathbf{v}) = \int_0^T \int_\Omega \mid \mathbf{y}(\,,t) - \mathbf{h}(\,.,t) \mid^2 d\psi d\eta dt \tag{2.4}$$

and

$$J_2(\mathbf{y};\, u^1;\, \mathbf{v}) = \int_0^T \int_\omega \mid \mathbf{y}(\,.,t) - \mathbf{k}(\,.,t) \mid^2 d\eta dt. \tag{2.5}$$

Let $D(\mathbf{y})$ be the deformation tensor of the flow $\mathbf{y}$; it represents the rate of energy dissipation due to the deformation and is given by

$$D(\mathbf{y}) = \frac{1}{2}(\nabla\mathbf{y} + (\nabla\mathbf{y})^t).$$

We shall associate with (2.3) the cost functional

$$J(\mathbf{y};u^1\,;\mathbf{v}) = \int_0^T \int_{Q(u^1)} (D(\mathbf{y}))^2 d\psi d\eta dt + J_1(\mathbf{y};\, u^1;\, \mathbf{v}) + J_2(\mathbf{y};\, u^1;\, \mathbf{v}). \tag{2.6}$$

The aim of this paper is to show the existence of an open loop of (2.3)–(2.5) and to establish the existence of an optimal control of (2.3)–(2.6), thereby extending Gunzburger and Kim's result.

DEFINITION 2.2. *A control* $\vec{u}_* = (u^1_*,\, \mathbf{u}_*)$ *is said to be an open loop control of* (2.3)–(2.5) *if*

- *there exists* $\tilde{\mathbf{y}}$*, weak solution of* (2.3) *in* $Q(u^1_*)$ *with inflow* $\mathbf{u}^2_*$ *and outflow* $\mathbf{u}^3_*$*, and*
- $$\begin{aligned} &J_1(\tilde{\mathbf{y}};\, u^1_*;\, \mathbf{u}_*) \le J_1(\mathbf{y};u^1_*;\, \mathbf{v}) \qquad \forall \mathbf{v} \in \mathcal{U}_2 \times \mathcal{U}_3, \\ &J_2(\tilde{\mathbf{y}};\, u^1_*;\, \mathbf{u}_*) \le J_2(\mathbf{x};\, u^1;\, \mathbf{u}_*) \qquad \forall u^1 \in \mathcal{U}_1. \end{aligned} \tag{2.7}$$

  $\mathbf{y}$ *is the solution of* (2.3) *with controls* $\{u^1_*,\, \mathbf{v}\}$*. Similarly for* $\mathbf{x}$*.*

We denote by $\tilde{Q}$ the rectangle with vertices at $(-2,2)$, $(-2,0)$, $(2,0)$, $(2,-2)$. We shall now state the main result of the paper.

THEOREM 2.1. *Let* $\{\mathbf{f},\, \mathbf{y}_0, \mathbf{h}, \mathbf{k}\}$ *be in*

$$L^2(0,T;L^2(\tilde{Q})) \times (\tilde{Q}) \times L^2(0,T;L^2(\Omega)) \times L^2(0,T;L^2(\omega))$$

and let J_1, J_2 *be as in* (2.4)–(2.5)*. Then there exists an open loop control* $\{\tilde{u}^1,\, \tilde{\mathbf{u}}\} \in \mathcal{U}$ *of* (2.3)–(2.5).

For the optimal control problem (2.3), (2.6) we have the following.

THEOREM 2.2. *Suppose all the hypotheses of Theorem* 2.1 *are satisfied. Then there exists*

- *a control* $\{\hat{u}^1, \hat{\mathbf{u}}\} \in \mathcal{U}$,
- *a weak solution* $\hat{\mathbf{y}}$ *of* (2.3), *in the sense of Definition* 2.1, *with controls* $\{\hat{u}^1, \hat{\mathbf{u}}\}$ *such that*

$$J(\hat{\mathbf{y}};\, \hat{u}^1;\, \hat{\mathbf{u}}) \le J(\mathbf{x};\, v^1;\, \mathbf{v}) \qquad \forall \{v^1,\, \mathbf{v})\} \in \mathcal{U}.$$

Consider the initial boundary-value problem for the Stokes equations:

$$\begin{gathered} \mathbf{w}' - \nu\Delta\mathbf{w} + \nabla p = 0 \ \text{ in } Q(u^1) \times (0,T), \\ \mathbf{w} = \mathbf{v}^j \ \text{ on}\, \Gamma_{j-1},\ \mathbf{w} = 0 \ \text{ on } \partial Q(u^1)/\Gamma_j,\ j = 2,3, \\ \nabla.\mathbf{w} = 0,\ \mathbf{w}(\,.\,,0) = 0 \ \text{ in } Q(u^1). \end{gathered} \tag{2.8}$$

We have the following result.

LEMMA 2.1. *Let* $\mathbf{v} = (\mathbf{v}^2, \mathbf{v}^3)$ *be in* $\mathcal{U}_2 \times \mathcal{U}_3$*; then there exists a unique solution* $\mathbf{w}$ *of* (2.8) *with*

$$\{\mathbf{w}, \mathbf{w}'\} \in L^2(0,T;{}^4(Q(u^1))) \times L^2(0,T;L^2(Q(u^1))).$$

Moreover,

$$\|\mathbf{w}\|^2_{L^2(0,T;{}^4(Q(u^1)))} + \|\mathbf{w}'\|^2_{L^2(0,T;L^2(Q(u^1)))} \le C\{1 + \|\mathbf{v}\|^2_{H^2(\Gamma)}\}.$$

The lemma is an immediate result of the theory of Stokes equations; cf. [14].

Set $\hat{\mathbf{y}} = \mathbf{y} - \mathbf{w}$; then (2.3) becomes

$$\begin{gathered} \hat{\mathbf{y}}' - \nu\Delta\hat{\mathbf{y}} + (\hat{\mathbf{y}}.\nabla)\hat{\mathbf{y}} + L(\hat{\mathbf{y}};u^1;\,\mathbf{v}) + \nabla p = \mathbf{f} \text{ in } Q(u^1) \times (0,T), \\ \nabla.\hat{\mathbf{y}} = 0;\ \hat{\mathbf{y}} = 0 \ \text{ on } \partial Q(u^1) \times (0,T), \\ \hat{\mathbf{y}}(\,.\,,0) = \mathbf{y}_0 \text{ in } Q(u^1) \end{gathered} \tag{2.9}$$

with

$$L(\hat{\mathbf{y}};\mathbf{v};\,u^1) = (\hat{\mathbf{y}}.\nabla)\mathbf{w} + (\mathbf{w}.\nabla)\hat{\mathbf{y}} + (\mathbf{w}.\nabla)\mathbf{w}. \tag{2.10}$$

The extension method, used in [4] and in many optimal design stationary problems, gives rise to difficulties when applied to time-dependent problems. In the case of the Navier–Stokes equations, one needs an estimate on the time derivative of the solution in a *control-free* space. We shall follow Lenhart, Protopopescu, and Yong [8] and transform (2.9) into a problem in a fixed domain by making a change of variable.

Let

$$\zeta = 2\eta/u^1(\psi), \qquad \psi \in I. \tag{2.11}$$

Set

$$\mathbf{y}(\psi,\zeta,t) = \mathbf{Y}(\psi,\eta,t) = \mathbf{Y}(\psi,\zeta u^1/2;\,t) \qquad \forall (t,\,\psi,\zeta) \in [0,T] \times Q, \tag{2.12}$$

where Q is the rectangle with vertices at $(-2,0), (-2,2), (2,0), (2,2)$. A calculation as in [8, p. 946] gives

$$\nabla_{\psi,\eta}\mathbf{Y}(\psi,\eta,t) = U(\psi,\zeta;u^1)\nabla_{\psi,\zeta}\mathbf{y}(\psi,\zeta;t)$$

with

$$U(\psi,\zeta;u^1) = \begin{pmatrix} I & -\zeta u^1_\psi/u^1 \\ 0 & 2/u^1 \end{pmatrix}.$$

Throughout the paper, we shall write U for $U(\psi, \zeta; u^1)$. Furthermore

$$\nabla_{\psi,\eta}.\mathbf{Y} = \nabla_{\psi,\zeta}.(U^t\mathbf{y}) + (U^t\mathbf{y}.\nabla_{\psi,\zeta}u^1)/u^1 \tag{2.13}$$

and, as computed in [8, p. 946],

$$\operatorname{div}(\mathbf{Y}) = (U\nabla).\mathbf{y} = \nabla.U^t\mathbf{y} + (U^t\mathbf{y}.\nabla u^1)/u^1. \tag{2.14}$$

U^t denotes the transpose of U.

In order to operate in a control-free space, we shall consider an approximate system of Cauchy–Kowaleska type, introduced by Lions [11, pp. 466–469]. Consider the system

$$\begin{aligned} &\hat{\mathbf{y}}'_\varepsilon - \nu\nabla(F(u^1)\nabla\hat{\mathbf{y}}_\varepsilon) + (\hat{\mathbf{y}}_\varepsilon.U\nabla)\hat{\mathbf{y}}_\varepsilon + \hat{L}(\hat{\mathbf{y}}_\varepsilon; u^1, \mathbf{v}) + \frac{1}{2}\{(\nabla.U)\hat{\mathbf{y}}_\varepsilon\}\hat{\mathbf{y}}_\varepsilon \\ &\qquad + (\nabla U)p_\varepsilon = \mathbf{f} \text{ in } Q \times (0,T), \\ &\hat{\mathbf{y}}_\varepsilon = 0 \text{ on } \partial Q \times (0,T), \qquad \hat{\mathbf{y}}_\varepsilon(.0) = \mathbf{y}_0 \text{ in } Q, \end{aligned} \tag{2.15}$$

and

$$\varepsilon p'_\varepsilon + (U\nabla).\hat{\mathbf{y}}_\varepsilon = 0 \text{ in } Q \times (0,T);\ p_\varepsilon(.,0) = 0, \tag{2.16}$$

with

$$\hat{L}(\hat{\mathbf{y}};\ \mathbf{w}; u^1) = (\mathbf{w}.u\nabla)\mathbf{w} + (\mathbf{y}.U\nabla)\mathbf{w} + (\mathbf{w}.U\nabla)\mathbf{w} - (F(u^1)\nabla\hat{\mathbf{y}}.\nabla u^1)/u^1. \tag{2.17}$$

We denote by $F(u^1)$ the matrix

$$\begin{pmatrix} 1 & -\zeta u^1_\psi/u^1 \\ 0 & \zeta^2 \mid \nabla u^1 \mid^2 /(u^1)^2 + 4/(u^1)^2 \end{pmatrix}.$$

DEFINITION 2.3. *Let* $\{\mathbf{f}, \mathbf{y}_0, \mathbf{v}^j, u^1\}$ *be as in Theorem* 2.1. *Then* $\{\mathbf{y}_\varepsilon, p_\varepsilon\}$ *is said to be a weak solution of* (2.15)–(2.16) *if*

- $\{\mathbf{y}_\varepsilon, p_\varepsilon\} \in L^2(0,T; H_0^1(Q)) \cap L^\infty(0,T; L^2(Q)) \times L^\infty(0,T; L^2(Q))$,
- $\{D_t^\gamma \mathbf{y}_\varepsilon, D_t^\gamma p_\varepsilon\} \in L^2(0,T; L^2(Q)) \times L^2(0,T; L^2(Q))$, $0 < \gamma < 1/4$, *with*
- $$\begin{aligned} -\int_0^T (\mathbf{y}_\varepsilon, \phi')dt + \nu\int_0^T (F(u^1)\nabla\mathbf{y}_\varepsilon, \phi)dt + \int_0^T ((\mathbf{y}_\varepsilon.U\nabla)\mathbf{y}_\varepsilon, \phi)dt \\ + \int_0^T (\hat{L}(\mathbf{y}_\varepsilon; u^1; \mathbf{w}),\ \phi)dt + \frac{1}{2}\int_0^T (\mathbf{y}_\varepsilon(U\nabla.\mathbf{y}_\varepsilon),\ \phi)dt \\ - \int_0^T (p_\varepsilon, U\nabla.\mathbf{y}_\varepsilon)dt \\ = (\mathbf{y}_0,\ \phi(0)) + \int_0^T (\mathbf{f},\ \phi)dt \end{aligned}$$

and

$$-\varepsilon\int_0^T (p_\varepsilon, q')dt + \int_0^T (u\nabla._\varepsilon,\ q)dt = 0$$

for all

$$\{\phi, q\} \in L^2(0,T; H_0^1(Q)) \times L^2(0,T; L^2(Q)), \quad \{\phi',\ q'\} \in (L^2(0,T; L^2(Q)))^2,$$

and $\phi(.,T) = 0 = q(.,T)$.

We shall show the existence of an open loop for the approximating system (2.15)–(2.16) and Theorem 2.1 is obtained by letting $\varepsilon \to 0$.

3. The approximating system. The main result of the section is the following theorem.

THEOREM 3.1. *Suppose all the hypotheses of Theorem* 2.1 *are satisfied. Then for each* ε, *there exists a weak solution*

$$\{\mathbf{y}_\varepsilon, p_\varepsilon\} \in L^2(0,T;H_0^1(Q)) \cap L^\infty(0,T;L^2(Q)) \times L^\infty(0,T;L^2(Q))$$

of (2.15)–(2.16) *in the sense of Definition* 2.3. *Moreover,*

$$\|\mathbf{y}_\varepsilon\|^2_{L^2(0,T;H_0^1(Q))} + \|\mathbf{y}_\varepsilon\|^2_{L^\infty(0,T;L^2(Q))} + \varepsilon\|p_\varepsilon\|^2_{L^\infty(0,T;L^2(Q))}$$

$$\le C\{1 + \|\mathbf{f}\|^2_{L^2(0,T;L^2(Q))} + \|\mathbf{y}_0\|^2_{L^2(Q)}\exp(\|\mathbf{v}\|^2_{H^2(\Gamma)} + \|u^1\|^2_{H^2(I)})\}.$$

The constant C *is independent of* $\varepsilon, \mathbf{v}, u^1$. *Furthermore*

$$\|D_t^\gamma \mathbf{y}_\varepsilon\|^2_{L^2(0,T;L^2(Q))} + \|\varepsilon\|D_t^\gamma p_\varepsilon\|^2_{L^2(0,T;L^2(Q))} \le C, \qquad 0 < \gamma < \frac{1}{4}.$$

We shall follow Lions's proof [11, pp. 466–469] and make the necessary modifications using Lenhart, Protopopescu, and Yong's [8] estimates.

LEMMA 3.1. *Suppose all the hypotheses of Theorem* 3.1 *are satisfied. Then there exists a positive constant* c, *independent of* u^1, $\mathbf{v}$ *such that*

$$(F(u^1)\nabla\mathbf{y}, \nabla\mathbf{y}) = \int_Q F(u^1) \mid \nabla\mathbf{y} \mid^2 d\psi d\zeta \ge c\|\mathbf{y}\|^2_{H_0^1(Q)} \qquad \forall \mathbf{y} \in H_0^1(Q).$$

Proof. In [8, p. 952], it was shown that

$$\begin{aligned}(F(u^1)\nabla\mathbf{y}, \nabla\mathbf{y}) &\ge c\int_Q \mid U\nabla\mathbf{y} \mid^2 d\psi d\zeta \\ &\ge \int_Q \{\mid \mathbf{y}_\psi - \zeta\mathbf{y}_\zeta u^1_\psi/u^1 \mid^2 + \mid \mathbf{y}_\zeta \mid^2\} d\psi d\zeta \\ &\ge \int_Q \mid \mathbf{y}_\zeta \mid^2 d\psi d\zeta \qquad \forall \mathbf{y} \in H_0^1(Q).\end{aligned}$$

On the other hand,

$$\begin{aligned}\|\mathbf{y}_\psi\|_{L^2(Q)} &\le \|\mathbf{y}_\psi - \zeta\mathbf{y}_\zeta u^1_\psi/u^1\|_{L^2(Q)} + \|\zeta u^1_\psi \mathbf{y}_\zeta/u^1\|_{L^2(Q)} \\ &\le C\{1 + \|u^1\|_{H^{1,\infty}(I)}\|U\nabla\mathbf{y}\|_{L^2(Q)}\} \\ &\le C\{1 + \|u^1\|_{H^2(I)}\|U\nabla\mathbf{y}\|_{L^2(Q)}\}.\end{aligned}$$

The lemma is proved.

LEMMA 3.2. *Suppose all the hypotheses of Theorem* 3.1 *are satisfied. Let* $\hat{L}(\mathbf{y}, u^1, \mathbf{w})$ *be as in* (2.17). *Then*

$$\|\hat{L}(\mathbf{y}; u^1; \mathbf{w})\|_{L^2(Q)} \le C\{1 + \|u^1\|_{H^2(I)} + \|\mathbf{v}^j\|_{H^2(\Gamma)}\}^2(1 + \|\mathbf{y}\|_{H^1(Q)}) \quad \forall \mathbf{y} \in H_0^1(Q).$$

The constant C *is independent of* u^1, $\mathbf{v}$.

Proof. The proof is trivial and we shall not reproduce it.

LEMMA 3.3. *Let* $U, \mathbf{y}$ *be as in Theorem* 3.1; *then*

$$((\mathbf{y}.U\nabla)\mathbf{y}, \mathbf{y}) = -\frac{1}{2}(\mathbf{y},\ \mathbf{y}(U\nabla.\mathbf{y})). \tag{3.1}$$

The inner product in $L^2(Q)$ *is denoted by* $(\ .,\ .)$.

Proof. A simple but lengthy calculation using (2.14) gives the stated result.

To prove Theorem 3.1, we shall use the Faedo–Galerkin method and establish the existence of $\{\mathbf{y}_\varepsilon, p_\varepsilon\}$, a weak solution of (2.15)–(2.16) in the sense of Definition 2.3.

LEMMA 3.4. *Let* $\{\mathbf{y}_\varepsilon^n,\ p_\varepsilon^n\}$ *be the approximate solution of* (2.15)–(2.16) *obtained from the Galerkin method. Then*

$$\|\mathbf{y}_\varepsilon^n\|^2_{L^2(0,T;H_0^1(Q))} + \|\mathbf{y}_\varepsilon^n\|^2_{L^\infty(0,T;L^2(Q))} + \varepsilon\|p_\varepsilon^n\|^2_{L^\infty(0,T;L^2(Q))}$$

$$\leq C\{1 + \|\mathbf{f}\|^2_{L^2(0,T;L^2(Q))}\} + \|\mathbf{y}_0\|^2_{L^2(Q)}\exp(1 + \|u^1\|_{H^2(I)} + \|\mathbf{v}\|^2_{H^2(\Gamma)}).$$

The constant C *is independent of* $\varepsilon, n, u^1,\ \mathbf{v}^j$.

Proof. With (2.14) and with Lemmas 3.1 and 3.2, the proof is almost the same as the one given in [11, pp. 466–469]. We shall not reproduce it.

LEMMA 3.5. *Let* $\{\mathbf{y}_\varepsilon^n,\ p_\varepsilon^n\}$ *be as in Lemma* 3.4. *Then*

$$\|D_t^\gamma \mathbf{y}_\varepsilon^n\|_{L^2(0,T;L^2(Q))} + \|D_t^\gamma p_\varepsilon^n\|_{L^2(0,T;L^2(Q))} \leq C.$$

The constant C *is independent of* $n,\ \varepsilon, u^1,\ \mathbf{v}$ *and* γ *is any number in* $(0, 1/4)$.

Proof. The proof is as that done in [11, pp. 466–469]. The changes are minor.

Proof of Theorem 3.1. Let $\{\mathbf{y}^n, p^n\}$ be as in Lemmas 3.4 and 3.5. From the estimates of the lemmas, we obtain, by taking subsequences if necessary,

$$\{\mathbf{y}^n, D_t^\gamma \mathbf{y}^n, p^n, D_t^\gamma p^n\} \to \{\mathbf{y}, D^\gamma \mathbf{y},\ y, D_t^\gamma p\}$$

in

$$(L^2(0,T;H_0^1(Q)))_{weak} \cap (L^\infty(0,T;L^2(Q)))_{weak^*} \times (L^2(0,T:L^2(Q)))_{weak}$$
$$\times (L^\infty(0,T;L^2(Q)))_{weak^*} \times (L^2(0,T;L^2(Q)))_{weak}.$$

Since the injection mapping of $H^1(Q)$ onto $L^4(Q)$ is compact, it follows from the above estimates and from [11, p. 61, Theorem 5.2] that

$$\mathbf{y}^n \to \mathbf{y} \text{ in } L^2(0,T;L^4(Q)) \text{ and a.e.}$$

Now a standard proof (e.g., [11, pp. 78–79]) shows that $\{\mathbf{y}, p\}$ is a solution of (2.14)–(2.15). The estimates of the theorem are direct consequences of those of Lemmas 3.4 and 3.5. The theorem is proved.

4. The equation (2.3) in a fixed domain. In this section, we shall study the problem (2.3) in the fixed domain $Q \times (0,T)$. Consider the initial boundary-value problem

$$\begin{aligned} &\mathbf{y}' - \nu\nabla.(F(u^1)\nabla\mathbf{y}) + (\mathbf{y}.U\nabla)\mathbf{y} + U\nabla p + \hat{L}(\mathbf{y}; u^1; \mathbf{w}) = \mathbf{y} \text{ in } Q \times (0,T), \\ &U\nabla.\mathbf{y} = 0 \text{ in } Q \times (0,T), \qquad \mathbf{y} = 0 \text{ on } \partial Q \times (0,T), \\ &\mathbf{y}(\ .,0) = \mathbf{y}_0 \text{ in } Q. \end{aligned} \tag{4.1}$$

DEFINITION 4.1. *The vector function* $\mathbf{y} \in L^2(0,T;H_0^1(Q)) \cap L^\infty(0,T;L^2(Q))$ *with* $D_t^\gamma \mathbf{y} \in L^2(0,T;L^2(Q))$ *is said to be a weak solution of* (4.1) *if*

$$-\int_0^T (\mathbf{y},\phi')dt + \nu\int_0^T (F(u^1)\nabla\mathbf{y},\nabla\phi)dt + \int_0^T ((\mathbf{y}.U\nabla)\mathbf{y},\phi)dt + \int_0^T (\hat{L}(\mathbf{y};u^1;\mathbf{w}),\phi)dt = (\mathbf{y}_0,\phi(\,.,0)) + \int_0^T (\mathbf{f},\phi)dt$$

for all

$$\phi \in L^2(0,T;H_0^1(Q)\cap H^2(Q)),\quad \phi' \in L^2(0,T;L^2(Q)),\quad \phi(\,.,T)=0,\quad U\nabla.\phi = 0.$$

The main result of the section is the following theorem.

THEOREM 4.1. *Suppose all the hypotheses of Theorem* 2.1 *are satisfied. Then there exists a weak solution of* (4.1) *in the sense of Definition* 4.1. *Moreover,*

$$\|\mathbf{y}\|^2_{L^2(0,T;H_0^1(Q))} + \|\mathbf{y}\|^2_{L^\infty(0,T;L^2(Q))} + \|D_t^\gamma\mathbf{y}\|^2_{L^2(0,T;L^2(Q))} \le C\{1+\|\mathbf{f}\|^2_{L^2(0,T;L^2(Q))}\} + C\|\mathbf{y}_0\|^2_{L^2(Q)}\exp(\|u^1\|^2_{H^2(I)} + \|\mathbf{v}^j\|^2_{H^2(\Gamma)})$$

with $0<\gamma<1/4$. *The constant* C *is independent of* $u^1,\mathbf{v}$.

Proof. (1) Let $\{\mathbf{y}_\varepsilon,\, p_\varepsilon\}$ be as in Theorem 3.1. From the estimates of the theorem, we obtain, by taking subsequences,

$$\{\mathbf{y}_\varepsilon,\, D_t^\gamma\mathbf{y}_\varepsilon,\, \varepsilon p_\varepsilon\} \to \{\mathbf{y},\, D_t^\gamma\mathbf{y}, 0\}$$

in

$$(L^2(0,T;H_0^1(Q)))_{weak} \cap (L^\infty(0,T;L^2(Q)))_{weak^*} \times (L^2(0,T;L^2(Q)))_{weak} \times L^2(0,T;L^2(Q)).$$

With our estimate on the fractional time derivative of $\mathbf{y}_\varepsilon$, it follows from the Sobolev imbedding theorem and from [11, p. 61, Theorem 5.2] that

$$\mathbf{y}_\varepsilon \to \mathbf{y} \text{ in } L^2(0,T;L^4(Q)) \text{ and a.e.}$$

The estimates of the theorem are direct consequences of those of Theorem 3.1.

(2) Since

$$\sqrt{\varepsilon}p_\varepsilon \to 0 \text{ in } (L^\infty(0,T;L^2(Q)))_{weak^*},$$

we have $\varepsilon p_\varepsilon' \to 0$ in the distribution sense in $Q\times(0,T)$. Hence

$$U\nabla.\mathbf{y}_\varepsilon \to U.\nabla\mathbf{y} \text{ in } (L^2(0,T;L^2(Q)))_{weak}.$$

Thus, $U\nabla.\mathbf{y} = 0$ in $Q\times(0,T)$.

(3) Since Q is a bounded subset of the plane, an application of the Sobolev imbedding theorem yields

$$\|\mathbf{y}_\varepsilon U\nabla.\mathbf{y}_\varepsilon\|_{H^{-1}(Q)} \le C\|\mathbf{y}_\varepsilon\|^{\frac12}_{L^2(0,T;L^2(Q))}\|\mathbf{y}_\varepsilon\|^{\frac32}_{H_0^1(Q)}.$$

Hence

$$\|\mathbf{y}_\varepsilon U\nabla.\mathbf{y}_\varepsilon\|_{L^{\frac43}(0,T;H^{-1}(Q))} \le C\{1+\|\mathbf{y}_\varepsilon\|^2_{L^\infty(0,T;L^2(Q))}\}\|\mathbf{y}_\varepsilon\|^2_{L^{\frac43}(0,T;H_0^1(Q))} \le C.$$

There exists a subsequence such that

$$\mathbf{y}_\varepsilon(U\nabla.\mathbf{y}_\varepsilon) \to \Psi \text{ in } (L^{\frac{4}{3}}(0,T;H^{-1}(Q)))_{weak},$$

and from part 2, we deduce that $\Psi = 0$.

(4) A similar argument shows that

$$(\mathbf{y}_\varepsilon.U\nabla)\mathbf{y}_\varepsilon \to (\mathbf{y}.U\nabla)\mathbf{y} \text{ in } (L^{\frac{4}{3}}(0,T;H^{-1}(Q)))_{weak}.$$

It is now trivial to check that $\mathbf{y}$ is a solution of (4.1) in the sense of Definition 4.1.

LEMMA 4.1. *The weak solution* $\mathbf{y}$ *of* (4.1)*, given by Theorem* 4.1*, is unique.*

Proof. (1) we have, by a simple calculation,

$$\begin{aligned}((\mathbf{y}.U\nabla),\,\phi) &= \sum_{j,k=1}^{2}(y_j(U\nabla)_j y_k,\,\phi_k)\\ &= -\sum_{k=1}^{2}(y_k,\phi_k\{\nabla.U^t\mathbf{y} + (U^t\mathbf{y}.\nabla u^1)/u^1\}) - ((\mathbf{y}.U\nabla)\phi,\mathbf{y})\\ &= -(\mathbf{y}U\nabla.\mathbf{y},\phi) - ((\mathbf{y}.U\nabla)\phi,\mathbf{y}) = -((\mathbf{y}.U\nabla)\phi,\,\mathbf{y}).\end{aligned}$$

Thus it follows from the Sobolev imbedding theorem that

$$\begin{aligned}\mid((\mathbf{y}.U\nabla)\mathbf{y},\phi)\mid &\le C\|\mathbf{y}\|^2_{L^4(Q)}\|\phi\|_{H^1_0(Q)}\\ &\le C\|\mathbf{y}\|_{L^2(Q)}\|\mathbf{y}\|_{H^1_0(Q)}\|\phi\|_{H^1_0(Q)}.\end{aligned}$$

(2) Let (u^1) be the completion of the set

$$\{\phi:\ \phi\in C_0^\infty(Q),\ U\nabla.\phi = 0\}$$

in the $H^1(Q)$-norm. Then we deduce from the above that

$$\|(\mathbf{y}.U\nabla)\mathbf{y}\|_{L^2(0,T;(\mathcal{V}(u^1))^*)} \le C\|\mathbf{y}\|_{L^\infty(0,T;L^2(Q))}\|\mathbf{y}\|_{L^2(0,T;H^1_0(Q))}.$$

It follows from (4.1) that $\mathbf{y}'$ exists and is in $L^2(0,T;((u^1))^*)$.

(3) Suppose that $\mathbf{x},\mathbf{y}$ are two solutions of (4.1), given by Theorem 4.1; then

$$\begin{aligned}(\mathbf{y}'-\mathbf{x}',\mathbf{y}-\mathbf{x}) &+ \nu(F(u^1)\nabla(\mathbf{y}-\mathbf{x}),\,\nabla(\mathbf{y}-\mathbf{x})) + ((\mathbf{y}-\mathbf{x}).U\nabla\mathbf{y},\mathbf{y}-\mathbf{x})\\ &+ (\mathbf{x}.U\nabla(\mathbf{y}-\mathbf{x}),\mathbf{y}-\mathbf{x}) + (\mathbf{y}-\mathbf{x}.U\nabla\mathbf{w},\mathbf{y}-\mathbf{x})\\ &+ (\mathbf{w}.U\nabla(\mathbf{y}-\mathbf{x}),\mathbf{y}-\mathbf{x}) = 0.\end{aligned}$$

We deduce that

$$\frac{d}{dt}\|\mathbf{y}-\mathbf{x}\|^2_{L^2(Q)} \le C\|\mathbf{y}-\mathbf{x}\|^2_{L^2(Q)}.$$

Hence $\mathbf{y}-\mathbf{x} = 0$. The lemma is proved.

5. Optimal control. The main results of the paper, which are Theorem 2.1 and Theorem 2.2, will be proved in this section. Let $\mathbf{y}$ be the solution of (4.1) with controls $\{u^1, \mathbf{v}\}$ and let

$$J_1(\mathbf{y}; u^1; \mathbf{v}) = \int_0^T \int_\Omega \mid \mathbf{Y}(\psi, 2\eta/u^1; t) - \mathbf{h}(\psi, \eta; t) \mid^2 d\psi d\eta dt, \tag{5.1}$$

$$J_2(\mathbf{y}; u^1; \mathbf{v}) = \int_0^T \int_\omega \mid \mathbf{Y}(3/2; 2\eta/u^1(3/2); t) - \mathbf{k}(\eta, t) \mid^2 d\eta dt$$

with

$$\mathbf{Y}(\psi, \eta; t) = \mathbf{y}(\psi, \zeta; t) + \mathbf{w}. \tag{5.2}$$

Set

$$\Psi(\vec{u}, \vec{v}) = J_1(; \mathbf{u}; v^1) + J_2(\mathbf{x}; u^1; \mathbf{v}),$$

where $\vec{u} = (u^1, \mathbf{u})$, $\vec{v} = (v^1, \mathbf{v})$ are in $\mathcal{U}$ and $\mathbf{x}$ is the unique solution of (4.1) with controls $\{v^1, \mathbf{u}\}$.

LEMMA 5.1. *Suppose all the hypotheses of Theorem* 4.1 *are satisfied. Then for each given* $\vec{u} = \{u^1, \mathbf{u}\}$ *in* $\mathcal{U}$, *there exists* $\vec{v}_* \in \mathcal{U}$ *such that*

$$\Psi(\vec{u}; \vec{v}_*) = d(u) = \inf\{\Psi(\vec{u}, \vec{v}) : \ \vec{v} \in \mathcal{U}\}. \tag{5.3}$$

Proof. First we note that the infimum exists. Let $\{\vec{v}_n\}$ be a minimizing sequence of the optimization problem (5.3) with

$$d(\vec{u}) \leq \Psi(\vec{u}; \vec{v}_n) \leq d(\vec{u}) + \frac{1}{n}.$$

Since $\{\vec{v}_n\}$ is in $\mathcal{U}$ and $\mathcal{U}$ is a compact subset of $H^1(I) \times H^1(\Gamma)$, we get, by taking subsequences,

$$\vec{v}_n \to \vec{v}_* \text{ in } H^1(I) \cap (H^2(I))_{weak} \times H^1(\Gamma) \cap (H^2(\Gamma))_{weak}$$

with $\vec{v}_* \in$. Let $\{\mathbf{y}_n, \mathbf{x}_n\}$ be the solutions of (4.1) with controls $\{u^1, \mathbf{v}_n\}$, $\{v_n^1, \mathbf{u}\}$, respectively. It follows from Theorem 4.1 that

$$\|\{\mathbf{y}_n, \mathbf{x}_n\}\|_{L^2(0,T;H_0^1(Q))} + \|\{\mathbf{y}_n, \mathbf{x}_n\}\|_{L^\infty(0,T;L^2(Q))} + \|\{D_t^\gamma \mathbf{y}_n, D_t^\gamma \mathbf{x}_n\}\|_{L^2(0,T;L^2(Q))} \leq C,$$

where C is independent of n.

Since the injection mapping of $H_0^1(Q)$ into $L^4(Q)$ is compact, it follows from the above estimates and from [11, p. 62, Theorem 5.2] that

$$\{\mathbf{y}_n, \mathbf{x}_n\} \to \{\mathbf{y}_*, \mathbf{x}_*\}$$
$$\text{in } (L^2(0,T; H_0^1(Q)))_{weak} \cap L^2(0,T; L^2(Q)) \cap (L^\infty(0,T; L^2(Q)))_{weak^*}$$

and a.e. with

$$\{\mathbf{y}_n, \mathbf{x}_n\} \mid_{\psi=3/2} \to \{\mathbf{y}_*, \mathbf{x}_*\} \mid_{\psi=3/2} \text{ in } L^2(0,T; L^2(0,2)).$$

Now we consider a typical term of (4.1). We have

$$\begin{aligned} &\|\mathbf{x}_n . U(v_n^1)\mathbf{x}_n\|_{L^{\frac{4}{3}}(0,T;H^{-1}(Q))} \\ &\quad \leq C\|U(v_n^1)\|_{L^\infty(Q)}(1 + \|\mathbf{x}_n\|^2_{L^\infty(0,T;L^2(Q))}\|\mathbf{x}_n\|^2_{L^{\frac{4}{3}}(0,T;H_0^1(Q))}) \\ &\quad \leq C. \end{aligned}$$

Since

$$\mathbf{x}_n.U(v_n^1)\nabla\mathbf{x}_n \to \mathbf{x}_*.U(v_*^1)\nabla\mathbf{x}_* \text{ a.e. in } Q\times(0,T),$$

it follows from the above estimate that

$$\mathbf{x}_n.U(v_n^1)\nabla\mathbf{x}_n \to \mathbf{x}_*.U(v_*^1)\nabla\mathbf{x}_* \text{ in } (L^{\frac{4}{3}}(0,T;H^{-1}(Q)))_{weak}.$$

It is now easy to check that $\{\mathbf{y}_*, \mathbf{x}_*\}$ is the unique solution of (4.1) with the corresponding controls. Moreover,

$$J_1(\mathbf{y}_*;\mathbf{u};v_*^1)+J_2(\mathbf{x};\mathbf{v}_*;u^1)\le \liminf \Psi(\vec{u};\vec{v}_n)\le d(\vec{u}).$$

It follows from the definition of $d(\vec{u})$ that

$$d(\vec{u})=\Psi(\vec{u};\vec{v}_*).$$

The lemma is proved.

Let

$$S=\{\vec{v}_*:\ \vec{v}_* \text{ as in (5.3)}\}.$$

LEMMA 5.2. *Suppose all the hypotheses of Theorem* 4.1 *are satisfied. Let* g_1, g_j^k *be weakly continuous functions from* $\mathcal{U}_1$, $_j^k$; $j=2,3,k=1,2$ *into* R^+, *and suppose that* g_1, g_j^k *are one-to-one. Then there exists a unique* $\vec{v}\in S$ *such that*

$$\begin{aligned} &g_1(\tilde{v}^1)=\alpha_1=\inf\{g_1(v^1):\ \forall\vec{v}\in S\},\\ &g_j^k(\tilde{v}_k^j)=\alpha_j^k=\inf\{g_j^k(v_k^j):\ \forall\vec{v}\in S\}, \qquad j=2,3,k=1,2. \end{aligned} \tag{5.4}$$

Proof. (1) First we note that there exists functions g_1, g_j^k satisfying the hypotheses of the lemma. If S is a finite set, then it is trivial to show the stated results. Let $\{\vec{v}_n\}$ be a minimizing sequence of the optimization problem (5.4) with

$$\alpha_1\le g_1(v_n^1)\le \alpha_1+\frac{1}{n}.$$

Since $\vec{v}_n$ are in $\mathcal{U}$, and since by hypothesis $\mathcal{U}$ is a compact subset of $H^2(I)\times H^2(\Gamma)$, there exists a subsequence such that

$$\vec{v}_n\to\hat{\vec{v}} \text{ in } H^1(I)\times\{H^1\cap(H^2(I))_{weak}\}\times(H^2(I))_{weak}.$$

It is clear that

$$g_1(v_n^1)\to g_1(\hat{v}^1)=\alpha_1.$$

We now show that $\hat{\vec{v}}\in S$, i.e., is such that

$$\Psi(\vec{u};\hat{\vec{v}})\le\Psi(\vec{u};\vec{v}) \qquad \forall\vec{v}\in\mathcal{U}.$$

From the definition of $\vec{v}_n$, we get

$$\begin{aligned}\Psi(\vec{u};\vec{v}_n)&=J_1(\mathbf{y}_n;v_n^1;\mathbf{u})+J_2(\mathbf{x}_n;u^1;\mathbf{v}_n)\\ &\le\Psi(\vec{u};\vec{v}) \qquad \forall\vec{v}\in\mathcal{U},\end{aligned}$$

where $\{\mathbf{y}_n, \mathbf{x}_n\}$ are the solutions of (4.1) with the controls $\{u^1, \mathbf{v}_n\}$, $\{v_n^1, \mathbf{u}\}$, respectively. An argument like that of Lemma 5.1 gives

$$\{\mathbf{y}_n, \mathbf{x}_n\} \to \{\mathbf{y}_*, \mathbf{x}_*\} \text{ in } L^2(0,T;L^2(Q)) \cap (L^2(0,T;H_0^1(Q)))_{weak}$$

and

$$\{\mathbf{y}_n, \mathbf{x}_n\} \mid_{\psi=3/2} \to \{\mathbf{y}_*, \mathbf{x}_*\} \mid_{\psi=3/2} \text{ in } L^2(0,T;L^2(0,2)).$$

Moreover, $\{\mathbf{y}_*, \mathbf{x}_*\}$ are the solutions of (4.1) with the corresponding controls. From the definition of J_1, J_2 and from the above, we deduce that

$$\Psi(\vec{u};\, \hat{\vec{v}}) \le \liminf \Psi(\vec{u};\, \vec{v}_n) \le \Psi(\vec{u};\, \vec{v}) \qquad \forall \vec{v} \in \mathcal{U}.$$

Since g_1 is one-to-one, $\hat{v}^1$ is unique.

(2) Now let

$$\alpha_2^1 = \inf\{g_2^1(v_1^2) :\ \forall \vec{\hat{v}} \in S \text{ with } \vec{\hat{v}} = (\hat{v}^1, .\,, .)\}.$$

Repeating the same argument as above, we have $\vec{v}^* \in S$ such that

$$\alpha_2^1 = g_2^1(v_1^{*,2}) = \inf\{g_2^1(v_1^2) :\ \forall \vec{\hat{v}} \in S,\, \vec{\hat{v}} = (\hat{v}^1, ., .)\}.$$

Then

$$g_2^1(v_1^{*,2}) = \alpha_2^1, \qquad g_1(v_*^1) = \alpha_1.$$

Repeating the process three more times, we get the stated result.

Let A be the nonlinear mapping of U, considered as a compact convex subset of $L^2(I) \times L^2(\Gamma)$ into $\mathcal{U}$, given by

$$A(\vec{u}) = \vec{\hat{v}}, \tag{5.5}$$

where $\vec{\hat{v}}$ is the unique element of $\mathcal{U}$ given by Lemma 5.2.

LEMMA 5.3. *Let A be as in* (5.5)*; then A has a fixed point $\vec{u}^* \in \mathcal{U}$.*

Proof. The nonlinear operator A, given by (5.5), maps a compact convex set U into itself. To show that A has a fixed point, we shall apply Schauder's theorem. Since U is compact, it suffices to show that it is continuous.

Let

$$\vec{v}_n^* = A(\vec{u}_n), \qquad \vec{u}_n \in \mathcal{U}.$$

Since both $\vec{u}_n$, $\vec{v}_n^*$ are in $\mathcal{U}$, we obtain, by taking subsequences,

$$\{\vec{u}_n, \vec{v}_n^*\} \to \{\vec{u}^*, \vec{v}^*\} \text{ in } H^1(I) \times \{H^1(\Gamma) \cap (H^2(I))_{weak}\} \times (H^2(I))_{weak}.$$

By definition, we have

$$\Psi(\vec{u}_n;\, \vec{v}_n^*) \le \Psi(\vec{u}_n;\, \vec{v}) \qquad \forall \vec{v} \in \mathcal{U}.$$

Thus, as in the proof of Lemma 5.1, we get

$$\Psi(\vec{u} :\, \vec{v}^*) \le \liminf \Psi(\vec{u}_n;\, \vec{v}_n^*). \tag{5.6}$$

It follows from the definition of $\vec{v}_n^*$ and from (5.6) that

$$\Psi(\vec{u};\, \vec{v}^*) \le \liminf \Psi(\vec{u}_n;\, \vec{v}) \qquad \forall \vec{v} \in \mathcal{U}. \tag{5.7}$$

On the other hand,

$$\Psi(\vec{u}_n;\, \vec{v}) = J_1(\mathbf{y}_n;\, {}_n; v^1) + J_2(\mathbf{x}_n;\, u_n^1;\, \mathbf{v}),$$

where $\{\mathbf{y}_n,\, \mathbf{x}_n\}$ are the solutions of (4.1) with controls $\{u_n^1,\, \mathbf{v}\}$, $\{v^1,\, \mathbf{u}_n\}$, respectively. A proof like that of Lemma 5.1 yields

$$\Psi(\vec{u};\, \vec{v}) = \lim \Psi(\vec{u}_n;\, \vec{v}) \qquad \forall \vec{v} \in \mathcal{U}. \tag{5.8}$$

It follows from (5.7)–(5.8) that

$$\Psi(\vec{u};\, \vec{v}^*) \le \Psi(\vec{u};\, \vec{v}) \qquad \forall \vec{v} \in \mathcal{U}.$$

The nonlinear mapping A satisfies all the hypotheses of the Schauder fixed point theorem. There exists $\vec{u}_* = \{u_*^1,\, \mathbf{u}_*\} \in \mathcal{U}$ such that $A(\vec{u}_*) = \vec{u}_*$.

Set $\vec{v} = (u^1,\, \mathbf{u}_*)$ in the formula and we obtain

$$\begin{aligned} &\Psi(\vec{u}_*;\, \vec{u}_*) \le \Psi(\vec{u}_*;\, \vec{v}), \\ &J_1(\mathbf{y};\, u_*^1;\, \mathbf{u}_*) \le J_1(\mathbf{x};\, u^1;\, \mathbf{u}_*) \qquad \forall u^1 \in \mathcal{U}_1. \end{aligned}$$

With $\vec{v} = (u_*^1,\, \mathbf{v})$, we have

$$J_2(\mathbf{y};\, u_*^1;\, \mathbf{u}_*) \le J_2(\mathbf{x};\, u_*^1;\, \mathbf{v}) \qquad \forall \mathbf{v} \in \mathcal{U}_2 \times \mathcal{U}_3.$$

The lemma is proved.

Proof of Theorem 2.1. From Theorem 4.1 and from Lemma 5.3, we deduce that there exists $\mathbf{y}$ and $\vec{u}_* = (u_*^1,\, \mathbf{u}_*) \in \mathcal{U}$ with $\mathbf{y}$ being the unique solution of (4.1) with control $\vec{u}_*$. Moreover,

$$\begin{aligned} &J_1(\tilde{\mathbf{y}};\, u_*^1;\, \mathbf{u}_*) \le J_1(\mathbf{x};\, u^1;\, \mathbf{u}_*) \qquad \forall u^1 \in \mathcal{U}_1, \\ &J_2(\tilde{\mathbf{y}};\, u_*^1;\, \mathbf{u}_*) \le J_2(\mathbf{z};\, u_*^1;\, \mathbf{v}) \qquad \forall \mathbf{v} \in \mathcal{U}_2 \times \mathcal{U}_3; \end{aligned}$$

$\mathbf{x}$, $\mathbf{z}$ are the unique solutions of (4.1) with the indicated controls.

Set

$$\begin{aligned} \mathbf{Y}(\psi; \eta; t) &= \tilde{\mathbf{y}}(\psi, \zeta, t) + \mathbf{w} \\ &= \tilde{\mathbf{y}}(\psi;\, 2\eta / u_*^1(\psi); t) + \mathbf{w}, \end{aligned}$$

where $\mathbf{w}$ is the solution of the Stokes equations, given by Lemma 2.1, with boundary controls $\mathbf{v} = \mathbf{u}_*$. Then $\mathbf{Y}$ is the solution of (2.3) and $\{u_*^1,\, \mathbf{u}_*\}$ is the open loop of the problem.

The theorem is proved.

We now turn to the case when the deformation vector function is involved in the cost functional. The open loop problem is open in that case as we have only weak convergence of the first order derivatives of the approximating sequences. Let

$$\tilde{J}(\mathbf{y};\, u^1;\, \mathbf{v}) = \int_0^T \int_Q \{U \nabla \mathbf{y} + (U \nabla \mathbf{y})^t\}^2 2u^1 d\psi d\zeta dt + J_1(\mathbf{y};\, u^1;\, \mathbf{v}) + J_2(\mathbf{y};\, u^1;\, \mathbf{v})$$

and consider the problem

$$\alpha = \inf\{\tilde{J}(\mathbf{y};\, u^1;\, \mathbf{u}) :\ \vec{u} \in \}, \tag{5.9}$$

where $\mathbf{y}$ is the unique solution of (4.1) with controls $\vec{u} = (u^1,\, \mathbf{u})$, given by Theorem 4.1 and by Lemma 4.1.

LEMMA 5.4. *Suppose all the hypotheses of Theorem* 4.1 *are satisfied. Then there exists* $\{\hat{\mathbf{y}},\, \hat{u}^1,\, \hat{\mathbf{u}}\}$ *such that*

$$\tilde{J}(\hat{\mathbf{y}};\, \hat{u}^1;\, \hat{\mathbf{u}}) = \alpha = \inf\{\tilde{J}(\mathbf{y};\, u^1;\, \mathbf{u}) :\ \forall \vec{u} \in \mathcal{U}\}.$$

Proof. It is clear that α is finite. Let $\{\vec{u}_n\}$ be a minimizing sequence of the optimization problem (5.8) with

$$\alpha \le \tilde{J}(\mathbf{y}_n;\, u_n^1;\, \mathbf{u}_n) \le \alpha + \frac{1}{n}.$$

Since $\vec{u}_n \in \mathcal{U}$, with our hypotheses on $\mathcal{U}$, we obtain, by taking subsequences if necessary,

$$\vec{u}_n \to \vec{u}_* \qquad \text{in } H^1(I) \times H^1(\Gamma).$$

From Theorem 4.1, we have the estimate

$$\|\mathbf{y}_n\|_{L^2(0,T;H^1(Q))} + \|\mathbf{y}_n\|_{L^\infty(0,T;L^2(Q))} + \|D_t^\gamma \mathbf{y}_n\|_{L^2(0,T;L^2(Q))} \le C.$$

Again by taking subsequences, we get

$$\mathbf{y}_n \to \hat{\mathbf{y}} \text{ in } L^2(0,T;L^2(Q)) \cap (L^2(0,T;H^1(Q)))_{weak} \cap (L^\infty(0,T;L^2(Q)))_{weak^*}$$

with

$$\mathbf{y}_n \to \hat{\mathbf{y}} \text{ a.e.}, \qquad \mathbf{y}_n\,|_{\psi=1.5} \to \hat{\mathbf{y}}\,|_{\psi=1.5} \qquad \text{in } L^2(0,T;L^2(0,2)).$$

It is not difficult to check that $\hat{\mathbf{y}}$ is the solution of (4.1) with control $\vec{u}_*$. We now have

$$\alpha = \lim \tilde{J}(\mathbf{y}_n;\, u_n^1;\, \mathbf{u}_n) = \tilde{J}(\hat{\mathbf{y}};\, u_*^1;\, \mathbf{u}_*).$$

Thus, by definition

$$\tilde{J}(\hat{\mathbf{y}};\, u_*^1;\, \mathbf{u}_*) \le \tilde{J}(\mathbf{x};\, v^1;\, \mathbf{v}) \qquad \forall \vec{v} \in \mathcal{U},$$

where $\mathbf{x}$ is the solution of (4.1) with control $\vec{v}$.

Proof of Theorem 2.2. Let $\hat{\mathbf{y}}$, $\mathbf{u}_*$ be as in Lemma 5.4 and set

$$\begin{aligned}\hat{\mathbf{Y}}(\psi;\, \eta, t) &= \hat{\mathbf{y}}(\psi; \zeta, t) + \mathbf{w} \\ &= \hat{\mathbf{y}}(\psi; 2\eta/u_*^1(\psi);\, t) + \mathbf{w},\end{aligned}$$

where $\mathbf{w}$ is the solution of the Stokes equations of Lemma 2.1 with $\mathbf{v} = \mathbf{u}_*$. Then $\{\hat{\mathbf{Y}},\, \mathbf{u}_*\}$ is the sought solution of the theorem.

Acknowledgment. The author wishes to thank the referees for their comments.

REFERENCES

[1] G. Armugan and O. Pironneau, *On the problem of riblets as a drag reduction device,* Optim. Control Appl. Methods, 10 (1989), pp. 93–112.
[2] J. Cea, *Problems of shape optimal design,* in Optimization of Distributed Parameters Structures, J. Cea and E.J. Haug, eds., Sijthoff and Noordhoff, Alphen aan den Rijn, the Netherlands, 1981, pp. 1049–1088.
[3] J.Chavent, *On parameter identifiability,* in Proceedings of the Seventh IFAC Symposium on Identification and System Parameter Estimations, Pergamon Press, New York, 1985, pp. 531–536.
[4] M. D. Gunzburger and H. Kim, *Existence of an optimal solution of a shape control problem for the stationary Navier–Stokes equations,* SIAM J. Control Optim., 36 (1998), pp. 895–909.
[5] M. D. Gunzburger and S. L. Hou, *Treating inhomogeneous essential boundary conditions in finite element methods and the calculation of boundary stresses,* SIAM J. Numer. Anal., 29 (1992), pp. 390–424.
[6] F. James and M. Sepulveda, *Parameter identification for a model of chromatographic column,* Inverse Problems, 10 (1994), pp. 367–385.
[7] H. Kim, *Penalized approach and analysis of an optimal shape control problem for the stationary Navier-Stokes equations,* J. Korean Math. Soc., 38 (2001), pp. 1–23.
[8] S. Lenhart, V. Protopopescu, and J. Yong, *Identification of boundary shape and reflectivity in a wave equation by optimal control techniques,* Differential Integral Equations, 13 (2000), pp. 941–972.
[9] S. Lenhart, V. Protopopescu, and J. Yong, *Solving inverse problems of identification type by optimal control techniques,* in Proceedings of the International Conference on Advances of Nonlinear Dynamics near the Millenium, AIP Press, Melville, NY, 1997, pp. 87–94.
[10] S. Lenhart, V. Protopopescu, and J. Yong, *Optimal control of a reflection boundary coefficient in an acoustic wave equation,* Appl. Anal., 68 (1998), pp. 179–194.
[11] J. L. Lions, *Quelques methodes de resolution des problemes aux limites non lineaire,* Dunod, Paris, 1969.
[12] W. Liu and J. E. Rubio, *Local convergences and optimal shape design,* SIAM J. Control Optim., 30 (1992), pp. 49–62.
[13] O. Pironneau, *On optimal design in fluid mechanics,* J. Fluid Mech., 64 (1974), pp. 97–110.
[14] R. Temam, *Navier-Stokes Equations,* North–Holland, Amsterdam, 1979.
[15] B. A. Ton, *Open loop equilibrium strategy for quasi-variational inequalities and for constrained non-cooperatives games,* Numer. Funct. Anal. Optim., 7 (1996), pp. 1053–1091.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 6, pp. 1748–1755

EXACT BOUNDARY CONTROLLABILITY FOR QUASI-LINEAR HYPERBOLIC SYSTEMS*

TA-TSIEN LI† AND BO-PENG RAO‡

Abstract. Using a result on the existence and uniqueness of the semiglobal C^1 solution to the mixed initial-boundary value problem for first order quasi-linear hyperbolic systems with general nonlinear boundary conditions, we establish the exact boundary controllability for quasi-linear hyperbolic systems if the C^1 norm of initial and final states is small enough.

Key words. semiglobal C^1 solution, exact boundary controllability, quasi-linear hyperbolic system

AMS subject classifications. 35L50, 49J20, 49N50

PII. S0363012901390099

1. Introduction and main result.

Consider the first order quasi-linear hyperbolic system

$$\frac{\partial u}{\partial t} + A(u)\frac{\partial u}{\partial x} = F(u), \tag{1.1}$$

where $u = (u_1, \dots, u_n)^T$ is a vector valued function of (t, x), $A(u) = (a_{ij}(u))$ is an $n \times n$ matrix with suitably smooth elements $a_{ij}(u)(i, j = 1, \dots, n)$, $F : \mathbb{R}^n \to \mathbb{R}^n$ is a vector valued function with suitably smooth components $f_i(u)(i = 1, \dots, n)$, and

$$F(0) = 0. \tag{1.2}$$

By the definition of hyperbolicity, on the domain under consideration, the matrix $A(u)$ has n real eigenvalues $\lambda_i(u)$ $(i = 1, \dots, n)$ and a complete set of left eigenvectors $l_i(u) = (l_{i1}(u), \dots, l_{in}(u))$ $(i = 1, \dots, n)$:

$$l_i(u)A(u) = \lambda_i(u)l_i(u), \tag{1.3}$$

and a complete set of right eigenvectors $r_i(u) = (r_{i1}(u), \dots r_{in}(u))^T$ $(i = 1, \dots, n)$:

$$A(u)r_i(u) = \lambda_i(u)r_i(u). \tag{1.4}$$

We have

$$\det |l_{ij}(u)| \neq 0 \qquad (\text{resp., } \det |r_{ij}(u)| \neq 0). \tag{1.5}$$

Without loss of generality, we may assume that

$$l_i(u)r_j(u) \equiv \delta_{ij} \qquad (i, j = 1, \dots n), \tag{1.6}$$

*Received by the editors May 30, 2001; accepted for publication (in revised form) April 25, 2002; published electronically February 4, 2003. This work was supported by the Special Funds for Major State Basic Research Projects of China.

http://www.siam.org/journals/sicon/41-6/39009.html

†Department of Mathematics, Fudan University, Shanghai 200433, People's Republic of China (dqli@fudan.edu.cn).

‡Institut de Recherche Mathématique Avancée, Université Louis Pasteur de Strasbourg, 7 Rue René-Descartes, 67084 Strasbourg, France (rao@math.u-strasbg.fr).

$$r_i^T(u)r_i(u) \equiv 1 \qquad (i = 1, \dots, n), \tag{1.7}$$

where δ_{ij} stands for the Kronecker symbol. Moreover, we assume that, on the domain under consideration, the eigenvalues satisfy the following conditions:

$$\lambda_r(u) < 0 < \lambda_s(u) \qquad (r = 1, \dots, m; \quad s = m+1, \dots, n). \tag{1.8}$$

Let

$$v_i = l_i(u)u \qquad (i = 1, \dots, n). \tag{1.9}$$

We consider the mixed initial-boundary value problem for the quasi-linear hyperbolic system (1.1) with the initial condition

$$t = 0: \quad u = \phi(x), \qquad 0 \le x \le 1, \tag{1.10}$$

and the boundary conditions

$$x = 0: \quad v_s = G_s(t, v_1, \dots, v_m) + H_s(t) \quad (s = m+1, \dots, n), \tag{1.11}$$

$$x = 1: \quad v_r = G_r(t, v_{m+1}, \dots, v_n) + H_r(t) \quad (r = 1, \dots, m). \tag{1.12}$$

Without loss of generality, we assume that

$$G_i(t, 0, \dots, 0) \equiv 0 \qquad (i = 1, \dots, n). \tag{1.13}$$

Let us recall the following result of Li–Jin [9] on the semiglobal C^1 solution, which will be used as the main tool in what follows.

LEMMA 1.1. *Assume that $l_{ij}(u), \lambda_i(u), f_i(u), G_i(t, \cdot), H_i(t)$ $(i, j = 1, \dots, n)$, and $\phi(x)$ are all C^1 functions with respect to their arguments. Assume, furthermore, that* (1.2), (1.5), (1.8), *and* (1.13) *hold. Assume finally that the conditions of C^1 compatibility are satisfied at the points $(0, 0)$ and $(0, 1)$, respectively. Then, for a given $T_0 > 0$, the mixed initial-boundary value problem* (1.1) *and* (1.10)–(1.12) *admits a unique C^1 solution $u = u(t, x)$ (called the semiglobal C^1 solution) with sufficiently small C^1 norm on the domain*

$$R(T_0) = \{(t, x) | \quad 0 \le t \le T_0, \quad 0 \le x \le 1\}, \tag{1.14}$$

provided that the C^1 norms $\|\phi\|_{C^1[0,1]}$ and $\|H\|_{C^1[0,T_0]}$ are small enough (depending on T_0). In particular, for any given $\epsilon_0 > 0$, we have

$$|u(t, x)| \le \epsilon_0 \tag{1.15}$$

on the domain $R(T_0)$ if $\|\phi\|_{C^1[0,1]}$ and $\|H\|_{C^1[0,T_0]}$ are small enough (depending on T_0 and ϵ_0).

Based on this result, we can consider the following problem of local exact boundary controllability.

For any given initial data $\phi \in C^1[0, 1]$ and final data $\psi \in C^1[0, 1]$ with small C^1 norm, can we find a time $T_0 > 0$ and boundary input controls $H_i \in C^1[0, T_0]$ $(i = 1, \dots, n)$ with small C^1 norm, such that the mixed initial-boundary value problem (1.1) *and* (1.10)–(1.12) *admits a unique C^1 solution $u = u(t, x)$ on the domain $R(T_0)$, which verifies the final condition*

$$u(T_0, x) = \psi(x), \quad 0 \le x \le 1? \tag{1.16}$$

Notice that, because of the finiteness of the speed of wave propagation, the exact boundary controllability of a hyperbolic system requires that the controllability time T_0 must be greater than a given constant. Let

$$T_0 > \max_{i=1,\dots,n} \frac{1}{|\lambda_i(0)|}. \tag{1.17}$$

In this paper, we will give an affirmative answer to the above problem of exact boundary controllability. The main result is the following theorem.

THEOREM 1.2. *Assume that $l_{ij}(u), \lambda_i(u), f_i(u), G_i(t,\cdot)$, and $H_i(t)$ $(i,j=1,\dots,n)$ are all C^1 functions with respect to their arguments. Assume, furthermore, that* (1.2), (1.5), (1.8), *and* (1.13) *hold. Let T_0 be defined by* (1.17). *Then, for any given initial data $\phi \in C^1[0,1]$ and final data $\psi \in C^1[0,1]$ with small C^1 norm, there exist boundary controls $H_i(t) \in C^1[0,T_0]$ $(i=1,\dots,n)$ with small C^1 norm, such that the mixed initial-boundary value problem* (1.1) *and* (1.10)–(1.12) *admits a unique C^1 solution $u = u(t,x)$ on the domain $R(T_0)$, which verifies the final condition* (1.16).

Remark 1.1. The exact controllability time T_0 given in this theorem is optimal.

Remark 1.2. The results given in [7], [8] can be regarded as a direct consequence of this theorem.

Remark 1.3. In some special cases, it is possible to use only boundary controls $H_s(t)$ $(s=m+1,\dots,n)$ on $x=0$ or $H_r(t)$ $(r=1,\dots,m)$ on $x=1$ to realize the exact boundary controllability, but the controllability time T_0 must be doubled (see [10]).

There are a number of publications concerning the exact controllability and the uniform stabilization for linear hyperbolic systems (see [11], [12], and the references therein). Furthermore, using the Hilbert uniqueness method (HUM) suggested by J.-L. Lions [11] and Schauder's fixed point theorem, Zuazua [13] proved the global (resp., local) exact boundary controllability for semilinear wave equations in the asymptotically linear case (resp., the superlinear case with suitable growth conditions). Later, using a global inversion theorem, Lasiecka–Triggiani [4] established an abstract result on the exact controllability for semilinear equations. As an application, they gave the global exact boundary controllability for the wave and plate equations in the asymptotically linear case. However, only a few results are known for quasi-linear hyperbolic systems. In the case in which $n=2$, the exact boundary controllability for *reducible* quasi-linear hyperbolic systems was proved in Li–Zhang [6] and Li–Rao–Jin [7], [8] by a constructive method which does not work in the general case of quasi-linear hyperbolic systems that we consider in this paper. A similar consideration can be found in Fursikov–Imanuvilov [3] for a class of one-dimensional semilinear wave equations. In earlier work, M-Cirinà [1], [2] considered the zero exact controllability for quasi-linear hyperbolic systems with linear boundary controls; however, his results are essentially valid only for the system of diagonal form. Moreover, if one applies the result of [1] twice for getting the general controllability, then the corresponding controllability time must be doubled. In this work, based on the existence result [9] for the semiglobal C^1 solution to the mixed initial-boundary value problem of quasi-linear hyperbolic systems, we establish the local exact boundary controllability for general quasi-linear hyperbolic systems with general nonlinear boundary controls.

2. Reduction of the problem. In order to prove the main theorem, it suffices to establish the following proposition.

PROPOSITION 2.1. *Let T_0 be defined by* (1.17), *and let $\epsilon_0 > 0$ be a given small number. For any given initial data $\phi \in C^1[0,1]$ and final data $\psi \in C^1[0,1]$ with small*

C^1 norm, the quasi-linear hyperbolic system (1.1) *admits a C^1 solution $u = u(t,x)$ on the domain $R(T_0)$ such that*

$$u(0,x) = \phi(x), \qquad 0 \le x \le 1, \tag{2.1}$$

$$u(T_0,x) = \psi(x), \qquad 0 \le x \le 1, \tag{2.2}$$

the C^1 norm of $u = u(t,x)$ is suitably small, and $u = u(t,x)$ satisfies (1.15) *on the domain $R(T_0)$.*

In fact, let $u = u(t,x)$ be a C^1 solution of (1.1) on the domain $R(T_0)$, given by Proposition 2.1. Set

$$H_r(t) = \big(v_r - G_r(t, v_{m+1}, \dots, v_n)\big)|_{x=1} \quad (r = 1, \dots, m), \tag{2.3}$$

$$H_s(t) = \big(v_s - G_s(t, v_1, \dots, v_m)\big)|_{x=0} \quad (s = m+1, \dots, n), \tag{2.4}$$

where v_i $(i = 1, \dots, n)$ are defined by (1.9). Noting (1.13), the C^1 norm of H_i $(i = 1, \dots, n)$ is small. Then, by Lemma 1.1, $u = u(t,x)$ is the semiglobal C^1 solution to the corresponding mixed initial-boundary value problem (1.1) and (1.10)–(1.12) on the domain $R(T_0)$, which also satisfies the final condition (2.2). Therefore, we obtain the desired exact boundary controllability, and the boundary controls H_i $(i = 1, \dots, n)$ are given by (2.3)–(2.4).

3. Exact boundary controllability. In this section, we will prove Proposition 2.1.

First, noting (1.17), there exists an $\epsilon_0 > 0$ so small that

$$T_0 > \max_{|u| \le \epsilon_0,\ i=1,\dots,n} \frac{1}{|\lambda_i(u)|}. \tag{3.1}$$

Let

$$T_1 = \max_{|u| \le \epsilon_0,\ i=1,\dots,n} \frac{1}{2|\lambda_i(u)|}. \tag{3.2}$$

We divide the proof into several steps.

(i) We first consider the forward auxiliary mixed initial-boundary value problem of (1.1) with the initial condition

$$t = 0: \quad u = \phi(x), \qquad 0 \le x \le 1, \tag{3.3}$$

and the boundary conditions

$$x = 0: \quad v_s = f_s(t) \quad (s = m+1, \dots, n), \tag{3.4}$$

$$x = 1: \quad v_r = \overline{f}_r(t) \quad (r = 1, \dots, m), \tag{3.5}$$

where v_i $(i = 1, \dots, n)$ are defined by (1.9) and $f_s, \overline{f}_r$ $(r = 1, \dots, m; s = m+1, \dots, n)$ are any given functions of t with small $C^1[0, T_1]$ norm. We assume that the conditions of C^1 compatibility are satisfied at the points $(0,0)$ and $(0,1)$, respectively. By Lemma 1.1, there exists a unique semiglobal C^1 solution $u = u^{(1)}(t,x)$ with small C^1 norm on the domain

$$\{(t,x)| \quad 0 \le t \le T_1, \quad 0 \le x \le 1\}. \tag{3.6}$$

In particular, the solution $u = u^{(1)}(t,x)$ satisfies (1.15) on the domain (3.6). Thus we can uniquely determine the corresponding value of u on $x = \frac{1}{2}$ as

$$x = \frac{1}{2}: \quad u = a(t), \quad 0 \le t \le T_1, \tag{3.7}$$

and the $C^1[0,T_1]$ norm of $a(t)$ is suitably small.

(ii) Similarly, we consider the backward auxiliary initial-boundary value problem of (1.1) with the initial condition

$$t = T_0: \quad u = \psi(x), \qquad 0 \le x \le 1, \tag{3.8}$$

and the boundary conditions

$$x = 0: \quad v_r = g_r(t) \quad (r = 1, \dots, m), \tag{3.9}$$

$$x = 1: \quad v_s = \overline{g}_s(t) \quad (s = m+1, \dots, n), \tag{3.10}$$

where v_i $(i = 1, \dots, n)$ are defined by (1.9) and $g_r, \overline{g}_s$ $(r = 1, \dots, m; s = m+1, \dots, n)$ are any given functions of t with small $C^1[T_0 - T_1, T_0]$ norm. We assume that the conditions of C^1 compatibility are satisfied at the points $(T_0, 0)$ and $(T_0, 1)$, respectively. Once again, by Lemma 1.1, there exists a unique semiglobal C^1 solution $u = u^{(2)}(t,x)$ with small C^1 norm on the domain

$$\{(t,x) | \quad T_0 - T_1 \le t \le T_0, \quad 0 \le x \le 1\}. \tag{3.11}$$

In particular, the solution $u = u^{(2)}(t,x)$ satisfies (1.15) on the domain (3.11). Thus we can uniquely determine the corresponding value of u on $x = \frac{1}{2}$ as

$$x = \frac{1}{2}: \quad u = b(t), \quad T_0 - T_1 \le t \le T_0, \tag{3.12}$$

and the $C^1[T_0 - T_1, T_0]$ norm of $b(t)$ is suitably small.

(iii) Now we change the order of variables t and x, and then the system (1.1) is rewritten in the following form:

$$\frac{\partial u}{\partial x} + A^{-1}(u)\frac{\partial u}{\partial t} = \widetilde{F}(u) := A^{-1}(u)F(u). \tag{3.13}$$

We notice that

$$\widetilde{F}(0) = 0. \tag{3.14}$$

Noting (1.8), the eigenvalues of the inverse matrix $A^{-1}(u)$ satisfy

$$\frac{1}{\lambda_r(u)} < 0 < \frac{1}{\lambda_s(u)} \qquad (r = 1, \dots, m; \quad s = m+1, \dots, n). \tag{3.15}$$

Moreover, since the matrices $A(u)$ and $A^{-1}(u)$ have the same left eigenvectors, we can still define the variables v_i $(i = 1, \dots, n)$ by the same formula (1.9).

Now we consider the mixed initial-boundary value problem for system (3.13) with the initial condition

$$x = \frac{1}{2}: \quad u = c(t), \qquad 0 \le t \le T_0, \tag{3.16}$$

and the boundary conditions

$$t = 0: \quad v_r = \Phi_r(x) \quad (r = 1, \dots, m), \quad 0 \le x \le \frac{1}{2}, \tag{3.17}$$

$$t = T_0: \quad v_s = \Psi_s(x) \quad (s = m+1, \dots, n), \quad 0 \le x \le \frac{1}{2}, \tag{3.18}$$

where $c(t)$ is a $C^1[0, T_0]$ function with small C^1 norm such that (noting (1.17))

$$c(t) = \begin{cases} a(t), & 0 \le t \le T_1, \\ b(t), & T_0 - T_1 \le t \le T_0; \end{cases} \tag{3.19}$$

moreover,

$$\Phi_i(x) = l_i(\phi(x))\phi(x) \qquad (i = 1, \dots, n), \tag{3.20}$$

$$\Psi_i(x) = l_i(\psi(x))\psi(x) \qquad (i = 1, \dots, n), \tag{3.21}$$

the $C^1[0, 1]$ norm of which is also small. Noting (3.19)–(3.21), we easily check that the mixed initial-boundary value problem (3.13) and (3.16)–(3.18) satisfies the conditions of C^1 compatibility at the points $(0, \frac{1}{2})$ and $(T_0, \frac{1}{2})$, respectively. Therefore, by Lemma 1.1, there exists a unique semiglobal C^1 solution $u = u_l(t, x)$ with small C^1 norm on the domain

$$R_l(T_0) = \left\{ (t, x) | \quad 0 \le t \le T_0, \quad 0 \le x \le \frac{1}{2} \right\}. \tag{3.22}$$

Moreover, $u = u_l(t, x)$ can be asked to satisfy (1.15) on the domain (3.22) if the C^1 norm of $\phi, \psi, \bar{f}_r, f_s, g_r$, and $\bar{g}_s (r = 1, \dots, m; s = m+1, \dots, n)$ is small enough.

(iv) Similarly, the mixed initial-boundary value problem (3.13) with the initial condition (3.16) and the boundary conditions

$$t = 0: \quad v_s = \Phi_s(x) \quad (s = m+1, \dots, n), \quad \frac{1}{2} \le x \le 1, \tag{3.23}$$

$$t = T_0: \quad v_r = \Psi_r(x) \quad (r = 1, \dots, m), \quad \frac{1}{2} \le x \le 1, \tag{3.24}$$

admits a unique semiglobal C^1 solution $u = u_r(t, x)$ with small C^1 norm on the domain

$$R_r(T_0) = \left\{ (t, x) | \quad 0 \le t \le T_0, \quad \frac{1}{2} \le x \le 1 \right\}, \tag{3.25}$$

and $u = u_r(t, x)$ can also be asked to satisfy (1.15) on the domain (3.25).

(v) Let

$$u(t, x) = \begin{cases} u_l(t, x), & (t, x) \in R_l(T_0), \\ u_r(t, x), & (t, x) \in R_r(T_0). \end{cases} \tag{3.26}$$

To complete the proof of Proposition 2.1, it is only necessary to check that

$$t = 0: \quad u = \phi(x), \qquad 0 \le x \le 1, \tag{3.27}$$

$$t = T_0: \quad u = \psi(x), \qquad 0 \le x \le 1. \tag{3.28}$$

In fact, the C^1 solutions $u = u_l(t,x)$ (resp., $u = u_r(t,x)$) and $u = u^{(1)}(t,x)$ satisfy the system (3.13) (and thus (1.1)), the initial condition

$$x = \frac{1}{2}: \quad u = a(t), \quad 0 \le t \le T_1, \tag{3.29}$$

and the boundary conditions

$$\begin{aligned} &t = 0: \quad v_r = \Phi_r(x) \quad (r = 1, \dots m), \quad \frac{1}{2} \le x \le 1 \\ &\left(\text{resp.,} \quad t = 0: \quad v_s = \Phi_s(x) \quad (s = m+1, \dots n), \quad \frac{1}{2} \le x \le 1\right). \end{aligned} \tag{3.30}$$

Because of the finiteness of the speed of wave propagation and the choice of T_1 given by (3.1), the mixed initial-boundary value problem (3.13) and (3.29)–(3.30) has a unique C^1 solution on the domain

$$\begin{aligned} &\left\{(t,x)| \quad 0 \le t \le 2T_1 x, \quad 0 \le x \le \frac{1}{2}\right\} \\ &\left(\text{resp.,} \quad \left\{(t,x)| \quad 0 \le t \le 2T_1(1-x), \quad \frac{1}{2} \le x \le 1\right\}\right) \end{aligned} \tag{3.31}$$

(see Li–Yu [5]). Then

$$u(t,x) \equiv u^{(1)}(t,x) \tag{3.32}$$

on these domains, and, in particular, we obtain (3.27).

On the other hand, the C^1 solutions $u = u_l(t,x)$ (resp., $u = u_r(t,x)$) and $u = u^{(2)}(t,x)$ satisfy the system (3.13) (and thus (1.1)), the initial condition

$$x = \frac{1}{2}: \quad u = b(t), \quad T_0 - T_1 \le t \le T_0, \tag{3.33}$$

and the boundary conditions

$$\begin{aligned} &t = T_0: \quad v_s = \Psi_s(x) \quad (s = m+1, \dots, n), \quad 0 \le x \le \frac{1}{2} \\ &\left(\text{resp.,} \quad t = T_0: \quad v_r = \Psi_r(x) \quad (r = 1, \dots, m), \quad \frac{1}{2} \le x \le 1\right). \end{aligned} \tag{3.34}$$

Similarly, the mixed initial-boundary value problem (3.13) and (3.33)–(3.34) has a unique C^1 solution on the domain

$$\begin{aligned} &\left\{(t,x)| \quad T_0 - 2T_1 x \le t \le T_0, \quad 0 \le x \le \frac{1}{2}\right\} \\ &\left(\text{resp.,} \quad T_0 + 2T_1(x-1) \le t \le T_0, \quad \frac{1}{2} \le x \le 1\right). \end{aligned} \tag{3.35}$$

Then it follows that

$$u(t,x) \equiv u^{(2)}(t,x) \tag{3.36}$$

on these domains, and, in particular, we obtain (3.28).

Thus we have finished the proof of Proposition 2.1, and then we get the local exact boundary controllability. In view of the proof, we see that the boundary controls are certainly not unique.

4. Remarks. We can consider the mixed initial-boundary value problem for the quasi-linear hyperbolic system (1.1) with the initial condition (1.10) and the following boundary conditions:

$$x = 0: \quad \widetilde{v}_s = \widetilde{g}_s(t, \widetilde{v}_1, \ldots, \widetilde{v}_m) + \widetilde{h}_s(t) \quad (s = m+1, \ldots, n), \tag{4.1}$$

$$x = 1: \quad \widetilde{v}_r = \widetilde{g}_r(t, \widetilde{v}_{m+1}, \ldots, \widetilde{v}_n) + \widetilde{h}_r(t) \quad (s = 1, \ldots, m), \tag{4.2}$$

where

$$\widetilde{v}_i = l_i(\phi(x))u \qquad (i = 1, \ldots, n), \tag{4.3}$$

and, without loss of generality, we assume that

$$\widetilde{g}_i(t, 0, \ldots, 0) \equiv 0 \qquad (i = 1, \ldots, n). \tag{4.4}$$

Following [9], for the C^1 solution $u = u(t, x)$ satisfying (1.15) with suitably small $\epsilon_0 > 0$, the mixed initial-boundary value problem (1.1), (1.10), and (4.1)–(4.2) is equivalent to the mixed initial-boundary value problem (1.1) and (1.10)–(1.12). Then we can also establish the exact boundary controllability for this problem. More precisely, T_0 being given by (1.17), for any given initial data ϕ and final data ψ with small $C^1[0, 1]$ norm, there exist boundary controls $\widetilde{h}_i$ $(i = 1, \ldots, n)$ with suitably small $C^1[0, T_0]$ norm such that the mixed initial-boundary value problem (1.1), (1.10), and (4.1)–(4.2) admits a unique C^1 solution $u = u(t, x)$ on the domain $R(T_0)$, which verifies the final condition (1.16).

Acknowledgments. The authors would like to thank the referees for their kind and valuable suggestions.

REFERENCES

[1] M. Cirinà, *Boundary controllability of nonlinear hyperbolic systems*, SIAM J. Control, 7 (1969), pp. 198–212.

[2] M. Cirinà, *Nonlinear hyperbolic problems with solutions on preassigned sets*, Michigan Math. J., 17 (1970), pp. 193–209.

[3] A. V. Fursikov and O. Yu. Imanuvilov, *Controllability of Evolution Equations*, Lecture Notes Ser. 34, Seoul National University, Seoul, 1996.

[4] I. Lasiecka and R. Triggiani, *Exact controllability of semilinear abstract systems with applications to waves and plates boundary control problems*, Appl. Math. Optim., 23 (1991), pp. 109–154.

[5] T.-T. Li and W.-C. Yu, *Boundary Value Problems for Quasilinear Hyperbolic Systems*, Duke University Math. Ser. V, Durham, NC, 1985.

[6] T.-T. Li and B.-Y. Zhang, *Global exact boundary controllability of a class of quasilinear hyperbolic systems*, J. Math. Anal. Appl., 225 (1998), pp. 289–311.

[7] T.-T. Li, B.-P. Rao, and Y. Jin, *Solution C^1 semi-global et contrôlabilité exacte frontière de systèmes hyperboliques quasi linéaires réductibles*, C. R. Acad. Sci. Paris Sér. I Math., 330 (2000), pp. 205–210.

[8] T.-T. Li, B.-P. Rao, and Y. Jin, *Semi-global C^1 solution and exact boundary controllability for reducible quasilinear hyperbolic systems*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 399–408.

[9] T.-T. Li and Y. Jin, *Semi-global C^1 solution to the mixed initial-boundary value problem for quasilinear hyperbolic systems*, Chinese Ann. Math. Ser. B, 22 (2001), pp. 325–336.

[10] T.-T. Li and B.-P. Rao, *Local exact boundary controllability for a class of quasilinear hyperbolic systems*, Chinese Ann. Math. Ser. B, 23 (2002), pp. 209–218.

[11] J.-L. Lions, *Contrôlabilité Exacte, Perturbations et Stabilisation de Systèmes Distribués,* Vol. I, Masson, Paris, 1988.

[12] D. L. Russell, *Controllability and stabilizability theory for linear partial differential equations: Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.

[13] E. Zuazua, *Exact controllability for the semilinear wave equation*, J. Math. Pures Appl. (9), 69 (1990), pp. 1–31.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 6, pp. 1756–1778

IMMERSION OF NONLINEAR SYSTEMS INTO LINEAR SYSTEMS MODULO OUTPUT INJECTION*

PHILIPPE JOUAN†

Abstract. The problem of the immersion of a SISO system into a linear up to an output injection one is studied in order to design Luenberger-like observers. Necessary and sufficient conditions are stated within a very general framework. Effective computations and examples are then provided.

Key words. immersion of nonlinear systems, linearization, output injection, observability

AMS subject classifications. 93C10, 93C18, 93B07

PII. S0363012901391706

1. Introduction.

It has been well known for a long time that a controlled and observed system in $\mathbb{R}^n$ whose nonlinear part depends only upon the output admits Luenberger-like observers. More specifically, let us consider

$$\begin{cases} \dot{z} = Az + \varphi(u, Cz), \\ \xi = Cz, \end{cases} \tag{1.1}$$

where $z \in \mathbb{R}^n$, $\xi \in \mathbb{R}$, $u \in \mathbb{R}^p$, $A \in \mathcal{M}(n, \mathbb{R})$, and $C \in \mathcal{M}(1 \times n, \mathbb{R})$. Such a system will be referred to as a linear up to an output injection system. If the pair (C, A) is observable, we can choose a vector-column K such that the eigenvalues of $A - KC$ have arbitrary prescribed values. If their real parts are negative, then the system

$$\dot{\widehat{z}} = A\widehat{z} + \varphi(u, \xi) - K(C\widehat{z} - \xi) \tag{1.2}$$

is an observer for (1.1): denoting the error by $e = \widehat{z} - z$, it is straightforward that

$$\dot{e} = (A - KC)e.$$

Thus the error dynamics is linear and the error converges exponentially to zero.

Up to now, mainly the possibility to transform a given system into the form (1.1) by diffeomorphism was studied. In the paper [KI83], Krener and Isidori gave necessary and sufficient conditions to transform an uncontrolled, single-output system into (1.1). Then Krener and Respondek extended the results to MIMO systems (see [KR85]). In these two papers the transformations under consideration are (local) diffeomorphisms in the state space, together with a diffeomorphism in the output space in the second one (see also [BZ83], [XG89], [Phe91], [GMP96], [HG88], [HP99]).

Some generalizations of the output injection method have been recently proposed in order to enlarge the class of systems for which one can design an observer with linear error dynamics. Among them let us quote the output-dependent time-scale transformation [Gua01], the generalized output injection (after transformation the nonlinear term depends upon the output, the input, and some of its derivatives) [Kel87], and the completely generalized output injection [LPG99] (the nonlinear term

*Received by the editors July 2, 2001; accepted for publication (in revised form) August 5, 2002; published electronically February 4, 2003.
http://www.siam.org/journals/sicon/41-6/39170.html
†Lab. R. Salem, CNRS UMR 6085, Université de Rouen, Mathématiques, site Colbert, 76821 Mont-Saint-Aignan Cedex, France (Philippe.Jouan@univ-rouen.fr).

depends, moreover, upon the derivatives of the output). In these papers, as well as in the previous ones, the state space is transformed by (local) diffeomorphism.

In the present paper we deal with the problem of the immersion of a given system into a linear up to an output injection one whose state space dimension may be greater, or lower, than the dimension of the initial state space. The precise definition of an immersion will be stated later (see Definition 2.2), but roughly speaking, a system Σ can be immersed into a system S if the input-output mapping of Σ, possibly followed by a diffeomorphism ψ of $\mathbb{R}$, is a restriction of the input-output mapping of S. The interest is twofold: on the one hand the class of systems into consideration is much wider; on the other hand it is well known that a generic uncontrolled system can be globally put in observable form by embedding but generically not by diffeomorphism. Therefore the transformation of systems by diffeomorphism will always keep a local character.

Among the literature let us mention [FK83], where the problem of the immersion into a bilinear system is studied, and [BRG89], where the immersion of a control-affine system into a linear up to an output injection one is studied under the very strong hypothesis that the drift system is linearizable.

We consider herein a smooth system defined on a manifold X,

$$\Sigma = \begin{cases} \dot{x} = f(x,u), \\ y = h(x), \end{cases}$$

where $u \in \mathbb{R}$ and $y \in \mathbb{R}$. (The paper mainly deals with control-affine systems; however, we start with general ones because the first reduction, stated in Theorem 2.3, is true in the general case.) This system will be said to be LIS (for the French *linéarisable par injection de sortie*) if it can be immersed into a linear up to an output injection one.

For instance, the reader can check that the uncontrolled system defined in $\mathbb{R}^2$ by

$$\begin{cases} \dot{x}_1 = x_2 \exp(-x_1), \\ \dot{x}_2 = \frac{1}{2}x_2^2 \exp(-x_1) + x_2 \exp(x_1) + 1, \\ y = h(x_1, x_2) = \exp(x_1) \end{cases}$$

is LIS because it can be immersed into the system defined in $\mathbb{R}^3$ by

$$\begin{cases} \dot{z_1} = z_2 + \frac{5}{6}y^2, \\ \dot{z_2} = z_3 + \ln y - \frac{2}{9}y^3, \\ \dot{z_3} = \quad -\frac{2}{3}y, \\ y = z_1, \end{cases}$$

the immersion being

$$\begin{cases} z_1 = \exp(x_1), \\ z_2 = x_2 - \frac{5}{6}\exp(2x_1), \\ z_3 = \frac{1}{2}x_2^2 \exp(-x_1) - \frac{2}{3}x_2 \exp(x_1) + \frac{2}{9}\exp(3x_1) - x_1 + 1. \end{cases}$$

We want to thank the anonymous reviewer who gave us this example, in a somewhat different form.

First at all we state in Theorem 2.3 that whenever this system is LIS, it can be

immersed into a system "in canonical form," that is, in the form

$$\begin{cases} \dot{z}_1 & = z_2 + \varphi_1(u, z_1), \\ \dot{z}_2 & = z_3 + \varphi_2(u, z_1), \\ \vdots & \vdots \\ \dot{z}_{n-1} & = z_n + \varphi_{n-1}(u, z_1), \\ \dot{z}_n & = \quad +\varphi_n(u, z_1), \\ \xi & = z_1. \end{cases}$$

This result leads in the uncontrolled case to Theorem 2.6: *An uncontrolled system is LIS if and only if there exist a smooth function ψ and n smooth functions φ_1, $\varphi_2, \ldots, \varphi_n$ such that*

$$(F) \qquad L_f^n \tilde{h} = L_f^{n-1}(\varphi_1 \circ \tilde{h}) + L_f^{n-2}(\varphi_2 \circ \tilde{h}) + \cdots + L_f(\varphi_{n-1} \circ \tilde{h}) + \varphi_n \circ \tilde{h},$$

where $\tilde{h} = \psi \circ h$.

Although this characterization involves unknown functions φ_i and ψ, it is fundamental for several reasons. In the first place it is very general because it is just as well global than local and because no assumption of observability is needed; then formula (F) will be systematically used to implement the computations; finally it is the starting point of the control-affine case. This last is studied in section 2.4, where Theorem 2.7 is stated: *A control-affine system whose dynamics is $\dot{x} = f(x) + ug(x)$ is LIS if and only if the following hold:*

1. *The uncontrolled part of the system is LIS.*
2. *If we denote by $\tau = (\tau_1, \ldots, \tau_n)$ the immersion of the uncontrolled part into a canonical system in $\mathbb{R}^n$, then for $i = 1, \ldots, n$ $L_g\tau_i$ is a smooth function of h.*

Moreover, at points x such that $dh(x) \neq 0$, condition 2 is equivalent to

$$dL_g\tau_i \wedge dh = 0.$$

After these theoretical results the second part of the paper is devoted to computational ones. The actual computation of the functions φ_i involved in formula (F), necessary to decide whether a given system can be immersed into a linear up to an output injection one, is not easy and we start with systems whose drift part is in observable form:

$$\Sigma = \begin{cases} \dot{x}_1 = x_2 & + ug_1(x), \\ \dot{x}_2 = x_3 & + ug_2(x), \\ \cdots & \cdots \\ \dot{x}_d = \Phi(x_1, x_2, \ldots, x_d) & + ug_d(x), \\ y = h(x) = x_1, \end{cases}$$

where $d = \dim(X)$, this last condition ensuring the uniqueness of $\Phi(x_1, x_2, \ldots, x_d)$. We also assume that no diffeomorphism in the output space is necessary ($\psi = Id_{\mathbb{R}}$). It turns out that the control-affine case is easier to deal with than the uncontrolled one because there are additional conditions that avoid to solve (untractable) differential equations. Thus the computation process is complete in the SISO control-affine case (with the slight restriction $g_1 \neq 0$).

2. Theoretical results.

2.1. Definitions and notations. Let X be a $\mathcal{C}^\infty$, connected manifold. We consider on X the system

$$\Sigma = \begin{cases} \dot{x} = f(x,u), \\ y = h(x), \end{cases} \tag{2.1}$$

where $x \in X$, $u \in \mathbb{R}$, and $y \in \mathbb{R}$. The parametrized vector field f and the output function h are assumed to be $\mathcal{C}^\infty$. We define the open interval $I_o =]a,b[$, where $a, b \in \overline{\mathbb{R}}$, w.r.t. the range $h(X)$ of h in the following way: if the lower bound α of $h(X)$ is finite and $h(X)$ is closed at α, then $a = -\infty$, otherwise $a = \alpha$; if the upper bound β of $h(X)$ is finite and $h(X)$ is closed at β, then $b = +\infty$, otherwise $b = \beta$.

In view of Definition 2.1 we denote by A_c the (n-dimensional) antishift matrix, that is, the $n \times n$ matrix

$$A_c = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ & & & \ddots & 1 \\ 0 & 0 & & & 0 \end{pmatrix},$$

and by C_c the $1 \times n$ matrix $C_c = (1, 0, \ldots, 0)$.

Definition 2.1. *Let E be a finite-dimensional vector space; let $A \in L(E)$ be an endomorphism of E and $C \in L(E, \mathbb{R})$ be a linear form. Let I be an open interval and $\varphi \in \mathcal{C}^\infty(I \times \mathbb{R}; E)$. The system Σ_L,*

$$\Sigma_L = \begin{cases} \dot{z} = Az + \varphi(Cz, u), \\ \xi = Cz, \end{cases} \tag{2.2}$$

defined on the pullback L of I by C is said to be linear up to an output injection.

When $E = \mathbb{R}^n$, the linear mappings A and C are identified with their matrices in the canonical basis, and Σ_L is said to be in canonical form if $A = A_c$ and $C = C_c$.

Remarks.

1. When a linear up to an output injection system Σ_{Lc} is in canonical form, the output function is $z \longmapsto z_1$ and the state space is $L = I \times \mathbb{R}^{n-1}$.
2. We will say that Σ_L is observable if so is the pair (C, A). It is easy to check that as well as in the linear case, whenever Σ_L is observable, it is observable for every input in the sense that for a L^∞ input being given, any two different states are distinguished by the output on any nontrivial time interval. Notice that, in particular, Σ_{Lc} is observable.

Let $u \in L^\infty([0, T_u[)$ be an input and let $x \in X$ (resp., $z \in E$). We denote by $x(t)$ (resp., $z(t)$) the solution of (2.1) (resp., (2.2)) for this input that verifies the initial condition $x(0) = x$ (resp., $z(0) = z$). These solutions are, respectively, defined on $[0, T_x[$ and $[0, T_z[$.

Definition 2.2. *An* immersion *of the system Σ into the system Σ_L is a couple (τ, ψ), where τ is a $\mathcal{C}^\infty$ mapping from X into E and ψ is a $\mathcal{C}^\infty$ diffeomorphism from I_o onto $\psi(I_o) \subset I$, that verifies the following property.*

For every input $u \in L^\infty([0, T_u[)$ and for every initial condition $x \in X$, if $z = \tau(x)$, then $T_x \leq T_z$ and

$$\psi \circ h[x(t)] = Cz(t) \qquad \forall t \in [0, T_x[.$$

This definition is rather weak because only the input-output mappings are involved; for a stronger one we could require that the trajectories of Σ are applied on those of Σ_L. But it is a general fact that these two requirements are equivalent whenever Σ_L is observable for every input and the forthcoming reduction will show that Σ_L can always be so chosen.

In what follows, a system that can be immersed into a linear up to an output injection one will be called LIS. A linear up to an output injection system in canonical form will be called a LIS system in canonical form.

2.2. First reduction.

THEOREM 2.3. *If the system Σ is LIS, then it can be immersed into a linear up to an output injection system in canonical form (hence observable). Moreover, the mapping τ applies in that case the trajectories of* (2.1) *onto the trajectories of* (2.2), *i.e., for every input $u \in L^\infty([0, T_u[)$ and for every initial condition $x \in X$ the image $\tau[x(t)]$ of $x(t)$ by τ is a trajectory of* (2.2).

In order to prove this theorem we state two very simple lemmas.

LEMMA 2.4. *If the system Σ can be immersed into a linear up to an output injection system, then it can be immersed into an observable one.*

Proof of Lemma 2.4. Let us assume that (τ, ψ) is an immersion of Σ into Σ_L and let

$$V = \bigcap_{i=0}^{+\infty} \ker CA^i$$

be the unobservable subspace of (C, A). If Σ_L is not observable, then $V \neq \{0\}$, but we can define an observable, linear up to an output injection system on the quotient space E/V in the same way as in the linear case. Let p be the projection of E onto E/V and consider

$$\begin{array}{lll} \widetilde{A} \in L(E/V) & \text{defined by} & \widetilde{A}(pz) = p(Az); \\ \widetilde{C} \in L(E/V; \mathbb{R}) & \text{defined by} & \widetilde{C}(pz) = Cz; \\ \widetilde{\varphi} \in \mathcal{C}^\infty(I \times \mathbb{R}; E/V) & \text{defined by} & \widetilde{\varphi} = p \circ \varphi; \\ \widetilde{\tau} \in \mathcal{C}^\infty(X; E/V) & \text{defined by} & \widetilde{\tau} = p \circ \tau. \end{array}$$

We have to show that the couple $(\widetilde{\tau}, \psi)$ is an immersion of Σ into the system

$$\widetilde{\Sigma}_L = \begin{cases} \dot{\widetilde{z}} = \widetilde{A}\widetilde{z} + \widetilde{\varphi}(\widetilde{C}\widetilde{z}, u), \\ \xi = \widetilde{C}\widetilde{z} \end{cases} \tag{2.3}$$

in the sense of Definition 2.2.

Let $u \in L^\infty([0, T_u[)$, $x \in X$, and $z = \tau(x)$. The solution of (2.3) issued from $p(z)$ is $p[z(t)]$; therefore

$$\forall t \in [0, T_x[\qquad \widetilde{C}p[z(t)] = Cz(t) = \psi \circ h[x(t)]$$

and Lemma 2.4 is proved. □

LEMMA 2.5. *We keep the conditions and notations of Lemma* 2.4. *Then for every $u \in L^\infty([0, T_u[)$ and $x \in X$*

$$\widetilde{\tau}[x(t)] = p[z(t)].$$

Proof of Lemma 2.5. Let $u \in L^\infty([0, T_u[)$ and $x \in X$. We want to show that

$$\forall t \in [0, T_x[\qquad \widetilde{\tau}[x(t)] = p[z(t)].$$

Let $t_0 \in [0, T_x[$, let $\widetilde{z}_0 = \widetilde{\tau}(x(t_0))$, and let $t \longmapsto \widetilde{z}(t)$ be the trajectory of (2.3) issued from $\widetilde{z}_0$ for the input v defined on $[0, T_u - t_0[$ by $v(t) = u(t_0 + t)$. We have

$$\begin{aligned} \forall t \in [0, T_x - t_0[\qquad \widetilde{C}\widetilde{z}(t) &= \psi \circ h[x(t_0 + t)] \\ &= Cz(t_0 + t) \\ &= \widetilde{C}p[z(t_0 + t)]. \end{aligned}$$

But the system $\widetilde{\Sigma}_L$ is observable and this equality implies

$$\widetilde{z}(0) = p[z(t_0)].$$

Therefore

$$\widetilde{\tau}(x(t_0)) = p[z(t_0)]. \qquad \square$$

Proof of Theorem 2.3. The system $\widetilde{\Sigma}_L$ being observable, so is the pair (C, A). We know by the linear theory that we can choose a basis in E/V in which

$$\widetilde{A} \equiv \begin{pmatrix} * & 1 & 0 & \cdots & 0 \\ * & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ & & & \ddots & 1 \\ * & 0 & & & 0 \end{pmatrix} \qquad \text{and } \widetilde{C} \equiv (1, 0, \ldots, 0).$$

The first column of $\widetilde{A}$ is a linear function of the output and can be added to $\widetilde{\varphi}$ in order to obtain in $\mathbb{R}^{\dim E/V}$ a system in canonical form. This choice of coordinates induces an isomorphism between E/V and $\mathbb{R}^{\dim E/V}$; $\widetilde{\tau}$ followed by this isomorphism, together with ψ, is clearly a suitable immersion. $\square$

Observability. This first reduction has some immediate consequences on the observability of LIS systems. Let us recall that an input is universal if it distinguishes between any two states that can be distinguished by at least one input (see [Sus79]). As a first result of Theorem 2.3 we can see at once that every input is universal for a LIS system. In particular, an observable and LIS system is observable for every input in the sense defined in the above remarks.

However, the universality of all inputs is not a characterization of LIS systems. Pick any uncontrolled system in observable form: it can be the drift part of a uniformly observable control affine system (see [GHO92]) without being LIS.

Despite this fact, no assumption of observability will be needed in the forthcoming results.

2.3. The uncontrolled case.

2.3.1. Main result. In this section we deal with uncontrolled systems:

$$\Sigma_u = \begin{cases} \dot{x} = f(x), \\ y = h(x). \end{cases} \tag{2.4}$$

THEOREM 2.6. *The uncontrolled system Σ_u is LIS if and only if there exist a function ψ belonging to $\mathcal{C}^\infty(I_o,\mathbb{R})$, an integer n, and n functions φ_1, $\varphi_2,\ldots,$ φ_n belonging to $\mathcal{C}^\infty(\psi(I_o),\mathbb{R})$ such that*

$$(F)\qquad L_f^n\tilde{h} = L_f^{n-1}(\varphi_1\circ\tilde{h}) + L_f^{n-2}(\varphi_2\circ\tilde{h}) + \cdots + L_f(\varphi_{n-1}\circ\tilde{h}) + \varphi_n\circ\tilde{h},$$

where $\tilde{h} = \psi\circ h$.

Proof of Theorem 2.6. Let us assume that Σ_u is LIS. Then, by Theorem 2.3, it can be immersed into a canonical system:

$$(2.5)\qquad \Sigma_{uc} = \begin{cases} \dot{z} = A_c z + \varphi(C_c z), \\ \xi = C_c z \end{cases} = \begin{cases} \dot{z}_1 &= z_2 + \varphi_1(z_1), \\ \dot{z}_2 &= z_3 + \varphi_2(z_1), \\ &\vdots \\ \dot{z}_{n-1} &= z_n + \varphi_{n-1}(z_1), \\ \dot{z}_n &= \varphi_n(z_1), \\ \xi = z_1. \end{cases}$$

Let (τ,ψ), with $\tau = (\tau_1,\ldots,\tau_n)$, be the immersion of Σ_u into Σ_{uc}. We have $\psi\circ h = C_c\circ\tau = \tau_1$.

Let us assume that for an index k, $1\le k<n$, the following equality holds:

$$(2.6)\qquad \tau_k = L_f^{k-1}\tilde{h} - L_f^{k-2}(\varphi_1\circ\tilde{h}) - \cdots - L_f(\varphi_{k-2}\circ\tilde{h}) - \varphi_{k-1}\circ\tilde{h}.$$

Let $x\in X$ and let us denote by $x(t)$ the solution of $\dot{x}=f(x)$ that verifies $x(0)=x$. Then

$$\begin{aligned}\tau_{k+1}(x) &= \frac{d}{dt}\tau_k(x(t))|_{t=0} - \varphi_k(\psi\circ h(x)) \\ &= L_f^k\tilde{h}(x) - L_f^{k-1}(\varphi_1\circ\tilde{h})(x) - \cdots - L_f(\varphi_{k-1}\circ\tilde{h})(x) - (\varphi_k\circ\tilde{h})(x).\end{aligned}$$

Moreover,

$$\begin{aligned}\varphi_n(\tilde{h}(x)) &= \frac{d}{dt}\tau_n(x(t))|_{t=0} \\ &= L_f^n\tilde{h}(x) - L_f^{n-1}(\varphi_1\circ\tilde{h})(x) - \cdots - L_f(\varphi_{n-1}\circ\tilde{h})(x)\end{aligned}$$

and formula (F) holds.

Conversely, if formula (F) holds, let

$$\begin{cases} \tau_1 = \psi\circ h, \\ \tau_{k+1} = L_f\tau_k - \varphi_k\circ\psi\circ h \qquad \text{for } 1\le k<n. \end{cases}$$

Let $\tau=(\tau_1,\ldots,\tau_n)$ and $\tilde{h}=\psi\circ h$.

For $k=1,\ldots,n-1$ we have by induction

$$\frac{d}{dt}\tau_k(x(t)) = L_f\tau_k(x(t)) = \tau_{k+1}(x(t)) + \varphi_k(\tilde{h}(x(t)));$$

hence

$$\tau_{k+1} = L_f^k\tilde{h} - L_f^{k-1}(\varphi_1\circ\tilde{h}) - \cdots - L_f(\varphi_{k-1}\circ\tilde{h}) - \varphi_k\circ\tilde{h}.$$

In the end

$$\begin{aligned}\frac{d}{dt}\tau_n(x(t)) &= L_f\tau_n(x(t))\\ &= \left(L_f^n\tilde{h} - L_f^{n-1}(\varphi_1\circ\tilde{h}) - L_f^{n-2}(\varphi_2\circ\tilde{h}) - \cdots - L_f(\varphi_{n-1}\circ\tilde{h})\right)(x(t))\\ &= \varphi_n(\tilde{h}(x(t)))\end{aligned}$$

and (τ,ψ) is an immersion of Σ_u into

$$\Sigma_{uc} = \begin{cases} \dot{z} = A_c z + \varphi(C_c z),\\ \xi = z_1, \end{cases}$$

where $\varphi = {}^T(\varphi_1,\ldots,\varphi_n)$. □

Remarks.

1. Clearly the functions $\varphi_1,\ldots,\varphi_{n-1}$ are defined up to a constant. But they need not be unique, even up to a constant, as the forthcoming Example 1 will show.
2. Formula (2.6), which gives τ_k as a function of the φ_i's, will be useful later to do the effective computations.

2.3.2. Existence of LIS systems. At a point x^0 where $f(x^0)\neq 0$ we can choose coordinates $(x_1,\ldots,x_d)$ such that $f\equiv\frac{\partial}{\partial x_1}$ in a neighborhood of x^0. In these coordinates, and in the case where $\psi = Id$, formula (F) becomes

$$(F') \qquad \frac{\partial^n}{\partial x_1^n}h = \frac{\partial^{n-1}}{\partial x_1^{n-1}}(\varphi_1\circ h) + \cdots + \frac{\partial}{\partial x_1}(\varphi_{n-1}\circ h) + \varphi_n\circ h.$$

Now we can consider $\varphi_1,\ldots,\varphi_n$ (defined on an interval I) as data and (F') as a differential equation whose unknown is h. Setting $\widehat{x} = (x_2,\ldots,x_d)$, we can choose initial conditions x_1^0 and $y_k(\widehat{x})$, for $k=0,\ldots,n-1$, defined in an open subset U of I^{d-1} and smoothly depending on $\widehat{x}$. Then there exists a smooth function h defined in a neighborhood of $\{x_1^0\}\times U$ that is solution of (F') and verifies

$$\frac{\partial^k h}{\partial x_1^k}(x_1^0,\widehat{x}) = y_k(\widehat{x}) \qquad \text{for } k=0,\ldots,n-1.$$

Thus for any choice of the integer n and of the functions φ_i, there exist functions h that verify equation (F'). All these solutions of (F'), possibly composed with diffeomorphisms ψ, form the class of the universal solutions of the LIS problem at a regular point of the vector field. Notice that we can, moreover, choose the $y_k(\widehat{x})$'s in such a way that the system is observable (if $n\geq d$), at least in a neighborhood of $\{x_1^0\}\times U$.

2.4. The control-affine case. In this section we consider systems whose vector field is affine w.r.t. the control. The general setting of these systems is

$$\Sigma = \begin{cases} \dot{x} = f(x) + ug(x),\\ y = h(x), \end{cases} \tag{2.7}$$

where f and g belong to $V^\infty(X)$, the set of C^∞-vector fields on X.

Let us assume that Σ is LIS and can be immersed into the system in canonical form:

$$S = \begin{cases} \dot{z} = A_c z + \varphi(C_c z) + u\gamma(C_c z), \\ \xi = C_c z, \end{cases} \tag{2.8}$$

where $z \in \mathbb{R}^n$.

Clearly the uncontrolled system

$$\Sigma U = \begin{cases} \dot{x} = f(x), \\ y = h(x) \end{cases} \tag{2.9}$$

is also LIS and can be immersed into

$$SU = \begin{cases} \dot{z} = A_c z + \varphi(C_c z), \\ \xi = C_c z. \end{cases} \tag{2.10}$$

This remark leads to the following statement.

THEOREM 2.7. *The control-affine system Σ is LIS if and only if the following hold:*

1. *The drift system ΣU is LIS.*
2. *If we denote by (τ, ψ), with $\tau = (\tau_1, \dots, \tau_n)$, the immersion of ΣU into a LIS system in canonical form in $\mathbb{R}^n$, then there exist smooth functions γ_i, $i = 1, \dots, n$, such that*

$$L_g \tau_i = \gamma_i \circ (\psi \circ h).$$

Moreover, at points x such that $dh(x) \neq 0$, condition 2 is locally equivalent to

$$dL_g \tau_i \wedge dh = 0.$$

Proof of Theorem 2.7.

(i) Proof of the necessity part. As we have just remarked, if (τ, ψ) is an immersion of Σ into S, then it is also an immersion of ΣU into SU and condition 1 holds. Let us show that condition 2 holds as well.

Let u be a constant control and $x \in X$. Let $x(t)$ be the solution of Σ for the constant input u and for the initial condition $x(0) = x$. By Theorem 2.3 the solution of S for the input u and the initial condition $z(0) = \tau(x)$ is $z(t) = \tau[x(t)]$.

Differentiating $\tau_i[x(t)] = z_i(t)$ at $t = 0$, we get for all x, u, and i

$$L_f \tau_i(x) + uL_g \tau_i(x) = L_F z_i(\tau(x)) + uL_\gamma z_i(\tau(x)),$$

where F (resp., γ) stands for the vector field $z \longmapsto A_c z + \varphi(C_c z)$ (resp., $z \longmapsto \gamma(C_c z)$) in $\mathbb{R}^n$.

Therefore

$$\begin{aligned} \forall x \in X \qquad L_g \tau_i(x) &= L_\gamma z_i(\tau(x)) \\ &= \gamma_i(C_c \tau(x)) \\ &= \gamma_i(\psi \circ h)(x), \end{aligned}$$

where γ_i denotes the ith component of γ.

(ii) Sufficiency. We assume now that conditions 1 and 2 hold. By condition 1, and because SU is observable, we know by Theorem 2.3 that the trajectories of the uncontrolled system ΣU are applied by τ on those of SU. Therefore we have

$$\forall x \in X \qquad L_f \tau_i(x) = L_F z_i(\tau(x)),$$

where again F stands for the vector field $z \longmapsto A_c z + \varphi(C_c z)$.

We also know that $C_c \circ \tau = \psi \circ h$ and, in order to prove that (τ, ψ) is an immersion of Σ into S, it is enough to show that for any input the mapping τ applies the trajectories of Σ onto those of S, where the vector field γ is defined by

$$\gamma(z) = (\gamma_1(C_c z), \gamma_2(C_c z), \ldots, \gamma_n(C_c z)) .$$

Let $u \in L^\infty[0, T_u[$ be an input and $t \mapsto x(t)$ a trajectory of Σ for this input.

We have almost everywhere

$$\begin{aligned} \frac{d}{dt}\tau_i[x(t)] &= L_f \tau_i[x(t)] + u(t) L_g \tau_i[x(t)] \\ &= L_F z_i[\tau(x(t))] + u(t)(\gamma_i \circ (\psi \circ h))[x(t)] \\ &= L_F z_i[\tau(x(t))] + u(t)\gamma_i(C_c \tau)[x(t)]. \end{aligned}$$

This computation proves that $\tau[x(t)]$ is a solution of S and ends the proof.

(iii) Regular points. The condition

$$L_g \tau_i = \gamma_i \circ (\psi \circ h)$$

always implies

$$dL_g \tau_i \wedge dh = d(\gamma_i \circ (\psi \circ h)) \wedge dh = 0.$$

Conversely let us assume that locally $dh(x) \neq 0$ and

$$dL_g \tau_i \wedge dh = 0.$$

Let us choose local coordinates $x_1, x_2, \ldots, x_d$ such that $h(x) = x_1$. In these coordinates, $L_g \tau_i$ is a function of x_1, hence, a function γ_i of $\psi \circ h = \psi(x_1)$. □

Remarks.

1. The conditions stated in Theorem 2.7 are very strong. Let us, for instance, assume that τ is an embedding from X into $\mathbb{R}^n$; $\tau(X)$ is then a submanifold of $\mathbb{R}^n$. The value of $\tau_* g$ at a point $z \in \tau(X)$ depends only on the value z_1 of the first coordinate of z. Thus all the subspaces of $\mathbb{R}^n$ tangent to $\tau(X)$ at points z such that $z_1 = z_1^0$ contain $\tau_* g(z^0)$ and, if we assume $\tau_* g(z^0) \neq 0$, their intersection must be different from $\{0\}$.
   These conditions, somewhat hidden, are used further to compute the immersion τ in the control-affine case.
2. In the paper [BRG89] the immersion of a control-affine system into a linear up to an output injection one is studied in the case where the drift system is linearizable. The necessary and sufficient conditions stated in that paper seem to be more intrinsic than those of Theorem 2.7 because they do not involve the immersion τ of the drift system. More specifically, the drift system is assumed to be linearizable, and the conditions stated in [BRG89] turn out to be

$$d\left(L_g L_f^k h\right) \wedge dh = 0 \quad \forall k.$$

But these conditions are no longer true in the more general case dealt with herein.

Example.

$$\Sigma = \begin{cases} \dot{z}_1 &= z_2 + z_1^2 + uz_1, \\ \dot{z}_2 &= z_3 + z_1^2 + uz_1, \\ & . \\ & . \\ \dot{z}_{n-1} &= z_n + z_1^2 + uz_1, \\ \dot{z}_n &= z_1^2 + uz_1, \\ y &= h(z) = z_1. \end{cases} \tag{2.11}$$

The system is LIS. However,

$$\begin{aligned} L_f h(z) &= z_2 + z_1^2, \\ L_f^2 h(z) &= z_3 + z_1^2 + 2z_1(z_2 + z_1^2) \\ &= z_3 + z_1^2 + 2z_1^3 + 2z_1 z_2, \\ L_g L_f^2 h(z) &= (1 + 2z_1 + 6z_1^2 + 2z_1 + 2z_2)z_1 \\ &= z_1 + 4z_1^2 + 6z_1^3 + 2z_1 z_2. \end{aligned}$$

Hence

$$dL_g L_f^2 h \wedge dh = -2z_1 dz_1 \wedge dz_2 \neq 0.$$

3. Computational results. The purpose of this section is to actually check if a given system is LIS and to compute the functions φ_k or, equivalently, the coordinate functions τ_k of the immersion in an LIS system in canonical form.

We shall always start with a system whose drift part is in observable form, that is, a system in the form

$$\Sigma = \begin{cases} \dot{x}_1 &= x_2 &+ ug_1(x), \\ \dot{x}_2 &= x_3 &+ ug_2(x), \\ \cdots & & \cdots \\ \cdots & & \cdots \\ \dot{x}_d &= \Phi(x) &+ ug_d(x), \\ y &= h(x) &= x_1, \end{cases} \tag{3.1}$$

where $x = (x_1, \ldots, x_d)$ and where d is the dimension of the state space X. This is of course a restriction because generically there exist points of the state space where the observable form can be obtained only in dimension strictly greater than the dimension of the state space, but at the present time we do not know how to do the computations at these points. Moreover, we shall always assume the function ψ that appears in Definition 2.2 to be equal to the identity of $\mathbb{R}$.

The computations are easier in the controlled case, but they make use of a general form of certain extensions of the drift system, and in the first place we are going to study uncontrolled systems.

3.1. The uncontrolled case. As we have just said, we assume that the system Σ is in observable form:

$$\Sigma = \begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = x_3, \\ \cdots \\ \dot{x}_d = \Phi(x_1, x_2, \dots, x_d), \\ y = h(x) = x_1. \end{cases} \tag{3.2}$$

The requirement $d = \dim X$, i.e., the requirement that the observable form is obtained by diffeomorphism, is very important because in that case the function Φ is unique. If Σ can be embedded in the system

$$\Sigma_n = \begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = x_3, \\ \cdots \\ \dot{x}_n = \theta(x_1, x_2, \dots, x_n), \\ y = h(x) = x_1, \end{cases} \tag{3.3}$$

we will say that Σ_n is *an extension at the order* n of Σ. If Σ_n is, moreover, LIS by diffeomorphism, we will say that Σ_n is *a LIS extension at the order* n of Σ. We can use existing algorithms to check whether Σ_n is LIS by diffeomorphism or not (see, for instance, [KI83], [Phe91], [GMP96]). But of course such an extension at order n is not unique as soon as $n > d$ and the problem is to choose the function θ in order to get an LIS one if it exists.

Let us denote by

$$f = \sum_{i=1}^{d-1} x_{i+1} \frac{\partial}{\partial x_i} + \Phi(x_1, x_2, \dots, x_d) \frac{\partial}{\partial x_d}$$

the vector field of (3.2) and by

$$F = \sum_{i=1}^{n-1} x_{i+1} \frac{\partial}{\partial x_i} + \theta(x_1, x_2, \dots, x_n) \frac{\partial}{\partial x_n}$$

the one of (3.3).

The first thing to notice is that (3.3) is an extension at the order n of (3.2) if and only if

$$\theta(x, L_f^d h(x), L_f^{d+1} h(x), \dots, L_f^{n-1} h(x)) = L_f^n h(x), \tag{3.4}$$

where $x = (x_1, x_2, \dots, x_d)$.

If, moreover, (3.3) is LIS by a change of variables, we can find functions φ_i, $i = 1, \dots, n$, such that

$$\theta(x_1, \dots, x_n) = L_F^{n-1}[\varphi_1(x_1)] + \cdots + L_F[\varphi_{n-1}(x_1)] + \varphi_n(x_1). \tag{3.5}$$

In this formula, we can compute the terms where $x_{d+1}, \dots, x_n$ appear, and then, combining this with formula (3.4), we can obtain the general form of θ in the case where (3.3) is an LIS extension of (3.2).

Let us first consider the case $n = d + 1$.

3.1.1. Extension at order $d+1$.

PROPOSITION 3.1. *If* (3.2) *admits an LIS extension* (3.3) *at order* $d+1$*, then the mapping* θ *that appears in* (3.3) *has the following form:*

$$\theta = L_f^{d+1}h(x) + \varphi_1'(x_1)\left(x_{d+1} - L_f^d h(x)\right),$$

where $x = (x_1, x_2, \dots, x_d)$*. Moreover,* φ_1' *is equal to*

$$\varphi_1' = \frac{\partial^2 L_f^{d+1}h}{\partial x_d^2}\left(\frac{\partial^2 L_f^d h}{\partial x_d^2}\right)^{-1}$$

if $\frac{\partial^2}{\partial x_d^2}L_f^d h(x) \neq 0$ *and* $d > 2$*. Otherwise* φ_1' *is solution of the first order linear differential equation*

$$\varphi_1'' + \varphi_1'\frac{\partial^2}{\partial x_2 \partial x_d}L_f^d h = \frac{\partial^2}{\partial x_2 \partial x_d}L_f^{d+1}h$$

if $d > 2$*, and*

$$2\varphi_1'' + \varphi_1'\frac{\partial^2}{\partial x_2^2}L_f^d h = \frac{\partial^2}{\partial x_2^2}L_f^{d+1}h$$

if $d = 2$.

Proof of Proposition 3.1.

1. Let us prove the first assertion: the only term of formula (3.5) in which x_{d+1} appears is $L_F^d[\varphi_1(x_1)]$ (see the appendix):

$$L_F^d[\varphi_1(x_1)] = x_{d+1}\varphi_1'(x_1) + \text{terms in } x_1, \dots, x_d.$$

Therefore we have

$$\theta(x, x_{d+1}) = x_{d+1}\varphi_1'(x_1) + \Lambda(x),$$

where Λ does not depend on x_{d+1}, and

$$L_f^{d+1}h(x) = \theta(x, L_f^d h(x)) = L_f^d h(x)\varphi_1'(x_1) + \Lambda(x).$$

Consequently

$$\Lambda(x) = L_f^{d+1}h(x) - L_f^d h(x)\varphi_1'(x_1)$$

and

$$\theta(x, x_{d+1}) = L_f^{d+1}h(x) + \varphi_1'(x_1)\left(x_{d+1} - L_f^d h(x)\right).$$

2. Let us now compute φ_1'. In formula (3.5) x_d appears only in the terms

$$L_F^d[\varphi_1(x_1)] = x_{d+1}\varphi_1'(x_1) + x_2 x_d \varphi_1''(x_1) + \text{terms in } x_1, \dots, x_{d-1}$$

and

$$L_F^{d-1}[\varphi_2(x_1)] = x_d\varphi_2'(x_1) + \text{terms in } x_1, \dots, x_{d-1}.$$

Replacing x_{d+1} by $L_f^d h(x)$ we get

$$L_f^{d+1}h(x) = L_f^d h(x)\varphi_1'(x_1) + x_2 x_d \varphi_1''(x_1) + x_d \varphi_2'(x_1) + \text{terms in } x_1, \ldots, x_{d-1}$$

and then, differentiating w.r.t. x_d,

$$\frac{\partial L_f^{d+1}h}{\partial x_d} = \frac{\partial L_f^d h}{\partial x_d}\varphi_1'(x_1) + x_2\varphi_1''(x_1) + \varphi_2'(x_1)$$

if $d > 2$, and

$$\frac{\partial L_f^3 h}{\partial x_2} = \frac{\partial L_f^2 h}{\partial x_2}\varphi_1'(x_1) + 2x_2\varphi_1''(x_1) + \varphi_2'(x_1)$$

if $d = 2$.

If $d > 2$ let us differentiate one more time w.r.t x_d. We get

$$\frac{\partial^2 L_f^{d+1}h}{\partial x_d^2} = \frac{\partial^2 L_f^d h}{\partial x_d^2}\varphi_1'$$

and, if $\frac{\partial^2}{\partial x_d^2}L_f^d h(x) \neq 0$, this equality gives the result.

Otherwise, differentiating w.r.t x_2 we get

$$\varphi_1'' + \varphi_1' \frac{\partial^2}{\partial x_2 \partial x_d} L_f^d h = \frac{\partial^2}{\partial x_2 \partial x_d} L_f^{d+1} h$$

if $d > 2$, and

$$2\varphi_1'' + \varphi_1' \frac{\partial^2}{\partial x_2^2} L_f^2 h = \frac{\partial^2}{\partial x_2^2} L_f^3 h$$

if $d = 2$. □

Example 1. Let us consider the system defined in $]-1, +\infty[\times\mathbb{R}$ by

$$\left\{ \begin{array}{l} \dot{x}_1 = x_2, \\ \dot{x}_2 = \frac{x_2^2}{2x_1+2} + x_1 x_2 + x_2 + 1, \\ y = h(x_1, x_2) = x_1. \end{array} \right. \tag{3.6}$$

Thus we have

$$L_f^2 h(x) = \frac{x_2^2}{2x_1 + 2} + x_1 x_2 + x_2 + 1$$

and we can first remark that the system (2.11) is not LIS by diffeomorphism; hence in dimension 2, in any open subset of $]-1, +\infty[\times\mathbb{R}$, because we would have in that case

$$\begin{aligned} \Phi(x_1, x_2) = L_f^2 h(x) &= L_f[\varphi_1(h(x))] + \varphi_2(h(x)) \\ &= \varphi_1'(x_1)x_2 + \varphi_2(x_1), \end{aligned}$$

one has

$$L_f^3 h(x) = \frac{5}{2}x_2^2 + \frac{x_2}{x_1+1} + x_1^2 x_2 + 2x_1 x_2 + x_1 + x_2 + 1,$$

$$\frac{\partial L_f^3 h}{\partial x_2}(x) = 5x_2 + \frac{1}{x_1+1} + x_1^2 + 2x_1 + 1,$$

$$\frac{\partial^2 L_f^3 h}{\partial x_2^2}(x) = 5.$$

If the system is LIS at order 3, then φ_1 must verify

$$\varphi_1'' + \frac{1}{2}\frac{\partial^2 L_f^2 h}{\partial x_2^2}\varphi_1' = \frac{1}{2}\frac{\partial^2 L_f^3 h}{\partial x_2^2},$$

that is,

$$\varphi_1'' + \frac{1}{2(x_1+1)}\varphi_1' = \frac{5}{2}.$$

The general solution of this equation is

$$\varphi_1' = \frac{C}{\sqrt{x_1+1}} + \frac{5}{3}(x_1+1),$$

where $C \in \mathbb{R}$, and we set

$$\theta(x_1, x_2, x_3) = L_f^3 h + \varphi_1'(x_3 - \Phi).$$

In order to check whether the system is LIS at order 3 and to compute φ_2 and φ_3, we have to compute $\theta - L_F^2(\varphi_1(x))$, where F is the vector field of the system

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = x_3, \\ \dot{x}_3 = \theta(x_1, x_2, x_3), \\ y = x_1. \end{cases}$$

We get

$$\begin{aligned} \theta - L_F^2(\varphi_1(x_1)) &= x_2\left(\tfrac{1}{x_1+1} + x_1^2 + 2x_1 + 1 - \tfrac{5}{3}(x_1+1)^2\right) \\ &\quad + x_1 + 1 - \tfrac{5}{3}(x_1+1) \\ &\quad - \tfrac{C}{\sqrt{x_1+1}}(x_1x_2 + x_2 + 1). \end{aligned}$$

Let us take $C = 0$, hence $\varphi_1' = \frac{5}{3}(x_1+1)$, and

$$\theta - L_F^2(\varphi_1(x_1)) = x_2\left(\frac{1}{x_1+1} - \frac{2}{3}x_1^2 - \frac{4}{3}x_1 - \frac{2}{3}\right) - \frac{2}{3}(x_1+1).$$

We obtain

$$\begin{aligned} &\varphi_2 = \ln(x_1+1) - \tfrac{2}{9}x_1^3 - \tfrac{2}{3}x_1^2 - \tfrac{2}{3}x_1, \\ &\varphi_3 = -\tfrac{2}{3}(x_1+1). \end{aligned}$$

At the end the system can be immersed into

$$\begin{cases} \dot{z}_1 = z_2 + \frac{5z_1^2}{6} + \frac{5z_1}{3}, \\ \dot{z}_2 = z_3 + \ln(z_1+1) - \frac{2}{9}z_1^3 - \frac{2}{3}z_1^2 - \frac{2}{3}z_1, \\ \dot{z}_3 = \quad -\frac{2}{3}(z_1+1), \end{cases}$$

the immersion being given by

$$\begin{aligned} &\tau_1 = x_1, \\ &\tau_2 = x_2 - \frac{5x_1^2}{6} - \frac{5x_1}{3}, \\ &\tau_3 = \frac{x_2^2}{2x_1+2} - \frac{2}{3}x_1x_2 - \frac{2}{3}x_2 - \ln(x_1+1) + \frac{2}{9}x_1^3 + \frac{2}{3}x_1^2 + \frac{2}{3}x_1 + 1. \end{aligned}$$

In order to verify the result, one can differentiate τ_1, τ_2, and τ_3. As expected we obtain

$$\begin{cases} \dot{\tau_1} = \tau_2 + \frac{5y^2}{6} + \frac{5y}{3}, \\ \dot{\tau_2} = \tau_3 + \ln(y+1) - \frac{2}{9}y^3 - \frac{2}{3}y^2 - \frac{2}{3}y, \\ \dot{\tau_3} = \quad -\frac{2}{3}(y+1). \end{cases}$$

Remark on the nonuniqueness. In fact it is possible to give any value to the integration constant C of the differential equation. Let us set

$$\varphi' = \frac{C}{\sqrt{x_1+1}}.$$

This function φ' verifies the homogeneous equation

$$\varphi'' + \frac{1}{2(x_1+1)}\varphi' = 0;$$

hence the degree of

$$L_f^2\varphi = (x_1x_2 + x_2 + 1)\varphi'$$

w.r.t. x_2 is one. Therefore we can add φ' to φ_1' if we modify φ_2 and φ_3. This is due to the fact that the degree of Φ w.r.t. x_2 is 2: if φ' is a solution of the homogeneous equation

$$\varphi'' + \frac{1}{2}\frac{\partial\Phi}{\partial x_2^2}\varphi' = 0,$$

then $L_f^2\varphi$ verifies

$$L_f^2\varphi = \left(-\frac{x_2^2}{2}\frac{\partial\Phi}{\partial x_2^2} + \Phi\right)\varphi',$$

where the second degree term w.r.t. x_2 clearly vanishes.

3.1.2. Extension at order $n \geq d+2$. We can compute the general form of the function θ of an LIS extension in the same manner as in the case of the extension at order $d+1$. Each term of the formula

$$\theta(x_1, \dots, x_n) = L_F^{n-1}[\varphi_1(x_1)] + \cdots + L_F[\varphi_{n-1}(x_1)] + \varphi_n(x_1)$$

is a polynomial in $x_{d+1}, \dots, x_n$ with coefficients in $\mathcal{C}^\infty(\mathbb{R})[x_2, \dots, x_d]$. Therefore θ can be written

$$\theta(x, x_{d+1}, \dots, x_n) = P_{n,d}(x_{d+1}, \dots, x_n) + \Lambda(x),$$

where $P_{n,d}$ is a polynomial in the variables $x_{d+1}, \dots, x_n$ without constant term. These polynomials are universal and their computation is postponed to the appendix. Now we have also

$$\theta(x, L_f^d h(x), L_f^{d+1}h(x), \dots, L_f^{n-1}h(x)) = L_f^n h(x);$$

hence

$$L_f^n h(x) = P_{n,d}(L_f^d h(x), L_f^{d+1}h(x), \dots, L_f^{n-1}h(x)) + \Lambda(x)$$

and

$$\Lambda(x) = L_f^n h(x) - P_{n,d}(L_f^d h(x), L_f^{d+1} h(x), \dots, L_f^{n-1} h(x)).$$

In the end we get

$$\begin{aligned} \theta(x, x_{d+1}, \dots, x_n) &= L_f^n h(x) + P_{n,d}(x_{d+1}, \dots, x_n) \\ &\quad - P_{n,d}(L_f^d h(x), L_f^{d+1} h(x), \dots, L_f^{n-1} h(x)). \end{aligned} \tag{3.7}$$

In particular, if $n < 2d$, the degree of the polynomial $P_{n,d}$ w.r.t. each of the variables $x_{d+1}, \dots, x_n$ is one (see the appendix):

$$P_{n,d} = x_{d+1} Q_{n,d}^{d+1}(x) + \cdots + x_n Q_{n,d}^n(x),$$

where $Q_{n,d}^k \in \mathcal{C}^\infty(\mathbb{R})[x_2, \dots, x_d]$, $k = d+1, \dots, n$, and θ becomes

$$\theta(x, x_{d+1}, \dots, x_n) = L_f^n h(x) + \sum_{k=d+1}^{n} (x_k - L_f^{k-1} h(x)) Q_{n,d}^k(x).$$

At this point we would like to get some equation allowing us to compute at least φ_1 or one of its derivatives. But whenever $n > d+1$, it turns out that the above formulas lead to partial differential equations, involving not only at least φ_1 and φ_2, but also the equivalent of the function θ in an intermediate extension, and we do not know how to solve these equations in the uncontrolled case.

Fortunately there are additional conditions in the controlled one and these can be used to compute the functions φ_i in most cases.

3.2. The controlled case. We consider controlled systems in observable form,

$$\Sigma = \begin{cases} \dot{x}_1 = x_2 & + u g_1(x), \\ \dot{x}_2 = x_3 & + u g_2(x), \\ \cdots & \cdots \\ \cdots & \cdots \\ \dot{x}_d = \Phi(x_1, x_2, \dots, x_d) & + u g_d(x), \\ y = h(x) = x_1, & \end{cases} \tag{3.8}$$

where d is again equal to the dimension of the state space X.

Following the results of section 2.4, if there exists an embedding $\tau = (\tau_1, \dots, \tau_n)$ into an n-dimensional system in LIS canonical form, then this mapping must verify

$$dL_g \tau_i \wedge dx_1 = 0 \qquad \text{for } i = 1, \dots, n.$$

But, by section 2.3, we also know that

$$\tau_k = L_f^{k-1} h - L_f^{k-2}(\varphi_1 \circ h) - \cdots - L_f(\varphi_{k-2} \circ h) - \varphi_{k-1} \circ h$$

for $k = 1, \dots, n$.

Let us first notice the following proposition.

PROPOSITION 3.2. *If Σ is LIS, then*

(i) *g_1 and g_2 are functions of x_1 only;*

(ii) *for $k = 3, \dots, d$, g_k depends on $x_1, \dots, x_{k-1}$ only; more specifically g_k is a polynomial in the variables $x_2, \dots, x_{k-1}$ and its degree is one w.r.t. x_{k-1}.*

In particular, g does not depend on x_d.

Proof of Proposition 3.2. The proof merely uses the fact that $L_g \tau_k$ is a function of x_1 only.

- $\tau_1 = x_1$; hence $L_g\tau_1 = g_1$ and g_1 is a function of x_1 only.
- $\tau_2 = L_f h - \varphi_1(x_1) = x_2 - \varphi_1(x_1)$; hence $L_g\tau_2 = g_2 - \varphi_1'(x_1)g_1(x_1)$ and $g_2 = L_g\tau_2 + \varphi_1'(x_1)g_1(x_1)$ is a function of x_1 only.
- Let $3 \le k \le d$ and let us assume the property true for $g_1, \dots, g_{k-1}$. We have

$$\begin{aligned}\tau_k &= x_k - L_f^{k-2}(\varphi_1(x_1)) - \cdots - L_f(\varphi_{k-2}(x_1)) - \varphi_{k-1}(x_1)\\ &= x_k - x_{k-1}\varphi_1'(x_1) + \text{terms without } x_k, x_{k-1}.\end{aligned}$$

  Therefore

$$\begin{aligned}L_g\tau_k = g_k - x_{k-1}\varphi_1''(x_1)g_1(x_1) - \varphi_1'(x_1)g_{k-1}(x_1, \dots, x_{k-2})\\ + \text{terms in } x_1, \dots, x_{k-2} \text{ and } g_1, \dots, g_{k-2}.\end{aligned}$$

  The fact that g_k is a polynomial in the variables $x_2, \dots, x_{k-1}$ comes from the similar property for the functions $L_f^p(\varphi_i(x_1))$. □

Remarks.

1. The independence of g on x_d is crucial in the forthcoming computations.
2. The necessary conditions of Proposition 3.2 are stronger than the criterion of uniform observability stated in [GHO92]: a control affine SISO system whose drift part is in observable form is uniformly observable if and only if g_k is a function of $x_1, \dots, x_k$ only for $k = 1, \dots, d$.

Now the same kind of computations of the τ_k's for $k \ge d+1$ will give conditions on the φ_i's. For instance,

$$\begin{aligned}\tau_{d+1} &= L_f^d h - L_f^{d-1}(\varphi_1(x_1)) - \cdots - \varphi_d(x_1)\\ &= L_f^d h - x_d\varphi_1'(x_1) + \text{ terms without } x_d.\end{aligned}$$

Therefore

$$L_g\tau_{d+1} = L_g L_f^d h - x_d\varphi_1''(x_1)g_1(x_1) + \text{ terms without } x_d.$$

Since $L_g\tau_{d+1}$ has to be a function of x_1 only, we have

$$\frac{\partial}{\partial x_d} L_g L_f^d h = \varphi_1''(x_1)g_1(x_1). \tag{3.9}$$

If the system is LIS, the left-hand side of (3.9) does depend only upon x_1, and, if $g_1(x_1) \neq 0$, this provides $\varphi_1''(x_1)$, hence $\varphi_1'(x_1)$ up to a constant.

We can check if the drift system is LIS at order $d+1$ for a value of this constant. If not, we do the same with τ_{d+2}:

$$\begin{aligned}\tau_{d+2} &= L_f^{d+1}h - L_f^d(\varphi_1(x_1)) - L_f^{d-1}(\varphi_2(x_1)) - \cdots - \varphi_{d+1}(x_1)\\ &= L_f^{d+1}h - L_f^d(\varphi_1(x_1)) - x_d\varphi_2'(x_1) + \text{ terms without } x_d\end{aligned}$$

and

$$L_g\tau_{d+1} = L_g\left(L_f^{d+1}h - L_f^d(\varphi_1(x_1))\right) - x_d\varphi_2''(x_1)g_1(x_1) + \text{ terms without } x_d.$$

As $L_g\tau_{d+2}$ has to be a function of x_1 only, we have

$$\varphi_2''(x_1)g_1(x_1) = \frac{\partial}{\partial x_d}\left(L_g\left(L_f^{d+1}h - L_f^d(\varphi_1(x_1))\right)\right)$$

and, again, $\varphi_2'(x_1)$ is known up to a constant if $g_1(x_1) \neq 0$ and if the right-hand side depends only on x_1 for a value of the constant of φ_1'.

We know $\varphi_1'(x_1)$ and $\varphi_2'(x_1)$ up to constants, and we can check if the drift system is LIS at order $d+2$ for some values of these constants.

By induction we will have

$$\begin{aligned}\tau_{k+1} &= L_f^k h - L_f^{k-1}(\varphi_1(x_1)) - \cdots - L_f^{d-1}(\varphi_{k-d+1}(x_1)) - \cdots - \varphi_k \circ h \\ &= \zeta(x) - x_d \varphi_{k-d+1}'(x_1) + \text{ terms without } x_d,\end{aligned}$$

where $\zeta = L_f^k h - L_f^{k-1}(\varphi_1(x_1)) - \cdots - L_f^d(\varphi_{k-d+2}(x_1))$ is known up to some constants. If the system is LIS we must have

$$\varphi_{k-d+1}''(x_1) g_1(x_1) = \frac{\partial}{\partial x_d}\zeta.$$

If it is not possible to choose the constants in such a way that $\frac{\partial}{\partial x_d}\zeta$ is a function of x_1 only, then the system is not LIS. If this function depends on x_1 only, one has to check whether the system is LIS at order $k+1$. If the answer is negative, the computation continues at order $k+2$.

Thus the computation of the functions φ_i is quite easy whenever the first component g_1 of the vector field g vanishes on no nontrivial interval.

Example 2. Let us consider the system defined by

$$\left\{\begin{array}{l} \dot{x}_1 = x_2 + u\gamma(x_1), \\ \dot{x}_2 = \sqrt{1-(x_1+x_2-\delta(x_1))^2} + (\delta'(x_1)-1)x_2 \\ \qquad\qquad + u(\delta'(x_1)-1)\gamma(x_1), \\ y = h(x_1,x_2) = x_1, \end{array}\right. \tag{3.10}$$

where γ and δ are smooth functions from $\mathbb{R}$ into $\mathbb{R}$. The system is defined for

$$-1 < x_1 + x_2 - \delta(x_1) < 1$$

and we assume that γ vanishes on no nontrivial interval.

If an immersion $\tau = (\tau_1, \tau_2, \tau_3, \dots)$ into an LIS canonical system exists, then $L_g\tau_3$ is a function of x_1 only. Hence we compute

$$\begin{aligned}\tau_3 &= L_f^2 h - L_f(\varphi_1 \circ h) - \varphi_2 \circ h \\ &= \sqrt{1-(x_1+x_2-\delta(x_1))^2} + (\delta'(x_1)-1)x_2 - x_2\varphi_1'(x_1) - \varphi_2(x_1)\end{aligned}$$

and

$$\begin{aligned}L_g\tau_3 = {} & x_2\delta''(x_1)\gamma(x_1) + (\delta'(x_1)-1)^2\gamma(x_1) - x_2\varphi_1''(x_1)\gamma(x_1) \\ & - \varphi_1'(x_1)\gamma(x_1)(\delta'(x_1)-1) - \varphi_2'(x_1)\gamma(x_1).\end{aligned}$$

As $L_g\tau_3$ does not depend on x_2 we get $\delta''(x_1) - \varphi_1''(x_1) = 0$; hence

$$\varphi_1(x_1) = \delta(x_1) + Cx_1,$$

where C is a constant to be determined.

If the drift part of the system is LIS at order 3, the extension will be

$$\left\{\begin{array}{l} \dot{x}_1 = x_2, \\ \dot{x}_2 = x_3, \\ \dot{x}_3 = L_f^3 h + \varphi_1'\left(x_3 - L_f^2 h\right). \end{array}\right.$$

An easy computation gives

$$\begin{aligned} L_f^3 h + \varphi_1' \left(x_3 - L_f^2 h\right) = & -x_1 - x_2 + \delta(x_1) + \delta''(x_1) x_2^2 + (\delta'(x_1) - 1)^2 x_2 \\ & - (C+1)\sqrt{1 - (x_1 + x_2 - \delta(x_1))^2} \\ & + (\delta'(x_1) + C)((x_3 - \delta'(x_1) x_2 + x_2). \end{aligned}$$

As this expression must be polynomial w.r.t. x_2, we take $C = -1$, we get

$$L_f^3 h + \varphi_1' \left(x_3 - L_f^2 h\right) = -x_1 - x_2 + \delta(x_1) + \delta''(x_1) x_2^2 + (\delta'(x_1) - 1) x_3,$$

and, as $\varphi_1(x_1) = \delta(x_1) - x_1$,

$$L_f^3 h + \varphi_1' \left(x_3 - L_f^2 h\right) - L_F^2(\varphi_1(x_1)) = -x_1 - x_2 + \delta(x_1).$$

Consequently

$$\begin{aligned} \varphi_2(x_1) &= -x_1, \\ \varphi_3(x_1) &= \delta(x_1) - x_1 \end{aligned}$$

and the drift system can be immersed into

$$\left\{ \begin{array}{l} \dot{z_1} = z_2 + \delta(z_1) - z_1, \\ \dot{z_2} = z_3 - z_1, \\ \dot{z_3} = \qquad \delta(z_1) - z_1. \end{array} \right.$$

The immersion is given by

$$\begin{array}{rcl} \tau_1 & = & x_1, \\ \tau_2 & = & L_f h - \varphi_1(x_1), \\ & = & x_1 + x_2 - \delta(x_1), \\ \tau_3 & = & L_f^2 h - L_f(\varphi_1(x_1)) - \varphi_2(x_1), \\ & = & x_1 + \sqrt{1 - (x_1 + x_2 - \delta(x_1))^2}; \end{array}$$

hence

$$\begin{array}{rcl} L_g \tau_1 & = & \gamma(x_1), \\ L_g \tau_2 & = & 0, \\ L_g \tau_3 & = & \gamma(x_1) \end{array}$$

and the system can be immersed into

$$\left\{ \begin{array}{l} \dot{z_1} = z_2 + \delta(z_1) - z_1 + u\gamma(x_1), \\ \dot{z_2} = z_3 - z_1, \\ \dot{z_3} = \qquad \delta(z_1) - z_1 + u\gamma(x_1). \end{array} \right.$$

□

Appendix. In what follows F stands for the vector field

$$F = \sum_{i=1}^{n-1} x_{i+1} \frac{\partial}{\partial x_i},$$

the possible last term $\theta(x_1, x_2, \ldots, x_n) \frac{\partial}{\partial x_n}$ being of no importance.

In order to compute the universal polynomials $P_{n,d}$ of section 3.1.2, the simplest way is to compute $L_F^k\varphi_{n-k}$ for $k = d, \dots, n-1$ and to sum the terms where $x_{d+1}, \dots, x_n$ appear on $k = d, \dots, n-1$. For example, let us compute $P_{4,2}$:

$$\begin{aligned} L_F^2\varphi_2 &= L_F(x_2\varphi_2') \\ &= x_3\varphi_2' + x_2^2\varphi_2'', \\ L_F^3\varphi_1 &= L_F(x_3\varphi_1' + x_2^2\varphi_1'') \\ &= x_4\varphi_1' + 3x_2x_3\varphi_1'' + x_2^3\varphi_1^{(3)}; \end{aligned}$$

hence

$$P_{4,2} = x_4\varphi_1' + x_3(3x_2\varphi_1'' + \varphi_2').$$

Of course this computation can be rather long, and, if d is great but n is not too much greater than d, it is more interesting to obtain $P_{n,d}$ without completely computing the $L_F^k\varphi_{n-k}$'s.

First of all we have the following lemma.

LEMMA A.1. *Let I be an open interval and $\varphi \in \mathcal{C}^\infty(I;\mathbb{R})$. Then, for $1 \le k \le n-1$,*

$$L_F^k(\varphi \circ h) = \sum_{p=1}^{k} Q_p^k \varphi^{(p)}(x_1),$$

where Q_p^k is a polynomial in the variables $x_2, \dots, x_n$, the monomials of which are

$$\zeta_{k,p,r} x_2^{r_2} x_3^{r_3} \dots x_{k+1}^{r_{k+1}},$$

where

$$\begin{aligned} r_2 + r_3 + \dots + r_{k+1} &= p, \\ r_2 + 2r_3 + \dots + kr_{k+1} &= k, \end{aligned}$$

$r = (r_2, r_3, \dots, r_{k+1})$, and $\zeta_{k,p,r}$ is an integer. We set $r = e_j$ if $r_j = 1$ and $r_i = 0$ for $i \ne j$, and we have the following:

- *If $p = 1$, then*

$$\zeta_{k,1,e_{k+1}} = 1$$

  and $\zeta_{k,1,r}$ vanishes for $r \ne e_{k+1}$.
- *If $p > 1$, then $\zeta_{k,p,r}$ can be defined by induction:*

$$\begin{aligned} \zeta_{k,p,(r_2,..,r_{k+1})} = \nu_2 \zeta_{k-1,p-1,(r_2-1,..,r_{k+1})} \\ + \textstyle\sum_{j=3}^{k+1} \nu_j (r_{j-1}+1) \zeta_{k-1,p,(..,r_{j-1}+1,r_j-1,...)}, \end{aligned}$$

  where $\nu_j = 1$ if $r_j > 0$; $\nu_j = 0$ otherwise.

Proof of Lemma A.1. As $L_F(\varphi \circ h) = x_2\varphi'(x_1)$, the lemma is clearly true for $k = 1$.

Let $\zeta x_2^{r_2} x_3^{r_3} \dots x_{k+1}^{r_{k+1}} \varphi^{(p)}$ be a monomial of $L_F^k(\varphi \circ h)$. The differentiation of this monomial along the vector field F gives monomials where $x_j^{r_j} x_{j+1}^{r_{j+1}}$ (with $r_j > 0$) is replaced by $x_j^{r_j-1} x_{j+1}^{r_{j+1}+1}$ and a monomial where $x_2^{r_2}\varphi^{(p)}(x_1)$ is replaced by

$x_2^{r_2+1}\varphi^{(p+1)}(x_1)$. This gives by induction the result about the r_j's and it remains to evaluate the $\zeta_{k,p,r}$'s.

If $p = 1$, the only nonzero monomial is $\zeta x_{k+1}\varphi'(x_1)$ and it is obtained by differentiating $k-1$ times $x_2\varphi'(x_1)$, so $\zeta_{k,1,e_{k+1}} = 1$ and $\zeta_{k,1,r} = 0$ if $r \neq e_{k+1}$.

If $p > 1$, the result is obtained by considering the monomials of $L_F^{k-1}(\varphi \circ h)$ whose differentiation along F gives the desired monomial. □

As a first consequence of Lemma A.1, $P_{n,d}$ is of the first degree w.r.t. $x_{d+1}, \ldots, x_n$ if $n \leq 2d-1$ because $r_j \geq 2$ with $j \geq d+1$ implies

$$n \geq k \geq (j-1)r_j \geq 2d.$$

We can then compute $P_{n,d}$ for $n = d+1, d+2, d+3, \ldots$.

1. $n = d+1$. The computation of $P_{d+1,d}$ involves only $L_f^d\varphi_1$ and the only term of this last quantity where x_{d+1} appears is $x_{d+1}\varphi_1'$. So we obtain the previously used formula:

$$P_{d+1,d} = x_{d+1}\varphi_1'.$$

2. $n = d+2$.

$$\begin{aligned} L_F^{d+1}\varphi_1 &= x_{d+2}\varphi_1' + \alpha x_{d+1}x_2\varphi_1'' + \ldots, \\ L_F^d\varphi_2 &= x_{d+1}\varphi_2' + \ldots, \end{aligned}$$

where $\alpha = \zeta_{d+1,2,(1,0,..,r_{d+1}=1,0)}$. But

$$\begin{aligned} \zeta_{d+1,2,(1,0,..,r_{d+1}=1,0)} &= \zeta_{d,1,(0,0,..,r_{d+1}=1,0)} + \zeta_{d,2,(1,0,..,r_d=1,0,0)} \\ &= 1 + \zeta_{d,2,(1,0,..,r_d=1,0,0)} \end{aligned}$$

and by induction

$$\zeta_{d+1,2,(1,0,..,r_{d+1}=1,0)} = d+1.$$

Hence

$$P_{d+2,d} = x_{d+2}\varphi_1' + x_{d+1}\left((d+1)x_2\varphi_1'' + \varphi_2'\right).$$

3. In the same manner, the reader can check that for $n = d+3$ (with $d \geq 3$),

$$\begin{aligned} P_{d+3,d} = {} & x_{d+3}\varphi_1' + x_{d+2}\left((d+2)x_2\varphi_1'' + \varphi_2'\right) \\ & + x_{d+1}\left(\tfrac{(d+1)(d+2)}{2}x_3\varphi_1'' + \tfrac{(d+1)(d+2)}{2}x_2^2\varphi_1^{(3)} + (d+1)x_2\varphi_2'' + \varphi_3'\right). \end{aligned}$$

REFERENCES

[BZ83] D. Bestle and M. Zeitz, *Canonical form observer design for non-linear time-variable systems,* Internat. J. Control, 38 (1983), pp. 419–431.

[BRG89] D. Bossane, D. Rakotopara, and J.P. Gauthier, *Local and global immersion into linear systems up to output injection*, in Proceedings of the 28th IEEE Conference on Decision and Control, 1989, pp. 2000–2004.

[FK83] M. Fliess and I. Kupka, *A finiteness criterion for nonlinear input-output differential systems*, SIAM J. Control Optim., 21 (1983), pp. 721–728.

[Gua01] M. Guay, *Observer linearization by output diffeomorphism and output-dependent time-scale transformations*, in Nonlinear Control Systems 2001, Pergamon Press, Oxford, 2002, pp. 1443–1446.

[GHO92] J.P. Gauthier, H. Hammouri, and S. Othman, *A simple observer for nonlinear systems. Applications to bioreactors*, IEEE Trans. Automat. Control, 37 (1992), pp. 875–880.

[GMP96] A. Glumineau, C.H. Moog, and F. Plestan, *New algebro-geometric conditions for the linearization by input-output injection*, IEEE Trans. Automat. Control, 41 (1996), pp. 598–603.

[HG88] H. Hammouri and J.P. Gauthier, *Bilinearization up to output injection*, Systems Control Lett., 11 (1988), pp. 139–149.

[HK77] R. Hermann and A.J. Krener, *Nonlinear controllability and observability*, IEEE Trans. Automat. Control, 22 (1977), pp. 728–740.

[HP99] M. Hou and A.C. Pugh, *Observer with linear error dynamics for nonlinear multi-output systems*, Systems Control Lett., 37 (1999), pp. 1–9.

[Kel87] H. Keller, *Non-linear observer design by transformation into a generalized observer canonical form*, Internat. J. Control, 46 (1987), pp. 1915–1930.

[KI83] A.J. Krener and A. Isidori, *Linearization by output injection and nonlinear observers*, Systems Control Lett., 3 (1983), pp. 47–52.

[KR85] A.J. Krener and W. Respondek, *Nonlinear observers with linearizable error dynamics*, SIAM J. Control Optim., 23 (1985), pp. 197–216.

[LPG99] V. López Morales, F. Plestan, and A. Glumineau, *Linearization by completely generalized input-output injection*, Kybernetika (Prague), 35 (1999), pp. 793–802.

[Phe91] A.R. Phelps, *On constructing nonlinear observers*, SIAM J. Control Optim., 29 (1991), pp. 516–534.

[Sus79] H.J. Sussmann, *Single-input observability of continuous-time systems*, Math. Syst. Theory, 12 (1979), pp. 371–393.

[XG89] X.-H. Xia and W.-B. Gao, *Nonlinear observer design by observer error linearization*, SIAM J. Control Optim., 27 (1989), pp. 199–216.

[XZ97] X. Xia and M. Zeitz, *On nonlinear continuous observers,* Internat. J. Control, 66 (1997), pp. 943–954.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 6, pp. 1779–1800

OPTIMAL STRATEGIES FOR RISK-SENSITIVE PORTFOLIO OPTIMIZATION PROBLEMS FOR GENERAL FACTOR MODELS*

HIDEO NAGAI†

Abstract. We consider constructing optimal strategies for risk-sensitive portfolio optimization problems on an infinite time horizon for general factor models, where the mean returns and the volatilities of individual securities or asset categories are explicitly affected by economic factors. The factors are assumed to be general diffusion processes. In studying the ergodic type Bellman equations of the risk-sensitive portfolio optimization problems, we introduce some auxiliary classical stochastic control problems with the same Bellman equations as the original ones. We show that the optimal diffusion processes of the problem are ergodic and that under some condition related to integrability by the invariant measures of the diffusion processes we can construct optimal strategies for the original problems by using the solution of the Bellman equations.

Key words. portfolio optimization, risk-sensitive control, infinite time horizon, Bellman equations, factor models

AMS subject classifications. 91B28, 93E20, 49L20, 35J60, 35K55, 60H30

PII. S0363012901399337

1. Introduction.

Risk-sensitive portfolio optimization problems for factor models have been studied by several authors, e.g., [4], [5], [7], [8], [9], [11], [12], [14], [15], etc., as the study of infinite time horizon versions of Merton terminal wealth problems for incomplete market models. In those works the problems have been mostly formulated for such models that the mean returns of the individual securities depend linearly on underlying economic factors formulated as the solutions of linear stochastic differential equations, except [15], which treats the case of discrete time and nonlinear factors. For these models they considered the problem maximizing the risk-sensitized expected growth rate per unit time:

$$J_\infty(v,x;h) = \liminf_{T\to\infty}\left(-\frac{2}{\theta T}\right)\log E\left[e^{-(\frac{2}{\theta})\log V_T(h)}\right], \tag{1.1}$$

where $V_T(h)$ denotes the capital at time T a investor possesses by selecting a portfolio proportion h. To discuss such portfolio optimization problems, employing the idea of Bellman's dynamic programming principle, they have constructed optimal strategies by using the solutions of relevant ergodic type Bellman equations, or more directly the ones of Riccati equations which express the solutions of the Bellman equations in the case of linear Gaussian factor models. However, it is to be noted that the solutions don't always straightforwardly construct the optimal strategies for the problems. In the case of linear Gaussian factor models, in [5], [9] they actually constructed nearly optimal strategies or optimal ones for small θ, while in [12], [14] the construction was done for general $\theta > 0$ under some condition related to integrability of the criterion function.

*Received by the editors December 7, 2001; accepted for publication (in revised form) August 6, 2002; published electronically February 6, 2003. Research supported in part by Grant-in-Aid for Scientific Research 13440033, JSPS.
http://www.siam.org/journals/sicon/41-6/39933.html
†Department of Mathematical Science, Graduate School of Engineering Science, Osaka University, Toyonaka, 560-8531, Japan (nagai@sigmath.es.osaka-u.ac.jp).

In the present paper we formulate a general factor model where the mean returns as well as the volatilities of security prices depend nonlinearly on the economic factors which are formulated as the solutions of general stochastic differential equations. For such a general model we consider the above-mentioned risk-sensitized portfolio optimization problem on infinite time horizon without strategy constraints. We shall study the ergodic type Bellman equation of the risk-sensitive control problem relevant to the portfolio optimization through asymptotic analysis of the solution of the Bellman equation corresponding to the portfolio optimization on a finite time horizon. Then, by using the solution of the ergodic type equation, we construct the optimal strategy for the problem on infinite time horizon under a similar condition to what was assumed in [12], [14]. We here notice that the condition suggests an integrability by the invariant measure of an underlying ergodic diffusion process. The ergodic diffusion process is the optimal one of the other classical ergodic control problem with the same Bellman equation of ergodic type as the original one. Furthermore, the integrability condition is checked precisely in the case of linear Gaussian models in section 5. We remark that such a situation occurs in discussing other stochastic control problems with an exponential type criterion as well (cf. [11]).

2. Finite time horizon case. We consider a market with $m+1 \geq 2$ securities and $n \geq 1$ factors. We assume that the set of securities includes one bond, whose price is defined by ordinary differential equation:

$$dS^0(t) = r(X_t)S^0(t)dt, \qquad S^0(0) = s^0, \tag{2.1}$$

where $r(x)$ is a nonnegative bounded function. The other security prices S^i_t, $i = 1, 2, \ldots, m$, and factors X_t are assumed to satisfy the following stochastic differential equations:

$$\begin{aligned} &dS^i(t) = S^i(t)\{g^i(X_t)dt + \textstyle\sum_{k=1}^{n+m} \sigma^i_k(X_t)dW^k_t\}, \\ &S^i(0) = s^i, \qquad i = 1, \ldots, m, \end{aligned} \tag{2.2}$$

and

$$\begin{aligned} &dX_t = b(X_t)dt + \lambda(X_t)dW_t, \\ &X_0 = x \in R^n, \end{aligned} \tag{2.3}$$

where $W_t = (W^k_t)_{k=1,\ldots,(n+m)}$ is an $m+n$ dimensional standard Brownian motion process defined on a filtered probability space $(\Omega, \mathcal{F}, P, \mathcal{F}_t)$. Here σ and λ are, respectively, $m \times (m+n)$, $n \times (m+n)$ matrix valued functions. We assume that

$$\begin{aligned} &g,\ \sigma,\ b,\ \lambda \ \text{ are locally Lipschitz}, \\ &c_1|\xi|^2 \leq \xi^*\sigma\sigma^*(x)\xi \leq c_2|\xi|^2, \qquad c_1, c_2 > 0, \\ &x^*b(x) + \tfrac{1}{2} \parallel \lambda\lambda^*(x) \parallel \leq K(1+|x|^2), \end{aligned} \tag{2.4}$$

where σ^* stands for the transposed matrix of σ.

Let us denote by $h^i(t)$ a portion of the capital invested to the ith security $S^i(t)$, $i = 0, 1, \ldots, m$ and set

$$\begin{aligned} &S(t) = (S^1(t), S^2(t), \ldots, S^m(t))^*, \\ &h(t) = (h^1(t), h^2(t), \ldots, h^m(t))^* \end{aligned}$$

and

$$\mathcal{G}_t = \sigma(S(u), X(u); u \le t).$$

Here S^* stands for transposed matrix of S.

DEFINITION 2.1. *$(h^0(t), h(t)^*)_{0\le t\le T}$ is called an investment strategy if the following conditions are satisfied:*

(i) *$h(t)$ is a R^m valued $\mathcal{G}_t$ progressively measurable stochastic process such that*

$$\sum_{i=1}^{m} h^i(t) + h^0(t) = 1; \tag{2.5}$$

(ii)

$$P\left(\int_0^T |h(s)|^2 ds < \infty\right) = 1.$$

The set of all investment strategies will be denoted by $\mathcal{H}(T)$. When $(h^0(t), h(t)^*)_{0\le t\le T} \in \mathcal{H}(T)$, we will often write $h \in \mathcal{H}(T)$ for simplicity since h^0 is determined by (2.5).

For given $h \in \mathcal{H}(T)$ the process $V_t = V_t(h)$ representing the investor's capital at time t is determined by the stochastic differential equation:

$$\begin{aligned} \frac{dV_t}{V_t} &= \textstyle\sum_{i=0}^m h^i(t) \dfrac{dS^i(t)}{S^i(t)} \\ &= h^0(t) r(X_t) dt + \textstyle\sum_{i=1}^m h^i(t)\{g^i(X_t)dt + \sum_{k=1}^{m+n} \sigma_k^i(X_t) dW_t^k\}, \\ V_0 &= v. \end{aligned}$$

Then, taking (2.5) into account, we find that it turns out to be a solution of

$$\begin{aligned} \frac{dV_t}{V_t} &= r(X_t)dt + h(t)^*(g(X_t) - r(X_t)\mathbf{1})dt + h(t)^*\sigma(X_t)dW_t, \\ V_0 &= v, \end{aligned} \tag{2.6}$$

where $\mathbf{1} = (1, 1, \ldots, 1)^*$.

We first consider the following problem. For a given constant $\theta > -2$, $\theta \ne 0$ maximize the following risk-sensitized expected growth rate up to time horizon T:

$$J(v, x; h; T) = -\frac{2}{\theta} \log E[e^{-\frac{\theta}{2} \log V_T(h)}], \tag{2.7}$$

where h ranges over the set $\mathcal{A}(T)$ of all admissible strategies defined later. Then we consider the problem of maximizing the risk-sensitized expected growth rate per unit time,

$$J(v, x; h) = \liminf_{T\to\infty} \left(\frac{-2}{\theta T}\right) \log E[e^{-\frac{\theta}{2} \log V_T(h)}], \tag{2.8}$$

where h ranges over the set of all investment strategies such that $h \in \mathcal{A}(T)$ for each T.

Since V_t satisfies (2.6) we have

$$\begin{aligned} V_t^{-\frac{\theta}{2}} &= v^{-\frac{\theta}{2}} \exp\{\tfrac{\theta}{2} \textstyle\int_0^t \eta(X_s, h_s) ds \\ &\quad - \tfrac{\theta}{2} \textstyle\int_0^t h_s^* \sigma(X_s) dW_s - \tfrac{\theta^2}{8} \int_0^t h_s^* \sigma\sigma^*(X_s) h_s ds\}, \end{aligned}$$

where

$$\eta(x,h)=\left(\frac{\theta+2}{4}\right)h^*\sigma\sigma^*(x)h-r(x)-h^*(g(x)-r(x)\mathbf{1}).$$

If a given investment strategy h satisfies

$$E[e^{-\frac{\theta}{2}\int_0^T h(s)^*\sigma^*(X_s)dW_s-\frac{\theta^2}{8}\int_0^T h(s)^*\sigma\sigma^*(X_s)h(s)ds}]=1, \tag{2.9}$$

then we can introduce a probability measure P^h given by

$$P^h(A)=E[e^{-\frac{\theta}{2}\int_0^T h^*(s)\sigma(X_s)dW_s-\frac{\theta^2}{8}\int_0^T h^*(s)\sigma\sigma^*(X_s)h(s)ds};A]$$

for $A\in\mathcal{F}_T$, $T>0$. By the probability measure P^h our criterion $J(v,x;h;T)$ and $J(v,x;h)$ can be written as follows:

$$J(v,x;h,T)=\log v-\frac{2}{\theta}\log E^h[e^{\frac{\theta}{2}\int_0^T \eta(X_s,h(s))ds}] \tag{2.7$'$}$$

and

$$J(v,x;h)=\liminf_{T\to\infty}-\frac{2}{\theta T}\log E^h[e^{\frac{\theta}{2}\int_0^T \eta(X_s,h(s))ds}]. \tag{2.8$'$}$$

On the other hand, under the probability measure,

$$\begin{aligned}W_t^h&=W_t-\langle W_\cdot,-\tfrac{\theta}{2}\textstyle\int_0^\cdot h^*(s)\sigma(X_s)dW_s\rangle_t\\&=W_t+\tfrac{\theta}{2}\textstyle\int_0^t\sigma^*(X_s)h(s)ds\end{aligned}$$

is a standard Brownian motion process, and therefore the factor process X_t satisfies the following stochastic differential equation:

$$dX_s=\left(b(X_s)-\frac{\theta}{2}\lambda\sigma^*(X_s)h(s)\right)ds+\lambda(X_s)dW_s^h. \tag{2.10}$$

We regard (2.10) as a stochastic differential equation controlled by h and the criterion function is written by P^h as follows:

$$J(v,x;h;T-t)=\log v-\frac{2}{\theta}\log E^h[e^{\frac{\theta}{2}\int_0^{T-t}\eta(X_s,h(s))ds}] \tag{2.11}$$

and the value function

$$u(t,x)=\sup_{h\in\mathcal{H}(T-t)}J(v,x;h;T-t),\qquad 0\le t\le T. \tag{2.12}$$

Then, according to Bellman's dynamic programming principle, it should satisfy the following Bellman equation:

$$\begin{aligned}&\frac{\partial u}{\partial t}+\sup_{h\in R^m}L^hu=0,\\&u(T,x)=\log v,\end{aligned} \tag{2.13}$$

where L^h is defined by

$$L^hu(t,x)=\frac{1}{2}\mathrm{tr}(\lambda\lambda^*(x)D^2u)+\left(b(x)-\frac{\theta}{2}\lambda\sigma^*(x)h\right)^*Du-\frac{\theta}{4}(Du)^*\lambda\lambda^*(x)Du-\eta(x,h).$$

Note that $\sup_{h\in R^m} L^h u$ can be written as

$$\sup_{h\in R^m} L^h u(t,x) = \tfrac{1}{2}\mathrm{tr}(\lambda\lambda^*(x)D^2u) + (b - \tfrac{\theta}{\theta+2}\lambda\sigma^*(\sigma\sigma^*)^{-1}(g - r\mathbf{1}))^* Du$$
$$- \tfrac{\theta}{4}(Du)^*\lambda(I - \tfrac{\theta}{\theta+2}\sigma^*(\sigma\sigma^*)^{-1}\sigma)\lambda^* Du + \tfrac{1}{\theta+2}(g - r\mathbf{1})^*(\sigma\sigma^*)^{-1}(g - r\mathbf{1}).$$

Therefore our Bellman equation (2.13) is written as follows:

$$\begin{aligned} &\tfrac{\partial u}{\partial t} + \tfrac{1}{2}\mathrm{tr}(\lambda\lambda^* D^2u) + B(x)^* Du - (Du)^*\lambda N^{-1}\lambda^* Du + U(x) = 0,\\ &u(T,x) = \log v, \end{aligned} \tag{2.14}$$

where

$$\begin{aligned} B(x) &= b(x) - \tfrac{\theta}{\theta+2}\lambda\sigma^*(\sigma\sigma^*)^{-1}(g(x) - r(x)\mathbf{1}),\\ N^{-1}(x) &= \tfrac{\theta}{4}(I - \tfrac{\theta}{\theta+2}\sigma^*(\sigma\sigma^*)^{-1}\sigma(x)),\\ U(x) &= \tfrac{1}{\theta+2}(g - r\mathbf{1})^*(\sigma\sigma^*)^{-1}(g - r\mathbf{1}) + r(x). \end{aligned} \tag{2.15}$$

As for (2.14), we note that if $\theta > 0$, then

$$\frac{\theta}{2(\theta+2)} I \le N^{-1} \le \frac{\theta}{4} I$$

and therefore we have

$$-\frac{\theta}{4}\lambda\lambda^* \le -\lambda N^{-1}\lambda^* \le -\frac{\theta}{2(\theta+2)}\lambda\lambda^*.$$

Such kinds of equations have been studied in Nagai [13], or Bensoussan, Frehse, and Nagai [3]. Here we can obtain the following result along the line of [3, Theorem 5.1] with refinement on estimate (2.17).

THEOREM 2.1. (i) *If, in addition to* (2.4), $\theta > 0$ *and*

$$\nu_r|\xi|^2 \le \xi^*\lambda\lambda^*(x)\xi \le \mu_r|\xi|^2, \quad r = |x|, \quad \nu_r,\ \mu_r > 0, \tag{2.16}$$

then we have a solution of (2.14) *such that*

$$\begin{aligned} &u,\ \tfrac{\partial u}{\partial t},\ D_k u,\ D_{kj}u \in L^p(0,T;L^p_{loc}(R^n)), \qquad 1 < \forall p < \infty,\\ &\tfrac{\partial^2 u}{\partial t^2},\ \tfrac{\partial D_k u}{\partial t},\ \tfrac{\partial D_{kj}u}{\partial t}, D_{kjl}u \in L^p(0,T;L^p_{loc}(R^n)), \qquad 1 < \forall p < \infty,\\ &u \ge \log v, \quad \tfrac{\partial u}{\partial t} \le 0. \end{aligned}$$

Furthermore, we have the estimate

$$\begin{aligned} |\nabla u|^2(t,x) - \frac{c_0}{\nu_r}\frac{\partial u}{\partial t}(t,x) &\le c_r(|\nabla Q|^2_{2r} + |Q|^2_{2r} + |\nabla(\lambda\lambda^*)|^2_{2r}\\ &\quad + |\nabla B|_{2r} + |B|^2_{2r} + |U|_{2r} + |\nabla U|^2_{2r} + 1), \qquad x \in B_r,\ \ t \in [0,T), \end{aligned} \tag{2.17}$$

where

$$\begin{aligned} &Q = \lambda N^{-1}\lambda^*, \quad c_0 = \tfrac{4(1+c)(\theta+2)}{\theta}, \quad c > 0,\\ &|\cdot|_{2r} = \|\cdot\|_{L^\infty(B_{2r})}, \end{aligned}$$

and c_r is a positive constant depending on n, r, ν_r, μ_r, and c.

(ii) *If, in addition to the above conditions,*

$$\inf_{|x|\geq r} U(x),\ \ \frac{r^2}{\mu_r}\inf_{|x|\geq r} U(x),\ \ r\inf_{|x|\geq r}\frac{U(x)}{|B(x)|}\to\infty,\quad as\ \ r\to\infty,$$

then the above solution u satisfies

$$\inf_{|x|\geq r, t\in(0,T)} u(x,t)\to\infty \qquad as\ \ r\to\infty.$$

Moreover, there exists at most one such solution in $L^\infty(0,T;W^{1,\infty}_{loc}(R^n))$.

Proof. We need only prove (2.17). Let us set

$$F(t,x)=t\left(|\nabla u|^2-\gamma\frac{\partial u}{\partial t}\right),\qquad \gamma>0,$$

and

$$\Gamma(F)=\frac{1}{2}(\lambda\lambda^*)^{ij}D_{ij}F+B^iD_iF-Q^{ij}D_juD_iF+\frac{\partial F}{\partial s}-\frac{F}{s},$$

where $\Lambda=\lambda\lambda^*$. Then, in a similar way to [3], [13] we have

$$\begin{aligned}\Gamma(F)\geq{}&\tfrac{2s}{n\mu_r}(\tfrac{\partial u}{\partial s}+B^iD_iu-\tfrac{1}{2}(Du)^*QDu+U)^2\\&-\tfrac{sn|\nabla\lambda\lambda^*|^2}{2\nu_r}|\nabla u|^2-2s|\nabla B||\nabla u|^2-2s|\nabla Q||\nabla u|^3-2s|\nabla u||\nabla U|.\end{aligned}$$

Note that

$$\frac{\theta\nu_r}{2(\theta+2)}|\nabla u|^2\leq\frac{1}{2}(Du)^*QDu\leq\frac{\theta\mu_r}{2}|\nabla u|^2$$

and that

$$-\frac{1}{2\delta}|B|^2-\frac{\delta}{2}|\nabla u|^2\leq B^*Du\leq\frac{1}{2\delta}|B|^2+\frac{\delta}{2}|\nabla u|^2.$$

By taking δ such that

$$\delta<\frac{\theta\nu_r}{2(\theta+2)},$$

we have

$$-\frac{3\theta\mu_r}{4(\theta+2)}|\nabla u|^2-\frac{1}{\delta}|B|^2\leq B^*Du-\frac{1}{2}(Du)^*QDu-\frac{1}{2\delta}|B|^2\leq-\frac{\theta\nu_r}{4(\theta+2)}|\nabla u|^2.$$

Set

$$|\nabla u|^2=\beta F$$

and take γ such that

$$\gamma=\frac{4(1+c)(\theta+2)}{\theta\nu_r},\qquad c>0;$$

then

$$\frac{1}{\gamma} < \frac{\theta\nu_r}{4(\theta+2)}$$

and we have

$$\frac{\partial u}{\partial s} + B^*Du - \frac{1}{2}(Du)^*QDu - \frac{1}{2\delta}|B|^2 \le -\left[\left(\frac{\theta\nu_r}{4(\theta+2)} - \frac{1}{\gamma}\right)\beta + \frac{1}{s\gamma}\right]F$$

and

$$-\left[\left(\frac{3\theta\mu_r}{4(\theta+2)} - \frac{1}{\gamma}\right)\beta + \frac{1}{s\gamma}\right]F - \frac{1}{\delta}|B|^2 \le \frac{\partial u}{\partial s} + B^*Du - \frac{1}{2}(Du)^*QDu - \frac{1}{2\delta}|B|^2.$$

Thus we obtain

$$\begin{aligned}
\Gamma(F) &\ge \tfrac{2s}{n\mu_r}\left(\tfrac{\partial u}{\partial s} + B^*Du - \tfrac{1}{2}(Du)^*QDu - \tfrac{1}{2\delta}|B|^2\right)^2 \\
&\quad - \tfrac{4s}{n\mu_r}\left(U + \tfrac{1}{2\delta}|B|^2\right)\left|\tfrac{\partial u}{\partial s} + B^*Du - \tfrac{1}{2}(Du)^*QDu - \tfrac{1}{2\delta}|B|^2\right|^2 \\
&\quad - \tfrac{sn|\nabla\lambda\lambda^*|^2}{2\nu_r}|\nabla u|^2 - 2s|\nabla B||\nabla u|^2 - 2s|\nabla Q||\nabla u|^3 - 2s|\nabla u||\nabla U| \\
&\ge \tfrac{2s}{n\mu_r}[(k_r - \tfrac{1}{\gamma})\beta + \tfrac{1}{s\gamma}]^2F^2 - \tfrac{4s}{n\mu_r}\left(U + \tfrac{1}{2\delta}|B|^2\right)[(k_r' - \tfrac{1}{\gamma})\beta + \tfrac{1}{s\gamma}]F \\
&\quad - \tfrac{4s}{n\mu_r\delta}|B|^2\left(U + \tfrac{1}{2\delta}|B|^2\right) - s(\tfrac{n|\nabla\lambda\lambda^*|^2}{2\nu_r} + 2|\nabla B|)\beta F \\
&\quad - 2s|\nabla Q|\beta^{\frac{3}{2}}F^{\frac{3}{2}} - 2s|\nabla U|\beta^{\frac{1}{2}}F^{\frac{1}{2}},
\end{aligned} \tag{2.18}$$

where

$$k_r = \frac{\theta\nu_r}{4(\theta+2)}, \qquad k_r' = \frac{3\theta\nu_r}{4(\theta+2)}.$$

Let $\alpha \in B_r$ and define the function

$$\tau = \begin{cases} (\frac{|x-\alpha|^2}{r^2} - 1)^2, & |x-\alpha| \le r, \\ 0, & |x-\alpha| > r. \end{cases}$$

Then we have

$$\begin{aligned}
\operatorname{tr}(\lambda\lambda^*D^2\tau) &\ge -\frac{4n}{r^2}\mu_r, \\
(D\tau)^*\lambda\lambda^*D\tau &\le \frac{16\mu_r}{r^2}\tau, \\
|D\tau|^2 &\le \frac{16\mu_r}{\nu_r r^2}\tau.
\end{aligned}$$

Now let (s,x) be a maximum point of τF in $[0,t) \times B_r(\alpha)$; then it suffices to prove that

$$\tau F(s,x) \le sc_r(|\nabla Q|^2 + |Q|^2 + |\nabla\lambda\lambda^*|^2 + |\nabla B| + |B|^2 + |U| + |\nabla U|^2 + 1)(x) \tag{2.19}$$

because

$$F(t,\alpha) = (\tau F)(t,\alpha) \le (\tau F)(s,x), \qquad s \le t, \;\; x \in B_{2r}.$$

Note that

$$D(\tau F)(s,x)=0,\quad \frac{\partial F}{\partial s}(s,x)\le 0,\quad \mathrm{tr}(\lambda\lambda^* D^2(\tau F))(s,x)\le 0.$$

Therefore

$$\begin{aligned}-\tfrac{\tau F}{s} &\ge \tfrac{1}{2}(\lambda\lambda^*)^{ij}D_{ij}(\tau F)+B^iD_i(\tau F)-Q^{ij}D_juD_i(\tau F)+\tau\tfrac{\partial F}{\partial s}-\tfrac{\tau F}{s}\\ &=\tau\Gamma(F)+(\tfrac{1}{2}(\lambda\lambda^*)^{ij}D_{ij}\tau)F-(\lambda\lambda^*)^{ij}\tfrac{D_i\tau D_j\tau}{\tau}F+(B^iD_i\tau-Q^{ij}D_juD_i\tau)F.\end{aligned}$$

Thus we obtain

$$-\frac{\tau F}{s}\ge\tau\Gamma(F)-\frac{c_1\mu_r}{r^2}F-(|B|+|Q|\beta^{\frac12}F^{\frac12})\frac{c_2\sqrt{\mu_r}}{r\sqrt{\nu_r}}\sqrt{\tau}F,$$

where c_1 and c_2 are global constants. Therefore

$$\begin{aligned}-\tfrac{\tau F}{s}\ge{}&\tfrac{2s}{n\mu_r}[(k_r-\tfrac{1}{\gamma})\beta+\tfrac{1}{s\gamma}]^2\tau F^2-\tfrac{4s}{n\mu_r}(U+\tfrac{1}{2\delta}|B|^2)[(k_r'-\tfrac{1}{\gamma})\beta+\tfrac{1}{s\gamma}]\tau F\\ &-(2s|\nabla Q|\beta^{\frac32}\tau+|Q|\beta^{\frac12}\tfrac{c_2\sqrt{\mu_r}}{r\sqrt{\nu_r}}\sqrt{\tau})F^{\frac32}\\ &-\{s(\tfrac{n|\nabla\lambda\lambda^*|^2}{2\nu_r}+2|\nabla B|)\beta\tau+\tfrac{c_1\mu_r}{r^2}+|B|\tfrac{c_2\sqrt{\mu_r}}{r\sqrt{\nu_r}}\sqrt{\tau}\}F\\ &-2s|\nabla U|\beta^{\frac12}\tau F^{\frac12}-\tfrac{4s}{n\mu_r\delta}|B|^2(U+\tfrac{1}{2\delta}|B|^2)\tau.\end{aligned}$$

We can asssume that

$$F\ge s|B|^2,\qquad F^{\frac12}\ge s^{\frac12}|\nabla U|. \tag{2.20}$$

Indeed, otherwise (2.19) is already proved. Then

$$\begin{aligned}-\tau\ge{}&\tfrac{2}{n\mu_r}[(k_r-\tfrac{1}{\gamma})\beta s+\tfrac{1}{\gamma}]^2\tau F-\tfrac{4s}{n\mu_r}(U+\tfrac{1}{2\delta}|B|^2)[(k_r'-\tfrac{1}{\gamma})\beta s+\tfrac{1}{\gamma}]\\ &-(2s^2|\nabla Q|\beta^{\frac32}+s|Q|\beta^{\frac12}\tfrac{c_2\sqrt{\mu_r}}{r\sqrt{\nu_r}})(\tau F)^{\frac12}\\ &-\{s^2(\tfrac{n|\nabla\lambda\lambda^*|^2}{2\nu_r}+2|\nabla B|)\beta+s\tfrac{c_1\mu_r}{r^2}+s|B|\tfrac{c_2\sqrt{\mu_r}}{r\sqrt{\nu_r}}\}\\ &-2s^{\frac32}\beta^{\frac12}-\tfrac{4s}{n\mu_r\delta}(U+\tfrac{1}{2\delta}|B|^2).\end{aligned}$$

Setting $X=(\tau F)^{\frac12}$, we have

$$\begin{aligned}0\ge{}&\tfrac{2}{n\mu_r}[(k_r-\tfrac{1}{\gamma})\beta s+\tfrac{1}{\gamma}]^2X^2-s^{\frac12}(2|\nabla Q|(\beta s)^{\frac32}+|Q|(\beta s)^{\frac12}\tfrac{c_2\sqrt{\mu_r}}{r\sqrt{\nu_r}})X\\ &-\tfrac{4s}{n\mu_r}(U+\tfrac{1}{2\delta}|B|^2)[(k_r'-\tfrac{1}{\gamma})\beta s+\tfrac{1}{\gamma}]\\ &-s\{(\tfrac{n|\nabla\lambda\lambda^*|^2}{2\nu_r}+2|\nabla B|)\beta s+\tfrac{c_1\mu_r}{r^2}+|B|\tfrac{c_2\sqrt{\mu_r}}{r\sqrt{\nu_r}}\}\\ &-2s((\beta s)^{\frac12}+\tfrac{2}{n\mu_r\delta}(U+\tfrac{1}{2\delta}|B|^2)).\end{aligned}$$

Since

$$\begin{aligned}&\tfrac{2}{n\mu_r}[(k_r-\tfrac{1}{\gamma})\beta s+\tfrac{1}{\gamma}]^2X^2-s^{\frac12}(2|\nabla Q|(\beta s)^{\frac32}+|Q|(\beta s)^{\frac12}\tfrac{c_2\sqrt{\mu_r}}{r\sqrt{\nu_r}})X\\ &\ge\tfrac{2}{n\mu_r\gamma^2}\{\tfrac12(k_r\gamma-1)\beta s+1\}\{(k_r\gamma-1)\beta s+1\}X^2-\tfrac{s\gamma^2(2\beta s|\nabla Q|+\frac{c_2\sqrt{\mu_r}}{r\sqrt{\nu_r}}|Q|)^2}{4(k_r\gamma-1)\{(k_r\gamma-1)\beta s+1\}}\end{aligned}$$

and $k_r\gamma - 1 = c > 0$, $k_r'\gamma - 1 = 2 + 3c$, we obtain

$$X^2 \le \frac{n\mu_r\gamma^2 s}{(c\beta s+2)(c\beta s+1)} \Big[(U + \tfrac{1}{2\delta}|B|^2)\tfrac{(2+3c)\beta s+1}{\gamma} + \{(\tfrac{n|\nabla\lambda\lambda^*|^2}{2\nu_r} + 2|\nabla B|)\beta s + \tfrac{c_1\mu_r}{r^2} + |B|\tfrac{c_2\sqrt{\mu_r}}{r\sqrt{\nu_r}}\} + 2\{(\beta s)^{\frac{1}{2}} + \tfrac{2}{n\mu_r\delta}(U + \tfrac{1}{2\delta}|B|^2)\} + \frac{\gamma^2(2\beta s|\nabla Q| + \frac{c_2\sqrt{\mu_r}}{r\sqrt{\nu_r}}|Q|)^2}{4c(c\beta s+1)} \Big].$$

Taking into account that

$$\frac{\beta s}{c\beta s+1}, \quad \frac{(\beta s)^{\frac{1}{2}}}{c\beta s+1}, \quad \frac{1}{c\beta s+1}$$

are bounded, we see that

$$X^2 \le sc_r(U + |B|^2 + |\nabla B| + |\nabla\lambda\lambda^*|^2 + |\nabla Q|^2 + |Q|^2 + 1)(x),$$

where c_r is a constant depending on n, r, c, ν_r, and μ_r. Including the other cases where (2.20) does not hold, we conclude (2.19). □

Remark. (i) If

$$\frac{1}{\nu_r},\ \mu_r \le M(1 + r^m), \qquad \exists m \ge 0, \tag{2.21}$$

then we have

$$c_r \le M'(1 + r^{m'}), \qquad \exists m'$$

in estimate (2.17). In particular, if $m = 0$, namely ν_r and μ_r are constants, then c_r can be taken independent of r.

(ii) In [13], [3] we have gotten a similar estimate to (2.17), where the dependence on the coefficients of the equation was not clear. However, to check the condition (2.22) in the following proposition, we need a precise estimate such as (2.17), which makes clear the dependence on the coefficients of (2.14). It is a key estimate to prove that the strategy defined by the solution of (2.14) forms the optimal one as we shall see in the following proposition. It is also the case in the proof of Theorem 4.1 since the estimate holds for the solution w of the limit equation (2.29) of (2.14).

Let us define a class of admissible investment strategy $\mathcal{A}_T$ as the set of investment strategies satisfying (2.9). Then, thanks to the above Theorem 2.1 and the above remark we have the following proposition.

PROPOSITION 2.1. (i) *We assume the assumptions in the above theorem and let u be a solution of* (2.14). *Define*

$$\hat{h}_t = \hat{h}(t, X_t),$$
$$\hat{h}(t,x) = \tfrac{2}{\theta+2}(\sigma\sigma^*)^{-1}(g - r\mathbf{1} - \tfrac{\theta}{2}\sigma\lambda^* Du)(t,x),$$

where X_t is the solution of (2.3); *then, under the assumption that*

$$E[e^{-\int_0^T (2N^{-1}\lambda^* Du+\theta K)^*(x_s)dW_s - \frac{1}{2}\int_0^T (2N^{-1}\lambda^* Du+\theta K)^*(2N^{-1}\lambda^* Du+\theta K)(x_s)ds}] = 1, \tag{2.22}$$

where

$$K = \frac{1}{\theta+2}\sigma^*(\sigma\sigma^*)^{-1}(g - r\mathbf{1}),$$

$\hat{h}_t \in \mathcal{A}_T$ *is an optimal strategy for the portfolio optimization problem of maximizing the criterion* (2.7).

(ii) *If*

$$\begin{aligned} &\nu_r \text{ and } \mu_r \text{ in (2.16) are constants and} \\ &g,\ b,\ \lambda,\ \sigma \text{ are globally Lipschitz,} \end{aligned} \tag{2.23}$$

then (2.22) *is valid.*

Proof. (i) Set

$$Z_T(\hat{h}) = \frac{\theta}{2}\int_0^T \eta(X_s, \hat{h}_s)ds - \frac{\theta}{2}\int_0^T \hat{h}_s^*\sigma(X_s)dW_s - \frac{\theta^2}{8}\int_0^T \hat{h}_s^*\sigma\sigma^*(X_s)\hat{h}_s ds.$$

Since $\hat{h}(t,x)$ attains the supremum in (2.13) we have

$$e^{Z_T(\hat{h})} = e^{\frac{\theta}{2}\{\log v - u(0,x) - \int_0^T [(Du)^*\lambda + \hat{h}_s^*\sigma]dW_s - \frac{\theta}{4}\int_0^T [(Du)^*\lambda + \hat{h}_s^*\sigma][(Du)^*\lambda + \hat{h}_s^*\sigma]^* ds\}}.$$

Note that

$$2N^{-1}\lambda^* Du + \theta K = \frac{\theta}{2}\lambda^* Du + \frac{\theta}{\theta+2}(\sigma\sigma^*)^{-1}\left(g - r\mathbf{1} - \frac{\theta}{2}\sigma\lambda^* Du\right);$$

then, under assumption (2.22), we obtain

$$-\frac{2}{\theta}\log E_x[e^{Z_T(\hat{h})}] = u(0,x) - \log v,$$

namely,

$$\frac{2}{\theta}\log E_x[e^{-\frac{\theta}{2}V_t(\hat{h})}] = -\frac{2}{\theta}\log E_x[v^{-\frac{\theta}{2}}e^{Z_T(\hat{h})}] = u(0,x).$$

On the other hand, for each h_s satisfying (2.9), we have

$$-\frac{2}{\theta}\log E_x[e^{Z_T(h)}] \le u(0,x) - \log v$$

in a similar way to the above because

$$\frac{\partial u}{\partial t} + L^h u \le 0$$

for each $h \in R^m$. Thus, we see that $\hat{h}_t$ is optimal.

(ii) Thanks to (2.17) and the above remark we see that $|Du|$ has at most linear growth under assumption (2.23) and so does $\hat{h}(t,x)^*\sigma(x)$. Moreover, $b(x)$ is globally Lipschitz and we can see that (2.22) holds under these conditions in a similar way to Lemma 4.1.1 in [2]. □

To discuss the problem on infinite time horizon we introduce another stochastic control problem on a finite time horizon with the same Bellman equation as (2.14) and then consider its ergodic counterpart. For that let us set

$$G = b - \lambda\sigma^*(\sigma\sigma^*)^{-1}(g - r\mathbf{1})$$

and rewrite (2.14) as

$$
\begin{aligned}
&\frac{\partial u}{\partial t} + \frac{1}{2}\mathrm{tr}(\lambda\lambda^*(x)D^2u) + G(x)^*Du \\
&\quad - (-\lambda^*Du + NK)^*N^{-1}(-\lambda^*Du + NK)(x) + \frac{\theta+2}{2}K^*NK(x) + r(x) = 0, \\
&u(T,x) = \log v.
\end{aligned} \tag{2.24}
$$

Since

$$
-(-\lambda^*Du + NK)^*N^{-1}(-\lambda^*Du + NK) = \inf_{z\in R^{n+m}}\{z^*Nz + 2z^*NK - 2(\lambda z)^*Du\},
$$

we can regard (2.24) as the Bellman equation of the following stochastic control problem. Set

$$
\begin{aligned}
&u(t,x) \\
&= \inf_{Z_\cdot} E_x\left[\int_0^{T-t}\left\{Z_s^*N(Y_s)Z_s + 2Z_s^*NK(Y_s) + \frac{\theta+2}{2}K^*NK(Y_s) + r(Y_s)\right\}ds + \log v\right],
\end{aligned} \tag{2.25}
$$

where Y_t is a controlled process governed by the stochastic differential equation

$$
dY_t = \lambda(Y_t)dW_t + (G(Y_t) - 2\lambda(Y_t)Z_t)dt, \qquad Y_0 = x, \tag{2.26}
$$

and Z_t is a control taking its value on R^{n+m}. We define the set of admissible controls Z_t as all progressively measurable processes satisfying

$$
E_x\left[\int_0^T |Z_s|^{2q}ds\right] < \infty \qquad \forall q \geq 1.
$$

An ergodic counterpart of the above problem is formulated as follows. Consider the problem

$$
\chi = \inf_{Z_\cdot}\liminf_{T\to\infty}\frac{1}{T}E_x\left[\int_0^T\left\{Z_s^*N(Y_s)Z_s + 2Z_s^*NK(Y_s) + \frac{\theta}{2}K^*NK(Y_s) + r(Y_s)\right\}ds\right] \tag{2.27}
$$

with controlled process Y_t governed by (2.26). Then, the corresponding Bellman equation is written as

$$
\begin{aligned}
\chi &= \frac{1}{2}\mathrm{tr}(\lambda\lambda^*(x)D^2w) + G(x)^*Dw \\
&\quad - (-\lambda^*Dw + NK)^*N^{-1}(-\lambda^*Dw + NK)(x) + \frac{\theta+2}{2}K^*NK(x) + r(x),
\end{aligned} \tag{2.28}
$$

whose original one is

$$
\chi = \frac{1}{2}\mathrm{tr}(\lambda\lambda^*(x)D^2w) + B(x)^*Dw - (Dw)^*\lambda N^{-1}\lambda^*(x)Dw + U(x) = 0, \tag{2.29}
$$

namely,

$$\chi = \frac{1}{2}\mathrm{tr}(\lambda\lambda^*(x)D^2w) + (b - \tfrac{\theta}{\theta+2}\lambda\sigma^*(\sigma\sigma^*)^{-1}(g - r\mathbf{1}))^*Dw$$
$$- \tfrac{\theta}{4}(Dw)^*\lambda(I - \tfrac{\theta}{\theta+2}\sigma^*(\sigma\sigma^*)^{-1}\sigma)\lambda^*Dw + \tfrac{1}{\theta+2}(g - r\mathbf{1})^*(\sigma\sigma^*)^{-1}(g - r\mathbf{1}) + r(x).$$

In the following section we shall analyze the Bellman equation of ergodic type (2.28). Indeed, we shall deduce (2.28), accordingly (2.29), as the limit of parabolic type equation (2.24) as $T \to \infty$ under suitable conditions.

Remark. To regard our Bellman equation as (2.24) has a meaning from a financial point of view. Indeed, under the minimal martingale measure $\tilde{P}$ (cf. [6, Proposition 1.8.2] as for minimal martingale measures), which is defined by

$$\left.\frac{d\tilde{P}}{dP}\right|_{\mathcal{F}_T} = e^{-\int_0^T \zeta(X_s)^* dW_s - \frac{1}{2}\int_0^T |\zeta(X_s)|^2 ds},$$

$\zeta(x) = \sigma^*(\sigma\sigma^*)^{-1}(x)(g(x) - r(x)\mathbf{1})$ factor process X_t is the diffusion process with the generator

$$L = \frac{1}{2}\mathrm{tr}(\lambda\lambda^*(x)D^2) + G(x)^*D,$$

namely, it is governed by the stochastic differential equation

$$dX_t = \lambda(X_t)d\tilde{W}_t + G(X_t)dt.$$

Here $\tilde{W}_t = W_t + \int_0^t \zeta(X_s)ds$ and it is a Brownian motion under the probability measure $\tilde{P}$.

3. Ergodic type Bellman equation. In what follows we assume that

$$\frac{1}{2}\mathrm{tr}(\lambda\lambda^*(x)) + x^*G(x) + \frac{\kappa}{2}\frac{x^*\lambda\lambda^*(x)x}{\sqrt{1+|x|^2}} \le 0, \qquad |x| \ge \exists r_0 > 0, \ \ \kappa > 0, \tag{3.1}$$

and set

$$L = \frac{1}{2}\mathrm{tr}(\lambda\lambda^*(x)D^2) + G^*(x)D.$$

Proposition 3.1. *We assume* (2.4), (3.1), *and* (2.16) *with*

$$\nu_r \ge e^{-\frac{\kappa - c}{8}r}, \qquad c > 0, \ \ r \ge \exists r_1 > 0; \tag{3.2}$$

then L diffusion process $(\tilde{P}_x, X_t)$ *is ergodic, namely, recurrent and admits a finite invariant measure (unique up to a constant multiple). Furthermore, it satisfies*

$$\tilde{E}_x[e^{\kappa\sqrt{1+|X_t|^2}}] \le e^{\kappa\sqrt{1+|x|^2}}. \tag{3.3}$$

Proof. Let us set

$$\alpha(r) = \inf_{|x|=r} \frac{x^*\lambda\lambda^*(x)x}{|x|^2},$$

and

$$I(u) = \int_{r_0}^{\infty} \frac{\beta(u)}{u} du,$$

where

$$\beta(r) = \sup_{|x|=r} \frac{\mathrm{tr}(\lambda\lambda^*(x)) - \frac{x^*\lambda\lambda^*(x)x}{|x|^2} + 2x^*G(x)}{\frac{x^*\lambda\lambda^*(x)x}{|x|^2}}.$$

According to [1], it is known that, if

$$\int_{r_0}^{\infty} e^{-I(u)} du = \infty \tag{3.4}$$

and

$$\int_{r_0}^{\infty} \frac{1}{\alpha(u)} e^{I(u)} du < \infty, \tag{3.5}$$

then the diffusion process with the generator

$$L = \frac{1}{2}\mathrm{tr}(\lambda\lambda^*(x)D^2) + G^*(x)D$$

is ergodic. From (3.1) it follows that

$$\mathrm{tr}(\lambda\lambda^*(x)) - \frac{x^*\lambda\lambda^*(x)x}{|x|^2} + 2x^*G(x) \le -\frac{\kappa}{4}\frac{x^*\lambda\lambda^*(x)x}{\sqrt{1+|x|^2}}, \qquad |x| > r_0.$$

Namely, we have

$$\frac{\mathrm{tr}(\lambda\lambda^*(x)) - \frac{x^*\lambda\lambda^*(x)x}{|x|^2} + 2x^*G(x)}{\frac{x^*\lambda\lambda^*(x)x}{|x|^2}} \le -\frac{\kappa|x|^2}{4\sqrt{1+|x|^2}}$$

and so $\beta(r) \le -\frac{\kappa r^2}{4\sqrt{1+r^2}}$, $r \ge r_0$. Then

$$I(r) \le -\int_{r_0}^{r} \frac{\kappa u}{4\sqrt{1+u^2}} du \le -\frac{\kappa}{8}(r - r_0), \qquad r > r_0 > 0,$$

which implies that

$$\int_{r_0}^{\infty} e^{-I(r)} dr = \infty.$$

On the other hand, because of (2.16), we have $\alpha(r) \ge \nu_r$ and obtain

$$\begin{aligned} \textstyle\int_{r_1}^{\infty} \frac{1}{\alpha(r)} e^{I(u)} du &\le \textstyle\int_{r_1}^{\infty} \frac{1}{\nu_r} e^{-\frac{\kappa}{8}(r-r_1)} dr \\ &= \textstyle\int_{r_1}^{\infty} e^{-\log \nu_r - \frac{\kappa}{8}(r-r_1)} dr < \infty \end{aligned}$$

by using (3.2). Hence we see that the diffusion process $(\tilde{P}_x, X_t)$ with the generator L is ergodic.

To see (3.3), let us set

$$\varphi(x) = e^{\kappa(1+|x|^2)^{\frac{1}{2}}};$$

then by Itô's formula we have

$$\begin{aligned}
&\varphi(X_t) - \varphi(X_0) \\
&\quad = \frac{\kappa\varphi(X_t)}{\sqrt{1+|X_t|^2}}\{\tfrac{1}{2}\mathrm{tr}(\lambda\lambda^*(X_t)) + X_t^* G(X_t) - \frac{X_t^*\lambda\lambda^*(X_t)X_t}{2(1+|X_t|^2)} + \frac{\kappa X_t\lambda\lambda^*(X_t)X_t}{2\sqrt{1+|X_t|^2}}\}dt \\
&\quad + \frac{\kappa\varphi(X_t)}{\sqrt{1+|X_t|^2}} X_t\lambda(X_t)d\tilde{W}_t.
\end{aligned}$$

Thus, we obtain

$$\tilde{E}_x[\varphi(X_{t\wedge\tau_R})] \le \varphi(x), \qquad \tau_R = \inf\{t; |X_t| \ge R\},$$

and so

$$\tilde{E}_x[\varphi(X_t)] \le \varphi(x). \qquad \square$$

THEOREM 3.1. *Assume the assumptions of Theorem* 2.1, (3.1) *and that*

$$\begin{aligned}
&\frac{1}{\nu_r},\ \mu_r \le K(1+r^m), \\
&|Q|,\ |\nabla Q|,\ |B|,\ |\nabla B|,\ U,\ |\nabla U|,\ |\nabla(\lambda\lambda^*)| \le K(1+|x|^m);
\end{aligned}$$

then, as $T \to \infty$,

$$\begin{aligned}
&u(0,x;T) - u(0,0;T) \to w(x), \\
&\tfrac{1}{T}u(0,x;T) \to \chi,
\end{aligned}$$

uniformly on each compact set, where (w,χ) *is the solution of* (2.28) *such that* $w \in C^2(R^n)$.

Proof. In a similar way to Lemma 3.1 in [13], we can see that there exists a subsequence $\{T_i\} \subset R_+$ such that $u(0,x;T_i) - u(0,0;T_i)$ converges to a function $w(x) \in C^2(R^n)$ uniformly on each compact set and strongly in $W^1_{2,loc}$ and $\frac{\partial u}{\partial t}(0,x;T_i)$ to $\chi(x) \in C(R^n)$ uniformly on each compact set by using estimate (2.17). We shall see that $\chi(x) = \chi$, a constant, in what follows.

We need the following lemma.

LEMMA 3.1. *Let* $\tilde{m}$ *be an invariant measure of* L *diffusion process* $(\tilde{P}_x, X_t)$*; then we have the following estimate:*

$$\tilde{m}(|x| \ge R) \le e^{\kappa - \kappa\sqrt{1+R^2}}. \tag{3.6}$$

Proof. Because of (3.3) we have

$$E_0[e^{\kappa\sqrt{1+|X_t|^2}}; |X_t| \ge R] \le e^{\kappa}.$$

Then we obtain

$$\tilde{P}_0(|X_t| \ge R) \le e^{\kappa - \kappa\sqrt{1+R^2}}.$$

Since the left-hand side converges to $\tilde{m}(|x| \ge R)$ as $t \to \infty$, we conclude our present lemma. $\square$

COROLLARY 3.1. *If* $|q(x)| \le c(1+|x|^l)$ *for* $\exists l > 0$, *then* $\int q(x)\tilde{m}(dx) < \infty$.

Now we shall complete the proof of Theorem 3.1. Let us introduce a function γ, which is a solution of the following linear partial differential equation:

$$\begin{aligned}&\tfrac{\partial\gamma}{\partial t} + \tfrac{1}{2}\mathrm{tr}(\lambda\lambda^*(x)D^2\gamma) + G(x)^*D\gamma + \tfrac{\theta+2}{2}K^*NK(x) = 0,\\ &u(T,x) = \log v.\end{aligned}$$

Then we have

$$\begin{aligned}&\tfrac{\partial(\gamma-u)}{\partial t} + \tfrac{1}{2}\mathrm{tr}(\lambda\lambda^*(x)D^2(\gamma-u))\\ &\quad + G(x)^*D(\gamma-u) + (-\lambda^*Du + NK)^*N^{-1}(-\lambda^*Du + NK) = 0,\\ &(\gamma-u)(T,x) = 0\end{aligned}$$

and so we see that

$$\gamma(t,x) \ge u(t,x) \ge \log v. \tag{3.7}$$

On the other hand, γ has an expression by L diffusion process $(\tilde{P}_x, X_t)$ such as

$$\gamma(t,x) = \tilde{E}_x\left[\int_0^{T-t} \frac{\theta+2}{2}K^*NK(X_s)ds\right] + \log v.$$

Therefore, by ergodic theorem

$$\begin{aligned}\lim_{T\to\infty}\tfrac{\gamma(0,x;T)}{T} &= \lim_{T\to\infty}\tfrac{1}{T}\tilde{E}_x[\textstyle\int_0^T \frac{\theta+2}{2}K^*NK(X_s)ds]\\ &= \textstyle\int \frac{\theta+2}{2}K^*NK(x)\tilde{m}(dx).\end{aligned} \tag{3.8}$$

Note that [10, IV, Theorem 5.1], and therefore its Corollary 1, extend to the present case (cf. also Proof of Theorem 3.5 in [1]) by using the above Corollary 3.1. Set

$$\bar{u}(T) = \frac{1}{|B_1|}\int_{B_1} u(0,y;T)dy.$$

Since

$$|u(0,x;T) - u(0,y;T)| \le K_R, \qquad x,y \in B_R,$$

because of (2.17), we have

$$|u(0,x;T) - \bar{u}(T)| \le K_R, \qquad x,y \in B_R, \tag{3.9}$$

for $R > 1$. Noting that $\{\frac{\bar{u}(T)}{T}\}_T$ is bounded due to (3.7) and (3.8), take a subsequence $T_i \subset R_+$ such that

$$\lim_{T_i\to\infty}\frac{\bar{u}(T_i)}{T_i} = \chi.$$

Then (3.9) implies that

$$\lim_{T_i\to\infty}\sup_{x\in B_R}\left|\frac{u(0,x;T_i)}{T_i} - \chi\right| = 0 \qquad \forall R.$$

Hence

$$\chi(x) = \lim_{T_i} \frac{\partial u}{\partial t}(0, x, T_i) = \lim_{T_i \to \infty} \frac{u(0, x; T_i)}{T_i} = \chi$$

uniformly on each compact set. On the other hand, in a similar way to Lemma 3.3 in [13], we can see that $w(x)$ is bounded below and so $w(x) \to \infty$ as $|x| \to \infty$ (cf. Remark 3.2 in [13]). Furthermore, such a solution (w, χ) is unique up to additive constant with respect to a function w (cf. Lemma 3.2 in [13]) and we conclude our present theorem. □

Our Bellman equation of ergodic type (2.28) is rewritten as

$$\begin{aligned} \chi = {} & \tfrac{1}{2}\mathrm{tr}(\lambda\lambda^*(x)D^2 w) + G(x)^* Dw \\ & + \inf_{z \in R^{n+m}} \{z^* N z + 2z^* N K - 2(\lambda z)^* Dw\} + \tfrac{\theta+2}{2} K^* N K(x), \end{aligned} \tag{3.10}$$

and the infimum is attained by

$$\hat{z}(x) = N^{-1}\lambda^*(x) Dw(x) - K(x),$$

which define the following elliptic operator considered as the generator of the optimal diffusion for (2.27):

$$\hat{L} = \frac{1}{2}\mathrm{tr}(\lambda\lambda^*(x)D^2) + G^*(x)D - 2(\lambda N^{-1}\lambda^*(x) Dw(x) - \lambda K(x))^* D.$$

Then we have the following proposition.

PROPOSITION 3.2. *Under the assumptions of Theorem* 3.1, $\hat{L}$ *diffusion process is ergodic.*

Proof. Thanks to Theorem 3.1 and our assumptions, the coefficients of $\hat{L}$ are locally Lipschitz and $\lambda\lambda^*$ is uniformly positive definite on each compact set. Then, we can check Has'minskii's conditions [10] for ergodicity as follows (cf. [10, III Theorem 7.1], its Corollary 1 and Theorem 4.1). Since $w(x) \to \infty$ as $|x| \to \infty$, it is bounded below. Moreover, by calculation we see that

$$\begin{aligned} \hat{L}w &= -\{(Dw)^*\lambda N^{-1}\lambda^* Dw + \tfrac{\theta}{2} K^* N K\} + \chi \\ &= -\tfrac{\theta}{4}(Dw)^*\lambda(I - \tfrac{\theta}{\theta+2}\sigma^*(\sigma\sigma^*)^{-1}\sigma)\lambda^* Dw - \tfrac{1}{\theta+2}(g - r\mathbf{1})^*(\sigma\sigma^*)^{-1}(g - r\mathbf{1}) + \chi. \end{aligned}$$

Since

$$\frac{1}{\theta+2}(g - r\mathbf{1})^*(\sigma\sigma^*)^{-1}(g - r\mathbf{1}) \to \infty, \qquad |x| \to \infty,$$

we see that

$$\hat{L}w \le -C, \ \ |x| >> 1, \quad C > 0.$$

Thus we conclude our present proposition. □

4. Optimal strategy for portfolio optimization on infinite time horizon. Define the set of admissible strategies $\mathcal{A}$ by

$$\mathcal{A} = \{h : h \in \mathcal{A}(T) \ \forall T\}$$

and set

$$\hat{H}_t = \hat{H}(X_t),$$
$$\hat{H}(x) = \tfrac{2}{\theta+2}(\sigma\sigma^*)^{-1}(g - r\mathbf{1} - \tfrac{\theta}{2}\sigma\lambda^* Dw)(x),$$

where X_t is the solution of stochastic differential equation (2.3); then we have the following theorem.

THEOREM 4.1. *In addition to the assumptions of Theorem* 3.1, *we assume* (2.23) *and that*

$$\frac{4}{\theta^2}(g - r\mathbf{1})^*(\sigma\sigma)^{-1}(g - r\mathbf{1}) - (Dw)^*\lambda\sigma^*(\sigma\sigma^*)^{-1}\sigma\lambda^* Dw \to \infty, \qquad |x| \to \infty; \tag{4.1}$$

then $\hat{H}_t$ *is an optimal strategy for portfolio optimization maximizing long run criterion* (2.8):

$$J(v, x; \hat{H}) = \sup_{h \in \mathcal{A}} J(v, x; h).$$

Proof. The Bellman equation (2.28) has the original form

$$\chi = \frac{1}{2}\mathrm{tr}(\lambda\lambda^*(x)D^2 w) + B(x)^* Dw - \frac{\theta}{4}(Dw)^*\lambda\lambda^*(x)Dw + \sup_{h \in R^m}\left\{-\frac{\theta}{2}(\lambda\sigma^* h)^* Dw - \eta(x, h)\right\}$$

and the supremum in this equation is attained by

$$\hat{H}(x) = \frac{2}{\theta+2}(\sigma\sigma^*)^{-1}\left(g - r\mathbf{1} - \frac{\theta}{2}\sigma\lambda^* Dw\right)(x).$$

We consider the stochastic differential equation

$$dX_t = \left(b(X_t) - \frac{\theta}{2}\lambda\sigma^*\hat{H}(X_t)\right)dt + \lambda(X_t)dW_t^{\hat{H}},$$

where

$$W_t^{\hat{H}} = W_t + \frac{\theta}{2}\int_0^t \sigma^*(X_s)\hat{H}(X_s)ds.$$

Then

$$\begin{aligned} w(X_t) - w(X_0) &= \int_0^t \{\tfrac{1}{2}\mathrm{tr}(\lambda\lambda^* D^2 w)(X_s) + (b^* - \tfrac{\theta}{2}\hat{H}^*\sigma\lambda^*)Dw(X_s)\}ds \\ &\quad + \int_0^t (Dw)^*\lambda(X_s)dW_s^{\hat{H}} \\ &= \int_0^t \{\chi + \tfrac{\theta}{4}(Dw)^*\lambda\lambda^* Dw(X_s) + \eta(X_s, \hat{H}_s)\}ds + \int_0^t (Dw)^*\lambda(X_s)dW_s^{\hat{H}}. \end{aligned} \tag{4.2}$$

Therefore, we have

$$\begin{aligned} &E_x^{\hat{H}}\left[e^{\frac{\theta}{2}\int_0^T \eta(X_s, \hat{H}_s)ds}\right] \\ &= E_x^{\hat{H}}\left[e^{-\frac{\theta}{2}\{\chi T + \int_0^T (Dw)^*\lambda(X_s)dW_s^{\hat{H}} + \frac{\theta}{4}\int_0^T (Dw)^*\lambda\lambda^* Dw(X_s)ds - w(X_T) + w(x)\}}\right]. \end{aligned}$$

We define new probability measure $\hat{P}$ by

$$\left.\frac{d\hat{P}}{dP^{\hat{H}}}\right|_{\mathcal{F}_t} = e^{-\frac{\theta}{2}\int_0^T (Dw)^*\lambda(X_s)dW_s^{\hat{H}} - \frac{\theta^2}{8}\int_0^T (Dw)^*\lambda\lambda^* Dw(X_s)ds}; \tag{4.3}$$

then $\hat{W}_t$ defined by

$$\hat{W}_t = W_s^{\hat{H}} + \frac{\theta}{2}\int_0^t \lambda^* Dw(X_s)ds$$

is a Brownian motion under the probability measure $\hat{P}$ and the above stochastic differential equation is described, by using $\hat{W}_t$, as

$$dX_t = \left(b(X_t) - \frac{\theta}{2}\lambda\sigma^*\hat{H}(X_t) - \frac{\theta}{2}\lambda\lambda^* Dw(X_t)\right)dt + \lambda(X_t)d\hat{W}_t.$$

Note that we can see that $|\nabla w|$ is at most linear growth because of the estimate (2.17) under the assumption (2.23) and the right-hand side of (4.3) is a martingale in a similar way to the proof of Proposition 2.1 (ii) (cf. the remark after Theorem 2.1). Thus,

$$E_x^{\hat{H}}[e^{\frac{\theta}{2}\int_0^T \eta(X_s,\hat{H}_s)ds}] = e^{-\frac{\theta}{2}\chi T - \frac{\theta}{2}w(x)}\hat{E}[e^{\frac{\theta}{2}w(X_T)}].$$

Then

$$\begin{aligned} e^{\frac{\theta}{2}w(X_T)} - e^{\frac{\theta}{2}w(X_0)} &= \tfrac{\theta}{2}\int_0^T e^{\frac{\theta}{2}w(X_s)}(Dw)^*\lambda(X_s)d\hat{W}_s \\ &\quad + \int_0^T e^{\frac{\theta}{2}w(X_s)}\{\tfrac{\theta}{4}\mathrm{tr}(\lambda\lambda^* D^2 w) + \tfrac{\theta^2}{8}(Dw)^*\lambda\lambda^* Dw \\ &\quad + \tfrac{\theta}{2}(B - \tfrac{\theta}{2}\lambda\sigma^*\hat{H} - \tfrac{\theta}{2}\lambda\lambda^* Dw)^* Dw\}(X_s)ds. \end{aligned}$$

Note that

$$\frac{\theta}{4}\mathrm{tr}(\lambda\lambda^* D^2 w) - \frac{\theta^2}{8}(Dw)^*\lambda\lambda^* Dw + \frac{\theta}{2}\left(B - \frac{\theta}{2}\lambda\sigma^*\hat{H}\right)^* Dw = \frac{\theta}{2}(\eta(x,\hat{H}(x)) + \chi)$$

and

$$\begin{aligned} \eta(x,\hat{H}(x)) &= \tfrac{\theta+2}{4}\hat{H}^*\sigma\sigma^*\hat{H}(x) - \hat{H}^*(g - r\mathbf{1})(x) - r(x) \\ &= -\tfrac{1}{\theta+2}(g - r\mathbf{1})^*(\sigma\sigma^*)^{-1}(g - r\mathbf{1})(x) \\ &\quad + \tfrac{\theta^2}{4(\theta+2)}(Dw)^*\lambda\sigma^*(\sigma\sigma^*)^{-1}\sigma\lambda^* Dw(x) - r(x). \end{aligned}$$

Then, under assumption (4.1), we see that

$$\eta(x,\hat{H}(x)) + \chi \le 0, \qquad x \in B_R^c, \ \ \exists R > 0,$$

which implies, by the arguments using a stopping time,

$$\hat{E}[e^{\frac{\theta}{2}w(X_T)}] \le e^{\frac{\theta}{2}w(x)} + MT, \qquad \exists M > 0.$$

Hence we conclude that

$$\lim_{T\to\infty} -\frac{2}{\theta T}\log E_x^{\hat{H}}[e^{\frac{\theta}{2}\int_0^T \eta(X_s,\hat{H}_s)ds}] = \lim_{T\to\infty} -\frac{2}{\theta T}\log\{e^{-\frac{\theta}{2}\chi T - \frac{\theta}{2}w(x)}\hat{E}[e^{\frac{\theta}{2}w(X_T)}]\} = \chi.$$

For $h \in \mathcal{A}$, inequality $\leq$ holds in (4.2) and we can see that

$$J(v, x; h) \leq \chi$$

in a similar way to the above since w is bounded below. □

Remark. Under the probability measure $\hat{P}_x$, the factor process is an ergodic diffusion process with the generator $\hat{L}$. In fact, by calculation, we can see that

$$\begin{aligned}&\tfrac{1}{2}\mathrm{tr}(\lambda\lambda^* D^2) + (b - \tfrac{\theta}{2}\lambda\sigma^*\hat{H} - \tfrac{\theta}{2}\lambda\lambda^* Dw)^* D \\ &\quad = \tfrac{1}{2}\mathrm{tr}(\lambda\lambda^*(x)D^2) + G^*(x)D - 2(\lambda N^{-1}\lambda^*(x)Dw(x) - \lambda K(x))^* D.\end{aligned}$$

Then, under assumption (4.1), $\hat{L}$ diffusion process $(\hat{P}_x, X_t)$ satisfies

$$\hat{E}_x[e^{\frac{\theta}{2}w(X_T)}] \to \int e^{\frac{\theta}{2}w(x)}\mu(dx) < \infty \qquad \text{as } T \to \infty,$$

where μ is the invariant measure of (P_x, X_t).

5. Example.

Example (linear Gaussian case). Let us consider the case where

$$\begin{aligned} g(x) = a + Ax,\ \sigma(x) = \Sigma, \\ b(x) = b + Bx,\ \lambda(x) = \Lambda, \\ r(x) = r, \end{aligned}$$

where A, B, Σ, Λ are all constant matrices and a and b are constant vectors. Such a case has been considered by Bielecki and Pliska [4], [5], Fleming and Sheu [8], [9], and Kuroda and Nagai [12].

In this case the solution $u(t, x)$ of (2.14) has the following explicit form:

$$u(t, x) = \frac{1}{2}x^* P(t)x + q(t)^* x + k(t),$$

where $P(t)$ is a solution of the Riccati differential equation

$$\begin{aligned} &\dot{P}(t) - P(t)K_0 P(t) + K_1^* P(t) + P(t)K_1 + \tfrac{2}{\theta+2}A^*(\Sigma\Sigma^*)^{-1}A = 0, \\ &P(T) = 0, \end{aligned} \tag{5.1}$$

and

$$\begin{aligned} K_0 &= \tfrac{\theta}{2}\Lambda(I - \tfrac{\theta}{\theta+2}\Sigma^*(\Sigma\Sigma^*)^{-1}\Sigma)\Lambda^*, \\ K_1 &= B - \tfrac{\theta}{\theta+2}\Lambda\Sigma^*(\Sigma\Sigma^*)^{-1}A. \end{aligned}$$

The term $q(t)$ is a solution of linear differential equation

$$\begin{aligned} &\dot{q}(t) + (K_1^* - P(t)K_0)q(t) + P(t)b + (\tfrac{2}{\theta+2}A^* - \tfrac{\theta}{\theta+2}P(t)\Lambda\Sigma^*)(\Sigma\Sigma^*)^{-1}(a - r\mathbf{1}) = 0, \\ &q(T) = 0 \end{aligned}$$

and $k(t)$ a solution of

$$\begin{aligned} &\dot{k}(t) + \frac{1}{2}\mathrm{tr}(\Lambda\Lambda^* P(t)) - \frac{\theta}{4}q(t)^*\Lambda\Lambda^* q(t) + b^* q(t) + r + \frac{1}{\theta+2}(a - r\mathbf{1})^*(\Sigma\Sigma^*)^{-1}(a - r\mathbf{1}) \\ &\quad + \frac{\theta^2}{4(\theta+2)}q(t)^*\Lambda\Sigma^*(\Sigma\Sigma^*)^{-1}\Sigma\Lambda^* q(t) - \frac{\theta}{\theta+2}(a - r\mathbf{1})^*(\Sigma\Sigma^*)^{-1}\Sigma\Lambda^* q(t) = 0, \\ &k(T) = \log v. \end{aligned}$$

If

$$G \equiv B - \Lambda\Sigma^*(\Sigma\Sigma^*)^{-1}A \quad \text{is stable,}$$

then

(i) $P(0) = P(0;T)$ converges, as $T \to \infty$, to a nonnegative definite matrix $\tilde{P}$, which is a solution of algebraic Riccati equation

$$K_1^*\tilde{P} + \tilde{P}K_1 - \tilde{P}K_0\tilde{P} + \frac{2}{\theta+2}A^*(\Sigma\Sigma^*)^{-1}A = 0.$$

Moreover, $\tilde{P}$ satisfies the estimate

$$0 \le \tilde{P} \le \frac{2}{\theta}\int_0^\infty e^{sG^*}A^*(\Sigma\Sigma^*)^{-1}Ae^{sG}ds. \tag{5.2}$$

(ii) $q(0) = q(0;T)$ converges, as $T \to \infty$, to a constant vector $\tilde{q}$, which satisfies

$$(K_1^* - \tilde{P}K_0)\tilde{q} + \tilde{P}b + \left(\frac{2}{\theta+2}A^* - \frac{\theta}{\theta+2}\tilde{P}\Lambda\Sigma^*\right)(\Sigma\Sigma^*)^{-1}(a - r\mathbf{1}) = 0.$$

(iii) $\frac{k(0;T)}{T}$ converges to a constant $\rho(\theta)$ defined by

$$\begin{aligned}\rho(\theta) = {} & \frac{1}{2}\mathrm{tr}(\tilde{P}\Lambda\Lambda^*) - \frac{\theta}{4}\tilde{q}^*\Lambda\Lambda^*\tilde{q} + b^*\tilde{q} + r + \frac{1}{\theta+2}(a - r\mathbf{1})^*(\Sigma\Sigma^*)^{-1}(a - r\mathbf{1}) \\ & + \tfrac{\theta^2}{4\theta+8}\tilde{q}^*\Lambda\Sigma^*(\Sigma\Sigma^*)^{-1}\Sigma\Lambda^*\tilde{q} - \tfrac{\theta}{\theta+2}(a - r\mathbf{1})^*(\Sigma\Sigma^*)^{-1}\Sigma\Lambda^*\tilde{q}.\end{aligned}$$

If, moreover,

$$(B^*, A^*(\Sigma\Sigma^*)^{-1}\Sigma) \text{ is controllable,} \tag{5.3}$$

then

(iv) the solution $\tilde{P}$ of the above algebraic Riccati equation is strictly positive definite.

Finally, if, in addition to the above conditions,

$$(B, \Lambda) \text{ is controllable,} \tag{5.4}$$

then

(v) the investment strategy $\tilde{h}_t$ defined by

$$\tilde{h}_t = \frac{2}{\theta+2}(\Sigma\Sigma^*)^{-1}\left[a - r\mathbf{1} - \frac{\theta}{2}\Sigma\Lambda^*\tilde{q} + \left(A - \frac{\theta}{2}\Sigma\Lambda^*\tilde{P}\right)X_t\right]$$

is optimal for the portfolio optimization on infinite time horizon maximizing the criterion (2.8)

$$\sup_{h\in\mathcal{A}} J(v,x;h) = J(v,x;\tilde{h}_\cdot) = \rho(\theta)$$

if and only if

$$\hat{P}\Lambda\Sigma^*(\Sigma\Sigma^*)^{-1}\Sigma\Lambda^*\hat{P} < A^*(\Sigma\Sigma^*)^{-1}A, \tag{5.5}$$

where $\hat{P} = \frac{\theta}{2}\tilde{P}$ (cf. [12]).

Set

$$w(x) = \frac{1}{2}x^*\tilde{P}x + \tilde{q}^*x;$$

then $w(x)$ satisfies (2.28) and (5.5) is equivalent to

$$\int e^{\frac{\theta}{2}w(x)}\mu(dx) < \infty$$

under the assumptions (5.3) and (5.4), where $\mu(dx)$ is the invariant measure of $\hat{L}$ diffusion process. We consider the case where $n = m = 1$. Then $\Sigma\Sigma^*$, $\Lambda\Sigma^*$, A, B are all scalars and (5.5) is written as

$$\frac{\theta^2}{4}\tilde{P}^2(\Lambda\Sigma^*)^2 < A^2. \tag{5.5$'$}$$

We can find sufficient condition for (5.5′) by using estimate (5.2). Indeed, If

$$A^2(\Lambda\Sigma^*)^2(\Sigma\Sigma^*)^{-2}\left(\int_0^\infty e^{2sG}ds\right)^2 < 1, \tag{5.6}$$

then (5.5′) holds. (5.6) is equivalent to

$$(2B(\Sigma\Sigma^*) - 3(\Lambda\Sigma^*)A)(2B(\Sigma\Sigma^*) - (\Lambda\Sigma^*)A) > 0,$$

which indicates that

$$B < \frac{1}{2}\Lambda\Sigma^*(\Sigma\Sigma^*)^{-1}A \quad \text{if} \quad \Lambda\Sigma^*A > 0, \tag{5.7}$$

$$B < \frac{3}{2}\Lambda\Sigma^*(\Sigma\Sigma^*)^{-1}A \quad \text{if} \quad \Lambda\Sigma^*A < 0 \tag{5.8}$$

since $G = B - \Lambda\Sigma^*(\Sigma\Sigma^*)^{-1}A < 0$ by the stability assumption.

We illustrate an example where (5.5′) is violated as follows. Set $\theta = 4$ and $B = \frac{2}{3}\Lambda\Sigma^*(\Sigma\Sigma^*)^{-1}A$; then we have

$$\tilde{P}^2(6\Lambda\Lambda^*\Sigma\Sigma^* - 4(\Lambda\Sigma^*)^2) = A^2$$

and therefore (5.5′) is violated if and only if

$$6\Lambda\Lambda^*\Sigma\Sigma^* - 4(\Lambda\Sigma^*)^2 \le 4(\Lambda\Sigma^*)^2,$$

namely,

$$4(\Lambda\Sigma^*)^2 \ge 3\Lambda\Lambda\Sigma\Sigma^*. \tag{5.9}$$

Set $\Lambda = (1, \lambda)$, $\Sigma = (1, \sigma)$; then (5.9) is equivalent to

$$\{\lambda\sigma + 1 + \sqrt{3}(\lambda - \sigma)\}\{\lambda\sigma + 1 - \sqrt{3}(\lambda - \sigma)\} \ge 0.$$

REFERENCES

[1] R.N. Bhattacharya, *Criteria for recurrence and existence of invariant measures for multidimensional diffusions*, Ann. Probab., 6 (1978), pp. 541–553.

[2] A. Bensoussan, *Stochastic Control of Partially Observable Systems*, Cambridge University Press, Cambridge, UK, 1992.
[3] A. Bensoussan, J. Frehse, and H. Nagai, *Some results on risk-sensitive with full observation*, Appl. Math. Optim., 37 (1998), pp. 1–41.
[4] T.R. Bielecki and S.R. Pliska, *Risk-sensitive dynamic asset management*, Appl. Math. Optim., 39 (1999), pp. 337–360.
[5] T.R. Bielecki and S.R. Pliska, *Risk-Sensitive Intertemporal CAPM, with Application to Fixed Income Management*, preprint.
[6] N. El Karoui and M.-C. Quenez, *Dynamic programming pricing of contingent claims in an incomplete market*, SIAM J. Control Optim., 33 (1995), pp. 29–66.
[7] W.H. Fleming and S.J. Sheu, *Optimal long term growth rate of expected utility of wealth*, Ann. Appl. Probab., 9 (1999), pp. 871–903.
[8] W.H. Fleming and S.J. Sheu, *Risk-sensitive control and an optimal investment model*, Math. Finance, 10 (2000), pp. 197–213.
[9] W.H. Fleming and S.J. Sheu, *Risk-Sensitive Control and an Optimal Investment Model* (II), Ann. Appl. Probab., 12 (2002), pp. 730–767.
[10] R.Z. Has'minskii, *Stochastic Stability of Differential Equations*, Sijthoff and Noordhoff, Alphen aan den Rijn, The Netherlands, 1980.
[11] K. Kuroda and H. Nagai, *Ergodic type Bellman equation of risk-sensitive control and portfolio optimization on infinite time horizon*, Optimal Control and Partial Differential Equations - Innovations & Applications, A. Sulem, J.L. Menaldi, and E. Rofman, eds., IOS Press, Amsterdam, 2000, pp. 530–538.
[12] K. Kuroda and H. Nagai, *Risk-sensitive portfolio optimization on infinite time horizon*, Stoch. Stoch. Rep., 73 (2002), pp. 309–331.
[13] H. Nagai, *Bellman equations of risk-sensitive control*, SIAM J. Control Optim., 34 (1996), pp. 74–101.
[14] H. Nagai and S. Peng, *Risk-sensitive dynamic portfolio optimization with partial information on infinite time horizon*, Ann. Appl. Probab., 12 (2002), pp. 173–195.
[15] L. Stettner, *Risk-sensitive portfolio optimization*, Math. Methods Oper. Res., 50 (1999), pp. 463–474.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 6, pp. 1801–1819

APPROXIMATE NONLINEAR FILTERING FOR A TWO-DIMENSIONAL DIFFUSION WITH ONE-DIMENSIONAL OBSERVATIONS IN A LOW NOISE CHANNEL*

PAULA MILHEIRO DE OLIVEIRA† AND JEAN PICARD‡

Abstract. The asymptotic behavior of a nonlinear continuous time filtering problem is studied when the variance of the observation noise tends to 0. We suppose that the signal is a two-dimensional process from which only one of the components is noisy and that a one-dimensional function of this signal, depending only on the unnoisy component, is observed in a low noise channel. An approximate filter is considered in order to solve this problem. Under some detectability assumptions, we prove that the filtering error converges to 0, and an upper bound for the convergence rate is given. The efficiency of the approximate filter is compared with the efficiency of the optimal filter, and the order of magnitude of the error between the two filters, as the observation noise vanishes, is obtained.

Key words. stochastic differential models, nonlinear filtering, approximate filters

AMS subject classifications. 93E11, 60G35, 60F99

PII. S0363012902363920

1. Introduction.

Due to its vast application in engineering, the problem of filtering a random signal X_t from noisy observations of a function $h(X_t)$ of this signal has been considered by several authors. In particular, the case of small observation noise has been widely studied, and several articles are devoted to the research of approximate filters which are asymptotically efficient when the observation noise vanishes. Among them, one notices a first group in which a one-dimensional system is observed through an injective observation function h (see [4, 5, 7, 1]); in this case, the filtering error is small when the observation noise is small, and one can find efficient suboptimal finite-dimensional filters. The multidimensional case appears later with [8, 9], but an assumption of injectivity of h is again required; in particular, the extended Kalman filter is studied in [9]. See also previous work by Krener [6] for systems with linear observations. When h is not injective, the process $\{X_t\}$ cannot always be restored from the observation of $\{h(X_t)\}$, so the filtering error is not always small; such a case is studied in [3]. However, there are some classes of problems in which $\{X_t\}$ can be restored from $\{h(X_t)\}$; in these cases, the filtering error is small, and one again looks for efficient suboptimal filters. For instance, $\{X_t\}$ is sometimes obtained from $\{h(X_t)\}$ and its quadratic variation; see [2, 10, 11, 13]. Here, we are interested in another case in which $h(X_t)$ is differentiable with respect to the time t, and $\{X_t\}$ is obtained from $\{h(X_t)\}$ and its derivative. As opposed to [9], the existence of a Lipschitz inverse of h is not assumed in this paper, as the dimension of the measurements that we consider is lower than that of the state. More precisely, we consider the framework of [12], which we now describe.

We consider the two-dimensional process $X_t = (x_t^{(1)}, x_t^{(2)})$ given by the Itô equa-

*Received by the editors February 4, 2002; accepted for publication (in revised form) June 26, 2002; published electronically February 6, 2003. This paper was partially supported by Fundação Calouste Gulbenkian and FCT–CEDEC.

http://www.siam.org/journals/sicon/41-6/36392.html

†Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, P–4200–465 Porto, Portugal (poliv@fe.up.pt).

‡Laboratoire de Mathématiques Appliquées (CNRS–UMR 6620), Université Blaise Pascal, F–63177 Aubière Cedex, France (Jean.Picard@math.univ-bpclermont.fr).

tion

$$\left\{\begin{array}{rcl} dx_t^{(1)} &=& f_1(x_t^{(1)}, x_t^{(2)})\, dt, \\ dx_t^{(2)} &=& f_2(x_t^{(1)}, x_t^{(2)})\, dt + \sigma(x_t^{(1)}, x_t^{(2)})\, dw_t, \end{array}\right. \tag{1.1}$$

with initial condition $X_0 = (x_0^{(1)}, x_0^{(2)})$, and we are concerned by the problem of estimating the signal X_t when the observation process is modelled by the equation

$$dy_t = h(x_t^{(1)})\, dt + \varepsilon\, d\bar{w}_t, \tag{1.2}$$

where $\{w_t\}$ and $\{\bar{w}_t\}$ are standard independent real-valued Wiener processes and ε is a small nonnegative parameter. In particular, if $f_1(x_1, x_2) = x_2$, then $x_t^{(1)}$ is the position of some moving body on $\mathbf{R}$, $x_t^{(2)}$ is its speed, the body is submitted to a dynamical force described by f_2 and to a random force described by σ, and one has a noisy observation of the position. This class of problems arises in practice in tracking RADAR applications, for instance, as well as in control and communications engineering. The use of the method of proof introduced in [7] and later extended to [9] in the class of systems (1.1)–(1.2) is not covered by previous work.

If $\varepsilon = 0$ and if the functions h and $x_2 \mapsto f_1(x_1, x_2)$ are injective, then the signal X_t can (at least theoretically) be exactly restored from the observation; we are here interested by the asymptotic case $\varepsilon \to 0$, and we look for a good approximation of the optimal filter

$$\hat{X}_t = (\hat{x}_t^{(1)}, \hat{x}_t^{(2)}) = E\big[X_t \;\big|\; y_s, 0 \le s \le t\big].$$

This approximation should be finite-dimensional (a solution of a finite-dimensional equation driven by y_t).

The same problem has been dealt with in [12] (with σ constant) by means of a formal asymptotic expansion of the optimal filter in a stationary situation. Our aim is to work out a rigorous mathematical study of the filter proposed by [12], namely the solution $M_t = (m_t^{(1)}, m_t^{(2)})$ of

$$dM_t = f(M_t)dt + R_t[dy_t - h(m_t^{(1)})dt], \tag{1.3}$$

$$R_t \stackrel{def}{=} \left[\begin{array}{c} \sqrt{\dfrac{2\sigma(M_t)F_{12}(M_t)}{h'(m_t^{(1)})\,\varepsilon}} \\ \dfrac{\sigma(M_t)}{\varepsilon} \end{array}\right], \tag{1.4}$$

with $F_{12} = \partial f_1/\partial x_2$ and with initial condition $M_0 = E[X_0]$. This filter does in fact correspond to the extended Kalman filter with stationary gain if one neglects the contribution of the derivatives of f other than $\partial f_1/\partial x_2$. The stability of this filter is not evident and requires some assumptions. When it is stable, we prove in this work that

$$x_t^{(1)} - m_t^{(1)} = \mathcal{O}(\varepsilon^{3/4}), \quad x_t^{(2)} - m_t^{(2)} = \mathcal{O}(\varepsilon^{1/4}), \tag{1.5}$$

and

$$\hat{x}_t^{(1)} - m_t^{(1)} = \mathcal{O}(\varepsilon), \quad \hat{x}_t^{(2)} - m_t^{(2)} = \mathcal{O}(\sqrt{\varepsilon}). \tag{1.6}$$

We also verify that (1.6) can be improved when σ is constant, h is linear, and f_1 is linear with respect to x_2. (This case will be referred to as the almost linear case.) The proofs follow the method of [9].

The contents are organized as follows. In section 2, we introduce the assumptions which will be needed in what follows, and we study the filtering error as ε converges to zero; more precisely, we obtain the rate (1.5). In section 3, the error between the approximate filter and the optimal filter is studied, and we prove (1.6). Section 4 is devoted to the almost linear case. Results of numerical simulations that illustrate the performance of this approach are included in section 5.

Notation. The following notation is used:

$$f = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 0 \\ \sigma \end{bmatrix}, \quad H = \begin{bmatrix} h' & 0 \end{bmatrix};$$

$F = \begin{smallmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{smallmatrix}$ and $\Sigma' = \begin{smallmatrix} 0 & 0 \\ \Sigma'_{21} & \Sigma'_{22} \end{smallmatrix}$ are the Jacobian matrices of f and Σ; $\nabla_0 \Phi = \frac{\partial \Phi}{\partial X_0}$ is either a 2×2 matrix (if Φ is $\mathbf{R}^2$-valued) or a line-vector (if Φ is real-valued); see section 3. The symbol * is used for the transposition of matrices.

When describing the behavior of approximate filters, we will write asymptotic expressions with the meaning given by the following definition.

DEFINITION 1.1. *Consider a real- or vector-valued stochastic process $\{\xi_t\}$. If β is real and $p \geq 1$, we will write that*

$$\xi_t = \mathcal{O}(\varepsilon^\beta) \quad \textit{in} \quad L^p$$

when, for some $q \geq 0$, $\alpha > 0$, and some positive constants C_1, C_2, c_3,

$$E\big[\|\xi_t\|^p\big]^{1/p} \leq \frac{C_1}{\varepsilon^q} e^{-c_3 t/\varepsilon^\alpha} + C_2 \varepsilon^\beta$$

for $t \geq 0$ and ε small. In this situation, the process $\{\xi_t\}$ is usually said to converge to zero with rate of order ε^β, in a time scale of order ε^α.

2. Estimation of $X_t - M_t$. The following assumptions will be used throughout this article. The last one depends on a parameter $\delta \geq 1$.

(H1) X_0 is a random variable, the moments of which are finite.

(H2) $\{w_t\}$ and $\{\bar{w}_t\}$ are standard independent Wiener processes independent of X_0.

(H3) The function h is C^3 with bounded derivatives, and h' is positive.

(H4) The function f is C^3 with bounded partial derivatives, and $F_{12} = \partial f_1 / \partial x_2$ is positive.

(H5) The function σ is C^2 with bounded partial derivatives.

(H6.δ) One has

$$\frac{1}{\delta} \leq \sigma(x) \leq \delta, \quad \frac{1}{\delta} \leq h'(x_1) \leq \delta, \quad \frac{1}{\delta} \leq F_{12}(x) \leq \delta$$

for any $x = (x_1, x_2)$.

Remark 2.1. In order to reduce the notation in (H6.δ), system (1.1)–(1.2) has been rescaled. Indeed, if we assume instead that one has

$$\frac{1}{\delta} \leq \frac{\sigma(x)}{\bar{\sigma}} \leq \delta, \quad \frac{1}{\delta} \leq \frac{h'(x_1)}{\bar{H}} \leq \delta, \quad \frac{1}{\delta} \leq \frac{F_{12}(x)}{\bar{F}} \leq \delta$$

for any $x = (x_1, x_2)$ and for some positive $\bar\sigma$, $\bar H$, and $\bar F$ and if we replace the processes $x_t^{(1)}$, $x_t^{(2)}$, and y_t by $x_t^{(1)}/(\bar\sigma\bar F)$, $x_t^{(2)}/\bar\sigma$, and $y_t/(\bar\sigma\bar F\bar H)$, then the functions f_1, f_2, σ, and h are replaced, respectively, by

$$f_1(\bar\sigma\bar F x_1, \bar\sigma x_2) \,/\, (\bar\sigma\bar F), \quad f_2(\bar\sigma\bar F x_1, \bar\sigma x_2) \,/\, \bar\sigma,$$

$$\sigma(\bar\sigma\bar F x_1, \bar\sigma x_2) \,/\, \bar\sigma, \quad h(\bar\sigma\bar F x) \,/\, (\bar\sigma\bar F\bar H),$$

and ε is replaced by $\varepsilon/(\bar\sigma\bar F\bar H)$. We can apply the filter (1.3) to this new system, and we obtain $m_t^{(1)}/(\bar\sigma\bar F)$ and $m_t^{(2)}/\bar\sigma$. This shows that the problem can be reduced to the case $\bar\sigma = \bar F = \bar H = 1$.

Assumption (H6.δ) says that the system does not contain too much nonlinearity; when it is not satisfied, there may be a small positive probability for the filter to lose the signal (see [10] for a similar problem). This is a rather restrictive condition, so we discuss at the end of the section the general case in which it does not hold.

We consider the system (1.1)–(1.2) and the filter (1.3). We let $\mathcal{F}_t$ be the filtration generated by $(X_0, w_t, \bar w_t)$ and $\mathcal{Y}_t$ the filtration generated by (y_t).

THEOREM 2.1. *Assume* (H1)–(H5). *For* $1 < \delta < 2^{1/5}$, *if* (H6.δ) *holds, then one has*

$$x_t^{(1)} - m_t^{(1)} = \mathcal{O}(\varepsilon^{3/4}), \quad x_t^{(2)} - m_t^{(2)} = \mathcal{O}(\varepsilon^{1/4})$$

in L^p *for any* $p \geq 1$.

Consider a change of basis defined by a matrix T and its inverse T^{-1}, where

$$T \overset{def}{=} \begin{bmatrix} \sqrt{2/\varepsilon} & -1 \\ 0 & 1 \end{bmatrix}, \qquad T^{-1} = \begin{bmatrix} \sqrt{\varepsilon/2} & \sqrt{\varepsilon/2} \\ 0 & 1 \end{bmatrix}.$$

Then consider the process

$$Z_t \overset{def}{=} T(X_t - M_t). \tag{2.1}$$

We are going to check that Z_t is the solution of a linear stochastic differential equation; the study of the exponential stability of this equation will enable the estimation of both components of Z_t, and the theorem will immediately follow.

An equation for Z_t. From (1.1)–(1.3), we have

$$\begin{aligned} d(X_t - M_t) = {} & \big(f(X_t) - f(M_t)\big)dt - R_t\big(h(x_t^{(1)}) - h(m_t^{(1)})\big)dt \\ & + \begin{bmatrix} 0 & -\sqrt{\dfrac{2\varepsilon\,\sigma(M_t)F_{12}(M_t)}{h'(m_t^{(1)})}} \\ \sigma(X_t) & -\sigma(M_t) \end{bmatrix} \begin{bmatrix} dw_t \\ d\bar w_t \end{bmatrix}. \end{aligned}$$

In this equation, we introduce the Taylor expansions for the functions f and h,

$$f(X_t) - f(M_t) = F(\xi_t, \mu_t)(X_t - M_t)$$

and

$$h(x_t^{(1)}) - h(m_t^{(1)}) = h'(\eta_t)(x_t^{(1)} - m_t^{(1)}),$$

where $\{\xi_t\}$, $\{\mu_t\}$, and $\{\eta_t\}$ are $\mathbf{R}^2$- and $\mathbf{R}$-valued processes depending on $\{X_t\}$ and $\{M_t\}$, and

$$F(\xi_t, \mu_t) \stackrel{def}{=} \begin{bmatrix} F_{11}(\xi_t) & F_{12}(\xi_t) \\ \\ F_{21}(\mu_t) & F_{22}(\mu_t) \end{bmatrix}.$$

We obtain a linear equation for $X_t - M_t$. By applying the transformation (2.1), we deduce for Z_t an equation of the type

$$dZ_t = A_t\, Z_t\, dt + U_t \begin{bmatrix} dw_t \\ d\bar{w}_t \end{bmatrix}. \tag{2.2}$$

The precise computation shows that

$$A_t = T\big(F(\xi_t, \mu_t) - R_t H(\eta_t)\big)T^{-1} = \frac{\bar{A}_t}{\sqrt{2\varepsilon}} + \widetilde{A}_t\,,$$

with

$$\bar{A}_t^{(11)} = -2h'(\eta_t)\sqrt{\frac{F_{12}(M_t)\sigma(M_t)}{h'(m_t^{(1)})}} + h'(\eta_t)\sigma(M_t), \quad \bar{A}_t^{(12)} = \bar{A}_t^{(11)} + 2F_{12}(\xi_t),$$

$$\bar{A}_t^{(21)} = \bar{A}_t^{(22)} = -h'(\eta_t)\sigma(M_t),$$

and where $\widetilde{A}_t$ is a 2×2 matrix-valued process which is uniformly bounded as ε converges to 0; similarly, the matrix-valued process U_t is also uniformly bounded.

Stability of A_t. If $\delta = 1$, then $h' = F_{12} = \sigma = 1$, so $\bar{A}_t$ is the constant matrix

$$\bar{A}_t = \begin{bmatrix} -1 & 1 \\ -1 & -1 \end{bmatrix},$$

and

$$\bar{A}_t + \bar{A}_t^* = -2\, I.$$

In the general case $\delta > 1$, the coefficients of $\bar{A}_t + \bar{A}_t^*$ can be controlled so that this matrix is uniformly close to $-2I$ if δ is close to 1; in particular, for $1 < \delta < 2^{1/5}$, there exists $0 < \alpha < \alpha' < \sqrt{2}$ such that

$$\bar{A}_t + \bar{A}_t^* \leq -\alpha'\sqrt{2}\, I$$

and, therefore,

$$A_t + A_t^* \leq -\frac{\alpha}{\sqrt{\varepsilon}} I \tag{2.3}$$

if ε is small.

End of the proof of Theorem 2.1. Our goal is now to deduce an estimate of Z_t in L^{2p} for the p integer. From Itô's formula and (2.2), the process $\|Z_t\|^2 = Z_t^* Z_t$ is the solution of

$$d\|Z_t\|^2 = Z_t^*(A_t + A_t^*)Z_t\, dt + \operatorname{trace}(U_t^* U_t)\, dt + 2\, Z_t^* U_t \begin{bmatrix} dw_t \\ d\bar{w}_t \end{bmatrix}.$$

We deduce that the moment of order p of $\|Z_t\|^2$ is finite and that

$$\frac{d}{dt}E[\|Z_t\|^{2p}] = p\,E[\|Z_t\|^{2p-2}Z_t^*(A_t+A_t^*)Z_t] + p\,E[\|Z_t\|^{2p-2}\mathrm{trace}(U_t^*U_t)]$$
$$+2\,p\,(p-1)\,E[\|Z_t\|^{2p-4}\,\|U_t^*Z_t\|^2]\,.$$

From (2.3), one has

$$Z_t^*(A_t+A_t^*)Z_t \le -\frac{\alpha}{\sqrt{\varepsilon}}\,\|Z_t\|^2.$$

As a consequence of the Cauchy–Schwarz inequality, one has

$$\|U_t^*Z_t\|^2 \le \mathrm{trace}(U_t^*U_t)\,\|Z_t\|^2\,.$$

Thus we obtain the inequality

$$\frac{d}{dt}E[\|Z_t\|^{2p}] \le -p\,\frac{\alpha}{\sqrt{\varepsilon}}E[\|Z_t\|^{2p}] + p\,(2p-1)E[\|Z_t\|^{2p-2}\mathrm{trace}(U_t^*U_t)]$$
$$\le -p\,\frac{\alpha}{\sqrt{\varepsilon}}E[\|Z_t\|^{2p}] + C_pE[\|Z_t\|^{2p-2}]\,.$$

Moreover, there exists C_p' such that

$$C_p\|Z_t\|^{2p-2} \le p\,\frac{\alpha}{2\sqrt{\varepsilon}}\|Z_t\|^{2p} + C_p'\varepsilon^{(p-1)/2},$$

and so

$$\frac{d}{dt}E[\|Z_t\|^{2p}] \le -\frac{\alpha}{2\sqrt{\varepsilon}}\,p\,E[\|Z_t\|^{2p}] + C_p'\varepsilon^{(p-1)/2}\,.$$

By solving this differential inequality, one obtains that, for some $C_p'' > 0$,

$$E[\|Z_t\|^{2p}] \le C_p''\varepsilon^{p/2} + C_p''\,E[\|Z_0\|^{2p}]\,e^{-\alpha\,p\,t/(2\sqrt{\varepsilon})}\,. \tag{2.4}$$

Thus Z_t is $\mathcal{O}(\varepsilon^{1/4})$, and the order of magnitude of the components of $X_t - M_t$ follows from (2.1) and the form of T^{-1}. □

We remark in (2.4) that the time scale of the estimation is of order $\sqrt{\varepsilon}$; one can compare it with the time scale ε obtained when the observation function is injective (see, for instance, [7]). This means that here it takes more time to estimate the signal, and this is not surprising since the second component of the signal is not well observed. There are also other systems where the time scale is not the same for the different components of the signal (see [10]).

In Theorem 2.1, we need the assumption (H6.δ), which is a restriction to the nonlinearity of the system; otherwise, it is difficult to ensure that the filter does not lose the signal. (This problem also occurs in [10].) Actually, we have chosen the filter (1.3) because it gives a good approximation of $\hat{X}_t$ (see the next section), but it is not the most stable one. If in (1.4) we replace the processes $\sigma(M_t)$, $F_{12}(M_t)$, and $h'(m_t^{(1)})$ by constant numbers $\bar{\sigma}$, $\bar{F}$, and $\bar{H}$, then we obtain a filter with constant gain; we can again work out the previous estimations and prove that the result of Theorem 2.1 holds for this filter without (H6.δ) as soon as

$$\frac{\max F_{12}}{\bar{F}} < 2\frac{\min h'}{\bar{H}}.$$

Thus we have two filters—a filter which is stable and tracks the signal under rather weak assumptions and the filter (1.3) which seems more fragile but gives (under good stability assumptions) a better approximation of the optimal filter.

3. Estimation of $\hat{X}_t - M_t$. The main result contained in this section is Theorem 3.1, which states the rate of convergence of the approximate filter considered in this paper toward the optimal filter. In order to give a proof of this theorem, a sequence of steps is needed: a change of probability measure, the differentiation with respect to the initial condition, and an integration by parts formula. A similar method of proof is adopted in [9]. As in Theorem 2.1, we may have a problem of stability in the general nonlinear case.

THEOREM 3.1. *Consider a finite time interval* $[0,\tau]$. *Assume* (H1)–(H6.δ) *and the following:*

(H7) *The law of* X_0 *has a* C^1 *positive density* p_0 *with respect to the Lebesgue measure and* $\nabla p_0(X_0)/p_0(X_0)$ *is in* L^2.

If δ *in* (H6.δ) *is close enough to* 1, *in the sense that* $1 < \delta < 2^{2/9}$, *then the filter* M_t *given by* (1.3) *satisfies*

$$\hat{x}_t^{(1)} - m_t^{(1)} = \mathcal{O}(\varepsilon), \qquad \hat{x}_t^{(2)} - m_t^{(2)} = \mathcal{O}(\sqrt{\varepsilon})$$

in L^2.

The rest of this section is devoted to the proof of this theorem.

Consider the matrix

$$P_t \overset{def}{=} \begin{bmatrix} \dfrac{1}{h'(m_t^{(1)})}\sqrt{\dfrac{2\sigma(M_t)F_{12}(M_t)}{h'(m_t^{(1)})}}\,\varepsilon^{3/2} & \dfrac{\sigma(M_t)}{h'(m_t^{(1)})}\varepsilon \\ \dfrac{\sigma(M_t)}{h'(m_t^{(1)})}\varepsilon & \sigma(M_t)\sqrt{\dfrac{2\sigma(M_t)}{h'(m_t^{(1)})F_{12}(M_t)}}\,\varepsilon^{1/2} \end{bmatrix},$$

which depends only on M_t. Notice that P_t is the solution of the stationary Riccati equation

$$-\frac{1}{\varepsilon^2}P_t H^*(M_t)H(M_t)P_t + \widetilde{F}(M_t)P_t + P_t\widetilde{F}^*(M_t) + \Sigma(M_t)\Sigma^*(M_t) = 0 \tag{3.1}$$

with

$$\widetilde{F}(M_t) = \begin{bmatrix} 0 & F_{12}(M_t) \\ 0 & 0 \end{bmatrix}$$

and that the process R_t of (1.4) is

$$R_t = \frac{P_t}{\varepsilon^2}H^*(M_t). \tag{3.2}$$

We will also need the inverse of P_t, namely,

$$P_t^{-1} = \begin{bmatrix} h'(m_t^{(1)})\sqrt{\dfrac{2h'(m_t^{(1)})}{\sigma(M_t)F_{12}(M_t)}}\,\varepsilon^{-3/2} & -\dfrac{h'(m_t^{(1)})}{\sigma(M_t)}\varepsilon^{-1} \\ -\dfrac{h'(m_t^{(1)})}{\sigma(M_t)}\varepsilon^{-1} & \dfrac{1}{\sigma(M_t)}\sqrt{\dfrac{2h'(m_t^{(1)})F_{12}(M_t)}{\sigma(M_t)}}\,\varepsilon^{-1/2} \end{bmatrix}.$$

Change of probability measure. Our random variables can be viewed as functions of the initial condition X_0 and of the Wiener processes w and $\bar{w}$. We are going to

make a change of variables; in view of the Girsanov theorem, this can be viewed as a change of probability measure; however, all the estimations will be made under the original probability P. Thus consider the new probability measure which is given on $\mathcal{F}_t$ by

$$\left.\frac{d\dot{P}}{dP}\right|_{\mathcal{F}_t} = L_t^{-1}\,,$$

where

$$L_t^{-1} = \exp\left\{-\frac{1}{\varepsilon}\int_0^t h(x_s^{(1)})\,d\bar{w}_s - \frac{1}{2\varepsilon^2}\int_0^t h^2(x_s^{(1)})\,ds\right\}.$$

The probability $\dot{P}$ is the so-called reference probability, and one checks easily from the Girsanov theorem that y_t/ε and w_t are standard independent Wiener processes under $\dot{P}$. Let us define now the probability measure $\widetilde{P}$ on $\mathcal{F}_t$ by

$$\left.\frac{d\widetilde{P}}{d\dot{P}}\right|_{\mathcal{F}_t} = \Lambda_t^{-1}\,,$$

where

$$\Lambda_t^{-1} = \exp\left\{\int_0^t \Sigma^*(M_s)P_s^{-1}(X_s - M_s)\,dw_s - \frac{1}{2}\int_0^t (\Sigma^*(M_s)P_s^{-1}(X_s - M_s))^2\,ds\right\}.$$

Then the processes

$$\widetilde{w}_t = w_t - \int_0^t \Sigma^*(M_s)P_s^{-1}(X_s - M_s)\,ds$$

and y_t/ε are standard independent Wiener processes under $\widetilde{P}$. On the other hand, one has

$$dX_t = f(X_t)\,dt + \Sigma(X_t)\Sigma^*(M_t)P_t^{-1}(X_t - M_t)\,dt + \Sigma(X_t)\,d\widetilde{w}_t \tag{3.3}$$

and

$$\begin{aligned}\log(L_t\Lambda_t) = {}& \frac{1}{\varepsilon^2}\int_0^t h(x_s^{(1)})\,dy_s - \frac{1}{2\varepsilon^2}\int_0^t h^2(x_s^{(1)})\,ds - \int_0^t \Sigma^*(M_s)P_s^{-1}(X_s - M_s)\,d\widetilde{w}_s \\ & -\frac{1}{2}\int_0^t (\Sigma^*(M_s)P_s^{-1}(X_s - M_s))^2\,ds\,.\end{aligned} \tag{3.4}$$

Differentiation with respect to the initial condition and an estimation. The random variables involved in our computation can now be viewed as functions of X_0, $\{\widetilde{w}_t\}$, and $\{y_t\}$; let us denote by ∇_0 the differentiation with respect to the initial condition X_0 (computed in L^p). In particular, we can see on (3.3) and (3.4) that the processes X_t and $\log(L_t\Lambda_t)$ are differentiable, and we obtain matrix- and vector-valued processes, respectively. Our aim is to estimate the process

$$V_t \overset{def}{=} (\nabla_0 \log(L_t\Lambda_t)\,(\nabla_0 X_t)^{-1} + (X_t - M_t)^* P_t^{-1})U \tag{3.5}$$

with

$$U \overset{def}{=} \begin{bmatrix} 1 & 1 \\ 0 & \sqrt{2/\varepsilon} \end{bmatrix}, \qquad U^{-1} = \begin{bmatrix} 1 & -\sqrt{\varepsilon/2} \\ 0 & \sqrt{\varepsilon/2} \end{bmatrix}.$$

Then an integration by parts will enable us to conclude.

By applying the operator ∇_0 to (3.4), one gets

$$\begin{aligned} \nabla_0 \log(L_t\Lambda_t) &= \frac{1}{\varepsilon^2}\int_0^t h'(x_s^{(1)})\nabla_0 x_s^{(1)}\,(dy_s - h(x_s^{(1)})ds) \\ &\quad - \int_0^t \Sigma^*(M_s)P_s^{-1}\nabla_0 X_s\,(d\widetilde{w}_s + \Sigma^*(M_s)P_s^{-1}(X_s - M_s)ds) \\ &= \frac{1}{\varepsilon}\int_0^t h'(x_s^{(1)})\nabla_0 x_s^{(1)}\,d\bar{w}_s - \int_0^t \Sigma^*(M_s)P_s^{-1}\nabla_0 X_s\,dw_s\,. \end{aligned} \tag{3.6}$$

We can also differentiate (3.3), and, if Σ' is the Jacobian matrix of Σ, we obtain

$$d(\nabla_0 X_t) = [F(X_t) + \Sigma(X_t)\Sigma^*(M_t)P_t^{-1}]\nabla_0 X_t\,dt + \Sigma'(X_t)\nabla_0 X_t\,dw_t.$$

The matrix $\nabla_0 X_t$ is invertible, and Itô's calculus shows that

$$\begin{aligned} d(\nabla_0 X_t)^{-1} &= -(\nabla_0 X_t)^{-1}[F(X_t) + \Sigma(X_t)\Sigma^*(M_t)P_t^{-1} - \Sigma'^2(X_t)]\,dt \\ &\quad -(\nabla_0 X_t)^{-1}\Sigma'(X_t)dw_t\,. \end{aligned} \tag{3.7}$$

From this equation and (3.6), one can write that

$$\begin{aligned} d\left(\nabla_0 \log(L_t\Lambda_t)(\nabla_0 X_t)^{-1}\right) &= \frac{1}{\varepsilon}H(X_t)\,d\bar{w}_t - \Sigma^*(M_t)P_t^{-1}\,dw_t \\ &\quad -\nabla_0 \log(L_t\Lambda_t)(\nabla_0 X_t)^{-1}\Sigma'(X_t)\,dw_t \\ &\quad -\nabla_0 \log(L_t\Lambda_t)(\nabla_0 X_t)^{-1} \\ &\qquad .[F(X_t) + \Sigma(X_t)\Sigma^*(M_t)P_t^{-1} - \Sigma'^2(X_t)]dt \\ &\quad +\Sigma^*(M_t)P_t^{-1}\Sigma'(X_t)dt \end{aligned} \tag{3.8}$$

since one has $h'(x_t^{(1)})\nabla_0 x_t^{(1)}(\nabla_0 X_t)^{-1} = H(X_t)$.

On the other hand, from the equations of X_t and M_t ((1.1) and (1.3), respectively), one has

$$d(X_t - M_t) = [f(X_t) - f(M_t)]\,dt - R_t[h(x_t^{(1)}) - h(m_t^{(1)})]\,dt - R_t\varepsilon\,d\bar{w}_t + \Sigma(X_t)\,dw_t\,.$$

By writing the differential of P_t^{-1} in the form

$$dP_t^{-1} = J_t^{(1)}\,dt + J_t^{(2)}\,d\bar{w}_t,$$

we obtain

$$\begin{aligned} d\left((X_t - M_t)^*P_t^{-1}\right) &= [f^*(X_t) - f^*(M_t) - R_t^*(h(x_t^{(1)}) - h(m_t^{(1)}))]P_t^{-1}\,dt \\ &\quad +\Sigma^*(X_t)P_t^{-1}\,dw_t - \varepsilon\,R_t^*P_t^{-1}\,d\bar{w}_t \\ &\quad +(X_t - M_t)^*[J_t^{(1)}\,dt + J_t^{(2)}\,d\bar{w}_t] - \varepsilon\,R_t^*J_t^{(2)}\,dt\,. \end{aligned} \tag{3.9}$$

One can write the Taylor expansions for f and h,

$$f(X_t) - f(M_t) = F(M_t)(X_t - M_t) + \phi_t\,,$$
$$h(x_t^{(1)}) - h(m_t^{(1)}) = H(M_t)(X_t - M_t) + \gamma_t,$$

with

$$\|\phi_t\| \le C\|X_t - M_t\|^2, \quad |\gamma_t| \le C|x_t^{(1)} - m_t^{(1)}|^2.$$

By using these expansions together with the consequence of (3.2),

$$H^*(M_t)R_t^*P_t^{-1} = \frac{1}{\varepsilon^2}H^*(M_t)H(M_t),$$

in (3.9), we obtain

$$\begin{aligned} d((X_t - M_t)^*P_t^{-1}) = (X_t - M_t)^*\left(F^*(M_t)P_t^{-1} - \frac{1}{\varepsilon^2}H^*(M_t)H(M_t)\right)dt \\ +\Sigma^*(X_t)P_t^{-1}\,dw_t - \varepsilon\,R_t^*P_t^{-1}d\bar{w}_t \\ +(X_t - M_t)^*[J_t^{(1)}\,dt + J_t^{(2)}\,d\bar{w}_t] - \varepsilon\,R_t^*J_t^{(2)}\,dt \\ +(\phi_t^* - \gamma_t R_t^*)P_t^{-1}\,dt\,. \end{aligned}$$

By adding this equation to (3.8), we obtain that the process V_t of (3.5) satisfies

$$\begin{aligned} d\big(V_tU^{-1}\big) = -V_tU^{-1}[F(X_t) + \Sigma(X_t)\Sigma^*(M_t)P_t^{-1} - \Sigma'^2(X_t)]dt - V_tU^{-1}\,\Sigma'(X_t)\,dw_t \\ +\frac{1}{\varepsilon}H(X_t)d\bar{w}_t - \Sigma^*(M_t)P_t^{-1}dw_t + \Sigma^*(M_t)P_t^{-1}\Sigma'(X_t)dt \\ +(X_t - M_t)^*S_tdt + (X_t - M_t)^*P_t^{-1}\Sigma'(X_t)dw_t + \Sigma^*(X_t)P_t^{-1}dw_t \\ -\varepsilon R_t^*P_t^{-1}d\bar{w}_t + (X_t - M_t)^*[J_t^{(1)}dt + J_t^{(2)}d\bar{w}_t] - \varepsilon R_t^*J_t^{(2)}dt \\ +(\phi_t^* - \gamma_t R_t^*)P_t^{-1}dt, \end{aligned} \tag{3.10}$$

where S_t is the matrix given by

$$\begin{aligned} S_t \stackrel{def}{=} -\frac{1}{\varepsilon^2}H^*(M_t)H(M_t) + F^*(M_t)P_t^{-1} + P_t^{-1}F(X_t) \\ +P_t^{-1}\Sigma(X_t)\Sigma^*(M_t)P_t^{-1} - P_t^{-1}\Sigma'^2(X_t)\,. \end{aligned} \tag{3.11}$$

Consider also the matrix-valued process

$$A_t \stackrel{def}{=} -U^{-1}[F(X_t) + \Sigma(X_t)\Sigma^*(M_t)P_t^{-1} - \Sigma'^2(X_t)]U\,. \tag{3.12}$$

Then (3.10) can be written in the form

$$\begin{aligned} dV_t &= V_tA_tdt - V_tU^{-1}\Sigma'(X_t)Udw_t + J_t^{(3)}dt + J_t^{(4)}dw_t + J_t^{(5)}d\bar{w}_t, \\ V_0 &= (X_0 - M_0)^*P_0^{-1}U, \end{aligned} \tag{3.13}$$

where

$$\begin{aligned} J_t^{(3)} &= \Sigma^*(M_t)P_t^{-1}\Sigma'(X_t)U + (X_t - M_t)^*S_tU + (X_t - M_t)^*J_t^{(1)}U \\ &\quad -\varepsilon R_t^*J_t^{(2)}U + (\phi_t^* - \gamma_t R_t^*)P_t^{-1}U, \\ J_t^{(4)} &= (\Sigma^*(X_t) - \Sigma^*(M_t))P_t^{-1}U + (X_t - M_t)^*P_t^{-1}\Sigma'(X_t)U, \\ J_t^{(5)} &= \frac{1}{\varepsilon}H(X_t)U - \varepsilon R_t^*P_t^{-1}U + (X_t - M_t)^*J_t^{(2)}U \\ &= \frac{1}{\varepsilon}(H(X_t) - H(M_t))U + (X_t - M_t)^*J_t^{(2)}U \end{aligned}$$

(apply (3.2) for the last line). We deduce that $E[\|V_0\|^2]$ is of order ε^{-3} and that

$$\begin{aligned}\frac{d}{dt}E[\|V_t\|^2] &= E[V_t(A_t + A_t^*)V_t^*] + 2E[J_t^{(3)}V_t^*] \\ &\quad + E[\|V_tU^{-1}\Sigma'(X_t)U + J_t^{(4)}\|^2] + E[\|J_t^{(5)}\|^2]\,.\end{aligned} \tag{3.14}$$

We have to estimate the terms of the right-hand side.

By computing the matrix A_t, we obtain that

$$A_t = \frac{\bar{A}_t}{\sqrt{2\varepsilon}} + \widetilde{A}_t$$

with

$$\bar{A}_t^{(11)} = -\bar{A}_t^{(21)} = -h'(m_t^{(1)})\sigma(X_t)\,,$$

$$\bar{A}_t^{(12)} = -2\,F_{12}(X_t) - h'(m_t^{(1)})\sigma(X_t) + 2\,\sigma(X_t)\sqrt{\frac{h'(m_t^{(1)})F_{12}(M_t)}{\sigma(M_t)}}\,,$$

$$\bar{A}_t^{(22)} = h'(m_t^{(1)})\sigma(X_t) - 2\,\sigma(X_t)\sqrt{\frac{h'(m_t^{(1)})F_{12}(M_t)}{\sigma(M_t)}}\,,$$

and $\widetilde{A}_t$ is uniformly bounded. As in the proof of Theorem 2.1, we see that, if $\delta = 1$, then the matrix $\bar{A}_t$ is simply

$$\bar{A}_t = \begin{bmatrix} -1 & -1 \\ 1 & -1 \end{bmatrix},$$

which satisfies

$$\bar{A}_t + \bar{A}_t^* = -2\,I\,.$$

Thus, for $0 < \alpha < \sqrt{2}$, when δ is close enough to 1, that is, $1 < \delta < 2^{2/9}$, and when ε is small enough, we have

$$A_t + A_t^* \le -\frac{\alpha}{\sqrt{\varepsilon}}I\,. \tag{3.15}$$

We also notice that

$$2J_t^{(3)}V_t^* \le \frac{\alpha}{3\sqrt{\varepsilon}}\|V_t\|^2 + C\sqrt{\varepsilon}\|J_t^{(3)}\|^2$$

and that

$$\|V_tU^{-1}\Sigma'(X_t)U + J_t^{(4)}\|^2 \le C\|V_t\|^2 + 2\|J_t^{(4)}\|^2$$

because $U^{-1}\Sigma'(X_t)U$ is bounded. Thus (3.14) implies that, for small ε,

$$\begin{aligned}\frac{d}{dt}E[\|V_t\|^2] &\le -\frac{\alpha}{3\sqrt{\varepsilon}}E[\|V_t\|^2] + C\sqrt{\varepsilon}\,E[\|J_t^{(3)}\|^2] \\ &\quad + 2\,E[\|J_t^{(4)}\|^2] + E[\|J_t^{(5)}\|^2]\,.\end{aligned} \tag{3.16}$$

Let us first estimate $J_t^{(3)}$. We deduce from the Riccati equation (3.1) satisfied by P_t that the process S_t defined in (3.11) satisfies

$$S_t = \big(F^*(M_t) - \widetilde{F}^*(M_t)\big)P_t^{-1} + P_t^{-1}\big(F(X_t) - \widetilde{F}(M_t)\big) \\ +P_t^{-1}\big(\Sigma(X_t) - \Sigma(M_t)\big)\Sigma^*(M_t)P_t^{-1} - P_t^{-1}\Sigma'^2(X_t).$$

By computing this matrix and applying Theorem 2.1, we check that

$$S_t = \begin{bmatrix} \mathcal{O}(\varepsilon^{-7/4}) & \mathcal{O}(\varepsilon^{-5/4}) \\ \mathcal{O}(\varepsilon^{-5/4}) & \mathcal{O}(\varepsilon^{-3/4}) \end{bmatrix}$$

in the spaces L^p. Thus

$$(X_t - M_t)^* S_t U = \mathcal{O}(\varepsilon^{-1}).$$

The term $\Sigma^*(M_t)P_t^{-1}\Sigma'(X_t)U$ is easily shown to have the same order of magnitude. On the other hand, by looking at the equation of M_t and by applying Itô's formula, we can prove that, for any C^2 function ρ with bounded derivatives, one has

$$d\rho(M_t) = \mathcal{O}(\varepsilon^{-1/4})dt + \mathcal{O}(1)d\bar{w}_t.$$

By applying this result to the functions involved in P_t^{-1}, it appears that

$$J_t^{(1)} = \begin{bmatrix} \mathcal{O}(\varepsilon^{-7/4}) & \mathcal{O}(\varepsilon^{-5/4}) \\ \mathcal{O}(\varepsilon^{-5/4}) & \mathcal{O}(\varepsilon^{-3/4}) \end{bmatrix}, \quad J_t^{(2)} = \begin{bmatrix} \mathcal{O}(\varepsilon^{-3/2}) & \mathcal{O}(\varepsilon^{-1}) \\ \mathcal{O}(\varepsilon^{-1}) & \mathcal{O}(\varepsilon^{-1/2}) \end{bmatrix}.$$

We deduce that the terms of $J_t^{(3)}$ involving $J_t^{(1)}$ and $J_t^{(2)}$ are also of order ε^{-1}. Finally, ϕ_t and γ_t are, respectively, of order $\varepsilon^{1/2}$ and $\varepsilon^{3/2}$, and so the last term is of order ε^{-1}, and we deduce that

$$J_t^{(3)} = \mathcal{O}(\varepsilon^{-1}).$$

We can also estimate $J_t^{(4)}$ and $J_t^{(5)}$ and check that they are of order $\varepsilon^{-3/4}$. Thus (3.16) enables us to conclude that

$$V_t = \mathcal{O}(1/\sqrt{\varepsilon})$$

in L^2. We can take the conditional expectation with respect to $\mathcal{Y}_t$ in this estimation because the conditional expectation is a contraction in L^2; thus $E[V_t|\mathcal{Y}_t]$ is $\mathcal{O}(1/\sqrt{\varepsilon})$ in L^2, and, therefore, we obtain from the definition (3.5) that

$$(\hat{X}_t - M_t)^* P_t^{-1} U = -E[\nabla_0 \log(L_t\Lambda_t)(\nabla_0 X_t)^{-1}U|\mathcal{Y}_t] + \mathcal{O}(1/\sqrt{\varepsilon}). \tag{3.17}$$

Application of an integration by parts formula. The estimation of the right-hand side of (3.17) can be completed by means of an integration by parts formula. It is proved in Lemma 3.4.2 of [9] that, if $G = G(X_0, \widetilde{w}, y)$ is a functional defined on the probability space which is differentiable with respect to the initial condition (in the spaces L^p) and if ∇_0^i is the differentiation with respect to the ith component of X_0, then

$$E[G\nabla_0^i \log(L_t\Lambda_t) + G(p_0^{-1}\,\partial p_0/\partial x_i)(X_0) + \nabla_0^i G|\mathcal{Y}_t] = 0. \tag{3.18}$$

We can write (3.7) in the form

$$\begin{aligned} d(\nabla_0 X_t)^{-1} = &-(\nabla_0 X_t)^{-1}(F(X_t) + \Sigma(X_t)\Sigma^*(M_t)P_t^{-1} - \Sigma'^2(X_t) \\ &+\Sigma'(X_t)(\Sigma^*(M_t)P_t^{-1}(X_t - M_t)))dt \\ &-(\nabla_0 X_t)^{-1}\Sigma'(X_t)d\widetilde{w}_t \end{aligned} \tag{3.19}$$

with $(\nabla_0 X_0)^{-1} = I$. This equation can be differentiated with respect to X_0, and so we can apply the integration by parts formula (3.18) to the coefficients of the matrix $(\nabla_0 X_t)^{-1}$. Denote by $(\nabla_0 X_t)_i^{-1}$ its ith line. Then

$$E[(\nabla_0 X_t)_i^{-1}\nabla_0^i \log(L_t\Lambda_t) + (\nabla_0 X_t)_i^{-1}(p_0^{-1}\,\partial p_0/\partial x_i)(X_0) + \nabla_0^i(\nabla_0 X_t)_i^{-1}|\mathcal{Y}_t] = 0.$$

By summing on i and multiplying by U, we have

$$\begin{aligned} E\Bigg[&\nabla_0 \log(L_t\Lambda_t)(\nabla_0 X_t)^{-1}U + (p_0^{-1}p_0')(X_0)(\nabla_0 X_t)^{-1}U \\ &+\sum_i \nabla_0^i(\nabla_0 X_t)_i^{-1}U \;\Big|\; \mathcal{Y}_t\Bigg] = 0. \end{aligned} \tag{3.20}$$

The first term of (3.20) is exactly the term that we want to estimate in (3.17).

For the second term of (3.20), if

$$\Psi_t \overset{def}{=} (p_0^{-1}p_0')(X_0)\,(\nabla_0 X_t)^{-1}U,$$

we have from (3.7) and (3.12) that

$$\Psi_0 = (p_0^{-1}p_0')(X_0)U, \quad d\Psi_t = \Psi_t A_t\,dt - \Psi_t U^{-1}\Sigma'(X_t)U\,dw_t.$$

We proceed as in the study of (3.13). The stability of the matrix A_t, which has been obtained in (3.15), and the boundedness of $U^{-1}\Sigma'(X_t)U$ imply that $(\nabla_0 X_t)^{-1}$ is exponentially small in L^2, and so the second term is negligible.

Let us study the third term of (3.20). If

$$\Phi_t^i = \nabla_0^i(\nabla_0 X_t)_i^{-1}U,$$

then by differentiating (3.19) and transforming $\widetilde{w}$ back into w, we get

$$\begin{aligned} d\Phi_t^i = &\Phi_t^i A_t dt - \Phi_t^i\,U^{-1}\Sigma'(X_t)U\,dw_t - (\nabla_0 X_t)_i^{-1}\,\nabla_0^i\rho(X_t, M_t)U\,dt \\ &-\Sigma^*(M_t)P_t^{-1}\nabla_0^i X_t(\nabla_0 X_t)_i^{-1}\Sigma'(X_t)U\,dt - (\nabla_0 X_t)_i^{-1}\nabla_0^i\big(\Sigma'(X_t)\big)U\,dw_t \end{aligned}$$

with

$$\rho(X_t, M_t) \overset{def}{=} F(X_t) + \Sigma(X_t)\Sigma^*(M_t)P_t^{-1} - \Sigma'^2(X_t).$$

By summing on i and using

$$\sum_i (\nabla_0 X_t)_i^{-1}\nabla_0^i\rho(X_t, M_t) = \sum_{i,j} \nabla_0^i X_t^j (\nabla_0 X_t)_i^{-1}\frac{\partial\rho}{\partial x_j}(X_t, M_t) = \sum_j \frac{\partial\rho_j}{\partial x_j}(X_t, M_t),$$

where ρ_j is the jth line of ρ, we obtain that $\Phi_t = \sum \Phi_t^i$ is the solution of

$$\begin{aligned}\Phi_0 = 0, \quad d\Phi_t = \Phi_t A_t\,dt - \Phi_t U^{-1}\Sigma'(X_t)U\,dw_t - \sum_j \frac{\partial \rho_j}{\partial x_j}(X_t, M_t)U\,dt \\ -\Sigma^*(M_t)P_t^{-1}\Sigma'(X_t)U\,dt - \frac{\partial \sigma'}{\partial x_2}(X_t)U\,dw_t,\end{aligned} \tag{3.21}$$

where σ' is the Jacobian of σ. A computation shows that

$$\frac{\partial \rho_j}{\partial x_j}(X_t, M_t) = \begin{bmatrix} \mathcal{O}(1) & \mathcal{O}(1) \\ \mathcal{O}(\varepsilon^{-1}) & \mathcal{O}(\varepsilon^{-1/2}) \end{bmatrix}.$$

The multiplication on the right by U yields a process of order ε^{-1}; the term $\Sigma^*(M_t)$ $.P_t^{-1}\Sigma'(X_t)U$ is also $\mathcal{O}(\varepsilon^{-1})$, and the term involving the second derivative of σ is $\mathcal{O}(\varepsilon^{-1/2})$. By proceeding again as in the study of (3.13), we deduce that Φ_t is of order $\varepsilon^{-1/2}$.

Thus (3.17), (3.20), and the estimation of Ψ_t and Φ_t yield

$$(\hat{X}_t - M_t)^* P_t^{-1} U = \mathcal{O}(1/\sqrt{\varepsilon}).$$

We multiply on the right by the matrix $U^{-1}P_t$, the coefficients of which are of order $\varepsilon^{3/2}$ for the first column and ε for the second column, and we deduce the order of $\hat{X}_t - M_t$ which was claimed in the theorem. ∎

4. An almost linear case. It is interesting to consider a particular case in which σ, h', and F_{12} are constant so that the system (1.1)–(1.2) is

$$\begin{cases} dx_t^{(1)} &= \big(f_1^0(x_t^{(1)}) + F_{12}x_t^{(2)}\big)\,dt, \\ dx_t^{(2)} &= f_2(x_t^{(1)}, x_t^{(2)})\,dt + \sigma\,dw_t, \\ dy_t &= h'x_t^{(1)}\,dt + \varepsilon\,d\bar{w}_t. \end{cases} \tag{4.1}$$

In particular, (H6.δ) holds with $\delta = 1$. Then it is possible to improve the upper bounds given in Theorem 3.1. The time interval that we consider may be infinite. The result is stated in the following proposition.

PROPOSITION 4.1. *Assuming that* (H1)–(H7) *hold for* (4.1)*, the filter* M_t *given by* (1.3) *verifies*

$$\hat{x}_t^{(1)} - m_t^{(1)} = \mathcal{O}(\varepsilon^{5/4}), \qquad \hat{x}_t^{(2)} - m_t^{(2)} = \mathcal{O}(\varepsilon^{3/4})$$

in L^2.

Proof. The proof closely follows the sequence of steps adopted in Theorem 3.1. The matrices $P_t = P$ and $R_t = R$ are now constant; the processes $J_t^{(1)}$, $J_t^{(2)}$, $J_t^{(4)}$, and $J_t^{(5)}$ are zero. The order of S_t is improved into

$$S_t = \begin{bmatrix} \mathcal{O}(\varepsilon^{-3/2}) & \mathcal{O}(\varepsilon^{-1}) \\ \mathcal{O}(\varepsilon^{-1}) & \mathcal{O}(\varepsilon^{-1/2}) \end{bmatrix},$$

and

$$|\phi_t^{(1)}| \le C|x_t^{(1)} - m_t^{(1)}|^2 = \mathcal{O}(\varepsilon^{3/2}), \quad |\phi_t^{(2)}| \le C\|X_t - M_t\|^2 = \mathcal{O}(\varepsilon^{1/2})$$

so that

$$J_t^{(3)} = (X_t - M_t)^* S_t U + \phi_t^* P^{-1} U$$

is of order $\varepsilon^{-3/4}$. Thus V_t is $\mathcal{O}(\varepsilon^{-1/4})$, and we obtain $\mathcal{O}(\varepsilon^{-1/4})$ in (3.17).

For the end of the proof, we see that

$$\rho(X_t, M_t) = F(X_t) + \Sigma\Sigma^* P^{-1},$$

and so

$$\frac{\partial \rho_j}{\partial x_j}(X_t, M_t) = \frac{\partial F_j}{\partial x_j}(X_t)$$

is bounded. Multiplication by U yields a process of order $\varepsilon^{-1/2}$, and so the process Φ_t of (3.21) is bounded for small ε. We can conclude that

$$(\hat{X}_t - M_t)^* P_t^{-1} U = \mathcal{O}(\varepsilon^{-1/4})$$

and deduce the proposition. □

With more computational effort, it is possible to extend these results to the case in which the component $x^{(1)}$ is driven by low noise:

$$\text{(4.2)} \qquad \begin{cases} dx_t^{(1)} &= \big(f_1^0(x_t^{(1)}) + F_{12} x_t^{(2)}\big)\, dt + \varepsilon^\gamma\, dw_t^{(1)}, \\ dx_t^{(2)} &= f_2(x_t^{(1)}, x_t^{(2)})\, dt + \sigma\, dw_t^{(2)}, \\ dy_t &= h' x_t^{(1)}\, dt + \varepsilon\, d\bar{w}_t \end{cases}$$

with M_t given by (1.5) and with the gain R_t given by (1.4), as before, if $\gamma > 1/2$, and with R_t given by

$$R_t \stackrel{def}{=} \begin{bmatrix} \sqrt{\dfrac{2\sigma F_{12}}{h'} + 1}\, \dfrac{1}{\sqrt{\varepsilon}} \\ \dfrac{\sigma}{\varepsilon} \end{bmatrix}$$

if $\gamma = 1/2$.

Clearly, Theorem 2.1 extends to system (4.2) as soon as $\gamma \geq 1/2$. This results from the fact that, in the SDE of Z_t, the matrices involved in the martingale terms are still uniformly bounded as ε converges to 0, and the matrix A_t of (2.2) has the same stability property as before.

Regarding the extension of Proposition 4.1 to system (4.2), one can see that, assuming $\gamma \geq 3/4$, the estimation in Proposition 4.1 still holds. This happens because the matrix $\bar{A}_t$ in the decomposition of A_t remains the same. More effort is needed if one considers the cases $1/2 < \gamma < 3/4$ and $\gamma = 1/2$.

Another class of almost linear filtering problems when some of the observations and driving noises are small is considered by Krener [6]. Krener studied the multidimensional case, where nonlinearities depend only on state variables which can be estimated quickly and accurately; that is, the only nonlinearity allowed in (4.2) is that of the function f_2 with respect to $x_t^{(1)}$. Observations with at least two components, instead of one, are also assumed.

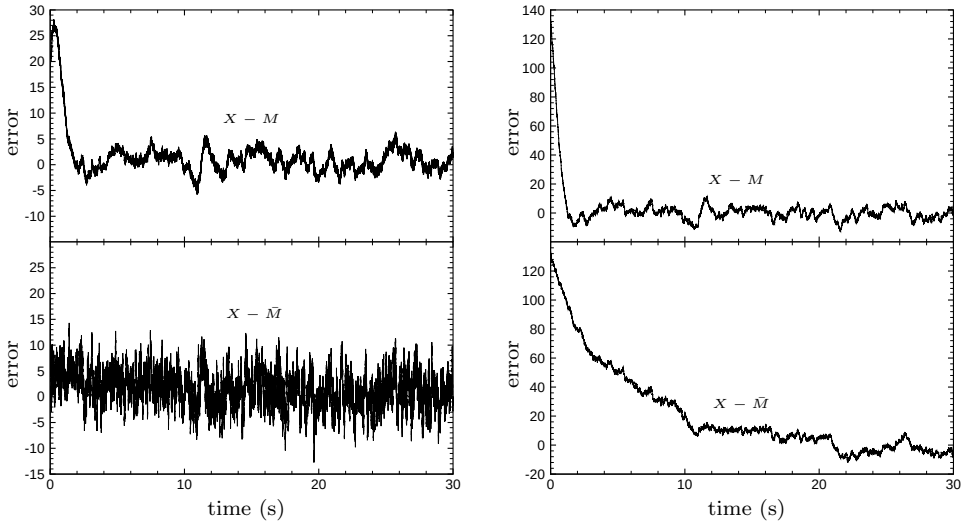


FIG. 5.1. *Estimation errors for the* (a) *first and* (b) *second components of X computed on a single trajectory.*

5. Numerical simulation results. Let us consider the following example illustrating the case of free fall of a body through the atmosphere:

$$\left\{ \begin{array}{rcl} dx_t^{(1)} & = & x_t^{(2)} dt, \\ dx_t^{(2)} & = & (\rho_0 e^{-x_t^{(1)}/k} (x_t^{(2)})^2/(2\beta) - g)\, dt + \sigma dw_t \end{array} \right.$$

and

$$dy_t = \sqrt{(x_t^{(1)})^2 + a^2}\, dt + \varepsilon\, d\bar{w}_t,$$

where $x_t^{(1)}$ is the position of the moving body and $x_t^{(2)}$ is its speed, ρ_0 being the reference air density, k the atmosphere thickness, β the ballistic coefficient of the body, g the acceleration due to gravity, and a the horizontal distance between the body and the measuring device ($\rho_0 = 3.4 \times 10^{-3}\, lb\, s^2/ft^4$, $k = 22 \times 10^3\, ft$, $\beta = 1.6 \times 10^3\, lb^2/ft^4$, $g = 32.2\, ft/s^2$, $\sigma = 5\, ft/s$, and $a = 10^4\, ft$). Figure 5.1 shows the estimation errors obtained from applying the two approximate filters (filter (1.5), noted M_t, and the constant gain filter mentioned at the end of section 2 with $\bar{H} = 0.02$, noted $\bar{M}_t$) to a single trajectory of the state with measurements taken each 0.001 s. The parameter ε is equal to 1 and

$$X_0 \sim \mathcal{N}\left(\left[\begin{array}{c} 3 \times 10^5 \\ -10^3 \end{array} \right], \left[\begin{array}{cc} 900 & 0 \\ 0 & 2 \times 10^4 \end{array} \right] \right).$$

It illustrates the fact that the errors get small very quickly, and one notices that the constant gain filter needs more time than filter (1.5) to attain small errors in the second component.

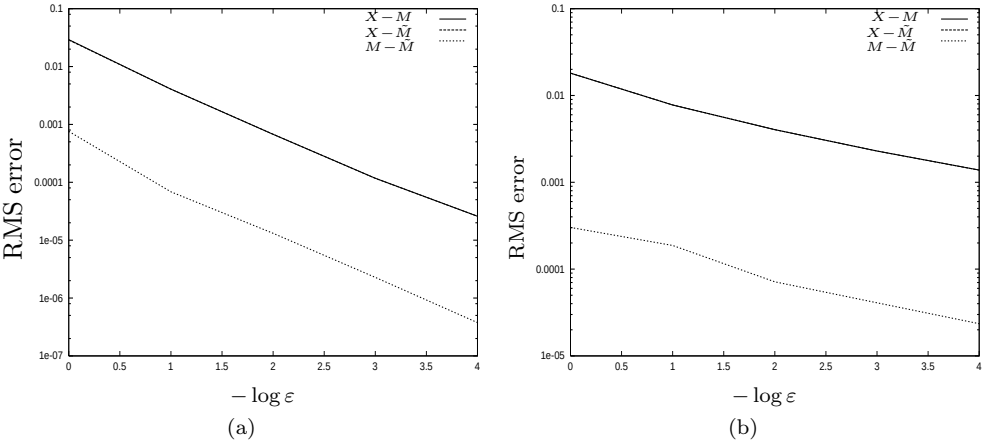


FIG. 5.2. *Estimation errors for the* (a) *first and* (b) *second components of* X.

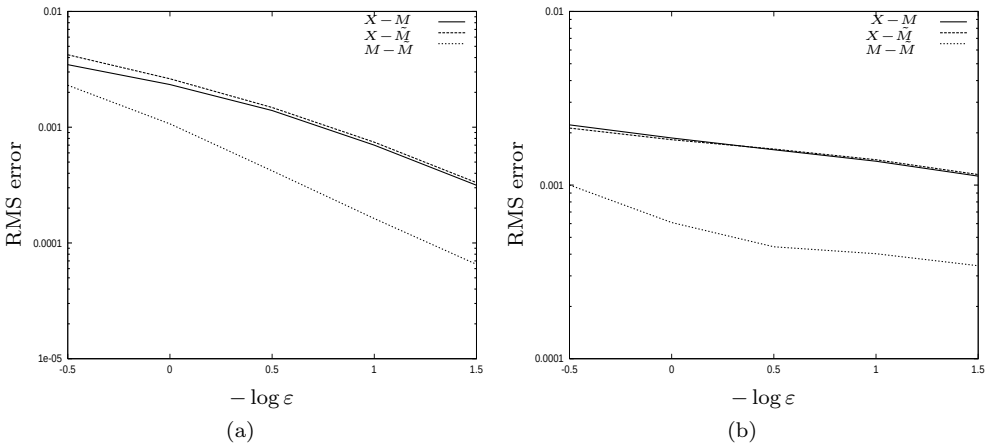


FIG. 5.3. *Estimation errors for the* (a) *first and* (b) *second components of* X ($G = 0.5$).

Figure 5.2 illustrates the asymptotic behavior of the estimation errors when system (1.1)–(1.2) with $f(x_1, x_2) = [x_2 \quad -1.5 \times 10^{-3} x_1^2]^*$, $\sigma = 2$, and $h(x_1) = \sqrt{x_1^2 + 10^8}$ is considered. Although f and h' fail to verify assumption (H3) and, in fact, $\inf h' = 0$, we will assume that the state remains in a bounded domain with high probability, thus assuming that $\inf h' > 1/\sqrt{120}$. The root mean square error between the two approximate filters (with $\bar{H} = 0.18$) was computed for $\varepsilon = 1, 10^{-1}, \ldots, 10^{-4}$ over 200 simulations for both components in the time interval $[0, 5]$. The solid lines exhibit approximate slopes of -0.76 (first component) and -0.28 (second component) which agree with the results in section 2. The error associated with the constant gain filter and that associated to filter (1.5) are very similar.

Figures 5.3–5.5 illustrate the van der Pol oscillator example presented in [12, section 6]: $f(x_1, x_2) = [x_2 \quad -x_1 - x_2]^*$, $\sigma = 1$, and $h(x_1) = 0.606(1 - G)x_1 +$

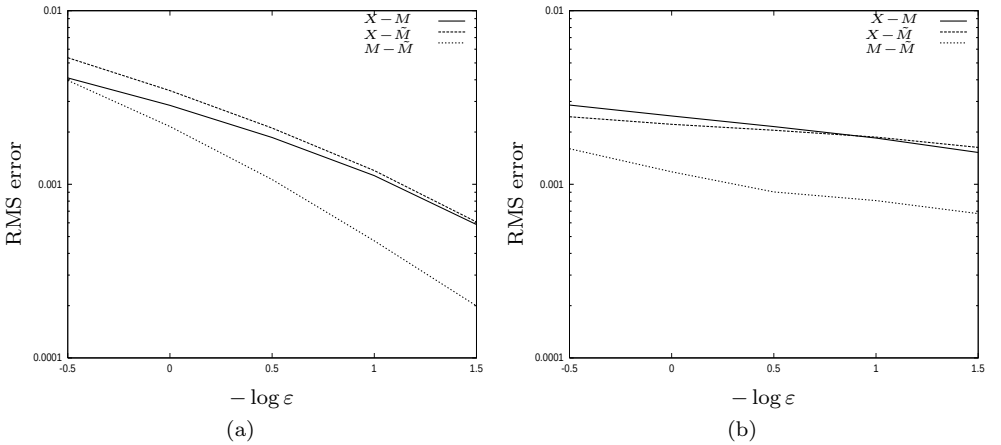


FIG. 5.4. *Estimation errors for the* (a) *first and* (b) *second components of* X $(G = 0.8)$.

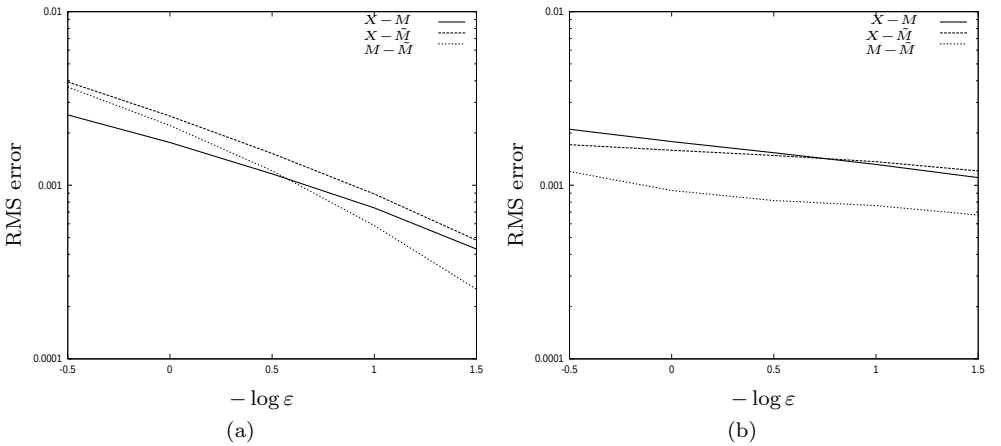


FIG. 5.5. *Estimation errors for the* (a) *first and* (b) *second components of* X $(G = 0.9)$.

Gx_1^3 with $G = 0.5, 0.8, 0.9$, respectively. The time interval $[0, 100]$ was considered. One can observe the increasing benefit of using filter (1.5) as the nonlinearity in the observations gets stronger. The results obtained by using the extended Kalman filter (EKF) are also shown in [12] for comparison.

REFERENCES

[1] A. BENSOUSSAN, *On some approximation techniques in nonlinear filtering*, in Stochastic Differential Systems, Stochastic Control Theory and Applications, IMA Vol. Math. Appl. 10, Springer-Verlag, New York, 1988, pp. 17–31.

[2] W. H. FLEMING AND É. PARDOUX, *Piecewise monotone filtering with small observation noise*, SIAM J. Control Optim., 27 (1989), pp. 1156–1181.

[3] A. GEGOUT-PETIT, *Approximate filter for the conditional law of a partially observed process in nonlinear filtering*, SIAM J. Control Optim., 36 (1998), pp. 1423–1447.

[4] R. Katzur, B. Z. Bobrovsky, and Z. Schuss, *Asymptotic analysis of the optimal filtering problem for one-dimensional diffusions measured in a low noise channel* I, SIAM J. Appl. Math., 44 (1984), pp. 591–604.

[5] R. Katzur, B. Z. Bobrovsky, and Z. Schuss, *Asypmptotic analysis of the optimal filtering problem for one-dimensional diffusions measured in a low noise channel* II, SIAM J. Appl. Math., 44 (1984), pp. 1176–1191.

[6] A. J. Krener, *The asymptotic approximation of nonlinear filters by linear filters*, in Theory and Applications of Nonlinear Control Systems (Stockholm, 1985), North–Holland, Amsterdam, 1986, pp. 359–378.

[7] J. Picard, *Nonlinear filtering of one-dimensional diffusions in the case of a high signal-to-noise ratio*, SIAM J. Appl. Math., 46 (1986), pp. 1098–1125.

[8] J. Picard, *Nonlinear filtering and smoothing with high signal-to-noise ratio*, in Stochastic Processes in Physics and Engineering (Bielefeld, 1986), D. Reidel, Dordrecht, The Netherlands, 1988, pp. 237–251.

[9] J. Picard, *Efficiency of the extended Kalman filter for nonlinear systems with small noise*, SIAM J. Appl. Math., 51 (1991), pp. 843–885.

[10] J. Picard, *Estimation of the quadratic variation of nearly observed semimartingales with application to filtering*, SIAM J. Control Optim., 31 (1993), pp. 494–517.

[11] M. C. Roubaud, *Filtrage linéaire par morceaux avec petit bruit d'observation*, Appl. Math. Optim., 32 (1995), pp. 163–194.

[12] I. Yaesh, B. Z. Bobrovsky, and Z. Schuss, *Asymptotic analysis of the optimal filtering problem for two-dimensional diffusions measured in a low noise channel*, SIAM J. Appl. Math., 50 (1990), pp. 1134–1155.

[13] Q. Zhang, *Nonlinear filtering and control of a switching diffusion with small observation noise*, SIAM J. Control Optim., 36 (1998), pp. 1638–1668.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 6, pp. 1820–1842

WEAK CONVERGENCE OF HYBRID FILTERING PROBLEMS INVOLVING NEARLY COMPLETELY DECOMPOSABLE HIDDEN MARKOV CHAINS*

G. YIN† AND S. DEY‡

Abstract. Concentrating on a class of hybrid discrete-time filtering problems that are modulated by a Markov chain, this work aims to reduce the complexity of the underlying problems. Since the Markov chain has a large state space, the solution of the problem relies on solving a large number of filtering equations. Exploiting the hierarchical structure of the system, it is noted that the transition probability matrix of the Markov chain can be viewed as a nearly decomposable one. It is shown that a reduced system of filtering equations can be obtained by aggregating the states of each recurrent class into one state. Extensions to inclusion of transient states and nonstationary cases are also treated.

Key words. Markov chain, filtering, near complete decomposability, weak convergence

AMS subject classifications. 60F05, 60G35, 60J10, 93E11

PII. S0363012901388464

1. Introduction.

In this work, we concern ourselves with hybrid filtering problems in discrete time. Since a wide variety of problems arising in target tracking, speech recognition, telecommunication, and manufacturing requires solutions of filtering problems involving a hidden Markov chain, in addition to the usual random system disturbances and observation noise, we assume that the system under consideration is influenced by a hidden Markov chain with finite state space. Due to the rapid advances in science and technology, various systems tend to be rather complex and large-scale in nature. As a result, although the state space of the Markov chain is finite, it inevitably contains a large number of states. Our main effort is devoted to reducing the complexity of such filtering problems involving large-scale hidden Markov chains.

In a recent paper, linear systems with coefficients driven by a hidden Markov chain were considered [21]. Discrete-time systems were studied in [1, 6, 12, 28] among others. In [33], Zhang studied hybrid filters in continuous time and treated problems involving non-Gaussian noise. Our study is motivated by these recent developments and stems from the needs in many applications mentioned above.

In the seminal paper [26], Simon and Ando pointed out that various large-scale systems have hierarchical structures. Some of the states vary rapidly, and others change slowly. In addition, these states are also naturally decomposable into different layers or a hierarchy. Such a hierarchy allows one to take advantage and to organize and reorganize the systems accordingly. Based on such ideas, Courtois dealt with the so-called *nearly completely decomposable* Markov chain models [7]. Recently, Dey derived reduced-complexity filtering results for hidden Markov models, in which the underlying Markov chains are nearly completely decomposable [9]. Such hierarchical

*Received by the editors April 16, 2001; accepted for publication (in revised form) June 26, 2002; published electronically February 6, 2003.
http://www.siam.org/journals/sicon/41-6/38846.html
†Department of Mathematics, Wayne State University, Detroit, MI 48202 (gyin@math.wayne.edu). The research of this author was supported in part by the National Science Foundation under grant DMS-9877090.
‡Department of Electrical and Electronic Engineering, University of Melbourne, Parkville, Victoria 3010, Australia (sdey@ee.mu.oz.au).

Markov chains have numerous applications in queueing and computer systems [7], multiple time-scale heterogeneous traffic modelling (e.g., variable bit rate video traffic [27]), manufacturing systems, operations research, and many other biological and physical systems in which a multiple time-scale or hierarchical behavior is involved; see also related work in [3, 5, 14, 18, 22, 23, 25] and the references therein. Taking the approaches of [7] and [9] as our point of departure, to reduce the complexity of the underlying problem, we introduce a small parameter $\varepsilon > 0$ into the system. Note that the small parameter is used to reflect the high contrast of the transition rates of the Markov chain. For the subsequent asymptotic analysis, to obtain the desired results, it is necessary to send $\varepsilon \to 0$, which can serve as a guideline for various applications and for approximation and heuristics. In real applications, however, ε might be a fixed constant, and only the relative order of magnitude of this parameter matters. In our setup, we also consider a nearly completely decomposable Markovian model, in which the hidden Markov chain has a large state space. The transition probability matrix is a sum of a completely decomposable transition matrix and a generator of a continuous-time Markov chain. Following our systematic studies on singularly perturbed Markov chains in both continuous time and discrete time [16, 29, 30, 31, 34], we investigate the asymptotic properties of the filtering problem by means of weak convergence methods. We show that a limit filtering problem can be derived in which the underlying Markov chain is replaced by an averaged chain and the system coefficients are averaged out with respect to the stationary measures of each ergodic class. The reduction of complexity is particularly pronounced when the transition matrix of the Markov chain consists of only one ergodic class. In this case, the limit filtering problem becomes a standard Kalman filter free of Markovian jump processes.

The rest of the paper is arranged as follows. Section 2 presents the precise formulation of the problem and a number of preliminary results that are to be used in our study. Section 3 is concerned with weak convergence analysis and the derivation of limit filtering problems or reduced systems. In order not to disrupt the flow of presentation, all proofs are placed in an appendix. Section 4 proceeds with numerical experiments and simulation studies that demonstrate the relationship between the original system and that of a reduced system. Section 5 gives remarks and a few extensions.

Throughout the paper, we use K to denote a generic positive constant, whose values may be different for different usage. For any $z \in \mathbb{R}^{\ell_1 \times \ell_2}$ with some positive integers ℓ_1 and ℓ_2, z' denotes its transpose. For a suitable function f, f_x and f_{xx} denote its first-order and second-order partial derivatives with respect to x.

2. Formulation and preliminaries. This section gives the precise formulation of the problem to be studied. It also presents some preliminary results needed in the analysis to follow.

2.1. Hybrid filtering problem. Let $\varepsilon > 0$ be a small parameter, and let $\{\alpha_n^\varepsilon\}$ be a (time) homogeneous singularly perturbed Markov chain in discrete time with a finite state space $\mathcal{M}$ having m elements and a transition matrix

$$P^\varepsilon = \widetilde{P} + \varepsilon Q, \tag{2.1}$$

where $\widetilde{P}$ is an $m \times m$ transition matrix and $Q = (q_{\iota\ell})$ is a generator of a continuous-time homogeneous Markov chain, i.e., $q_{\iota\ell} \geq 0$ for $\iota \neq \ell$ and $\sum_\ell q_{\iota\ell} = 0$ for each ι.

Suppose that, for some $T > 0$ and $0 \le n \le \lfloor T/\varepsilon \rfloor$ (where $\lfloor z \rfloor$ denotes the largest integer part of z), $x_n^\varepsilon \in \mathbb{R}^r$ is the state to be estimated, y_n^ε is the corresponding observation, and $A(\iota)$, $C(\iota)$, $\sigma_w(\iota)$, and $\sigma_v(\iota)$ are well defined for each $\iota \in \mathcal{M}$ (i.e., they are finite for each $\iota \in \mathcal{M}$). With initial data x_0 and y_0, the hybrid filtering problem is concerned with the linear system of equations

$$\begin{aligned} x_{n+1}^\varepsilon &= x_n^\varepsilon + \varepsilon A(\alpha_n^\varepsilon) x_n^\varepsilon + \sqrt{\varepsilon}\sigma_w(\alpha_n^\varepsilon) w_n, \\ y_{n+1}^\varepsilon &= y_n^\varepsilon + \varepsilon C(\alpha_n^\varepsilon) x_n^\varepsilon + \sqrt{\varepsilon}\sigma_v(\alpha_n^\varepsilon) v_n, \end{aligned} \tag{2.2}$$

where $\{w_n\}$ and $\{v_n\}$ are the system disturbance and the observation noise, respectively. For ease of presentation, in what follows, we will suppress the floor-function notation $\lfloor \cdot \rfloor$ and write it as $0 \le n \le T/\varepsilon$ throughout. The use of the $\sqrt{\varepsilon}$ in the noise terms stems from the central limit scaling. Precise conditions on the noises will be provided later. In what follows, we will show that, as $\varepsilon \to 0$, the above filtering problem has a limit. The limit filtering problem is still modulated by a Markov chain. However, the total number of states of the limit Markov chain is equal to the number of recurrent groups or clusters l. As mentioned before, typically $l \ll m$, and by considering this limit filtering problem, substantial computational savings can be obtained. Although (2.2) is a discrete-time filtering problem, the limit under appropriate scaling is a continuous-time hybrid filtering problem. In the rest of the paper, our main effort is devoted to deriving the limit filtering problem. For solutions of continuous-time hybrid filtering problems involving jump Markov processes, see [4, 8, 11, 12, 21]; see also [2, 10] and the references therein for discrete-time results.

2.2. Nearly completely decomposable Markov chain $\boldsymbol{\alpha_n^\varepsilon}$. In view of (2.1), the transition probabilities of α_n^ε are dominated by $\widetilde{P}$. The structure of $\widetilde{P}$ is thus important. Since α_n^ε is a finite-state Markov chain, the Markov chain corresponding to the transition matrix $\widetilde{P}$ either consists of all recurrent states or includes transient states in addition to recurrent states (see [15]). We first consider the case of inclusion of recurrent states only. Later we will discuss a generalization to the case in which transient states are also included. Suppose that the matrix $\widetilde{P}$ is given by

$$\widetilde{P} = \operatorname{diag}(\widetilde{P}^1, \dots, \widetilde{P}^l) = \begin{pmatrix} \widetilde{P}^1 & & \\ & \ddots & \\ & & \widetilde{P}^l \end{pmatrix}, \tag{2.3}$$

where each $\widetilde{P}^i \in \mathbb{R}^{m_i \times m_i}$ is itself a transition matrix and $\sum_{i=1}^l m_i = m$. Here and henceforth, by $\operatorname{diag}(Z^1, \dots, Z^l)$, we mean a diagonal block matrix with matrix entries Z^1 through Z^l of appropriate dimensions. It is clear that, for sufficiently small $\varepsilon > 0$, P^ε is close to $\widetilde{P}$, and so P^ε is a nearly completely decomposable transition matrix (see [7]). Note that, typically for large scale Markovian systems, $l \ll m$, and therein lies the motivation for reducing computational complexity. Concerning the Markov chain, we assume the following condition.

(A1) The transition probability matrix of the Markov chain α_n^ε is given by (2.1) with $\widetilde{P}$ specified in (2.3), and the state space of the Markov chain is

$$\begin{aligned} \mathcal{M} &= \mathcal{M}_1 \cup \mathcal{M}_2 \cap \cdots \cup \mathcal{M}_l \\ &= \{s_{11}, \dots, s_{1m_1}\} \cup \cdots \cup \{s_{l1}, \dots, s_{lm_l}\}. \end{aligned} \tag{2.4}$$

For each $i = 1, \dots, l$, $\mathcal{M}_i = \{s_{i1}, \dots, s_{im_i}\}$ is the state space corresponding to the transition matrix $\widetilde{P}^i$, and $\widetilde{P}^i$ is irreducible and aperiodic.

Note that the probability vector

$$p_n^\varepsilon = (P(\alpha_n^\varepsilon = s_{11}), \dots, P(\alpha_n^\varepsilon = s_{1m_1}), \dots, P(\alpha_n^\varepsilon = s_{l1}), \dots, P(\alpha_n^\varepsilon = s_{lm_l})) \in \mathbb{R}^{1\times m}$$

satisfies

$$p_{n+1}^\varepsilon = p_n^\varepsilon P^\varepsilon, \qquad p_0^\varepsilon = p_0, \tag{2.5}$$

such that p_0 is the initial probability distribution. By (A1), the result in [30] yields the following lemma.

LEMMA 2.1. *Assume condition* (A1). *Then the following assertions hold:*

(1) *Denote by* ν^i *the stationary distribution corresponding to the transition matrix* $\widetilde{P}_i$ *for each* $i = 1, \dots, l$. *Then, for some* $0 < \lambda < 1$,

$$p_n^\varepsilon = \theta(t)\,\mathrm{diag}(\nu^1, \dots, \nu^l) + O(\varepsilon + \lambda^n), \tag{2.6}$$

where $\theta(t) = (\theta_1(t), \dots, \theta_l(t)) \in \mathbb{R}^{1\times l}$ *(with* $t = \varepsilon n$*) satisfies*

$$\frac{d\theta(t)}{dt} = \theta(t)\overline{Q}, \quad \theta_i(0) = x_0^i \mathbf{1}_{m_i},$$

with

$$\begin{aligned} \overline{Q} &= \mathrm{diag}(\nu^1, \dots, \nu^l) Q \widetilde{\mathbf{1}}, \\ \widetilde{\mathbf{1}} &= \mathrm{diag}(\mathbf{1}_{m_1}, \dots, \mathbf{1}_{m_l}), \end{aligned} \tag{2.7}$$

where $\mathbf{1}_\ell$ *denotes an* ℓ*-dimensional column vector with all entries being* 1.

(2) *For* $n \le T/\varepsilon$, *the n-step transition probability matrix* $(P^\varepsilon)^n$ *satisfies*

$$(P^\varepsilon)^n = \Phi(t) + O\left(\varepsilon + \lambda^n\right), \tag{2.8}$$

where

$$\begin{aligned} P^0\Phi(t) &= \widetilde{\mathbf{1}}\Theta(t)\,\mathrm{diag}(\nu^1, \dots, \nu^l), \\ \frac{d\Theta(t)}{dt} &= \Theta(t)\overline{Q}, \quad \Theta(0) = I. \end{aligned} \tag{2.9}$$

Remark 2.2. Since we are primarily concerned with the form of the limit distribution, only the leading terms are presented in the lemma, although a full asymptotic expansion can be obtained. See [30] for more details.

Starting from the Markov chain α_n^ε, define an aggregated process $\overline{\alpha}_n^\varepsilon$ by setting $\overline{\alpha}_n^\varepsilon = i$ if $\alpha_n^\varepsilon \in \mathcal{M}_i$. Define piecewise constant interpolated processes $\alpha^\varepsilon(\cdot)$ and $\overline{\alpha}^\varepsilon(\cdot)$ by

$$\alpha^\varepsilon(t) = \alpha_n, \ \overline{\alpha}^\varepsilon(t) = \overline{\alpha}_n^\varepsilon, \quad t \in [n\varepsilon, n\varepsilon + \varepsilon).$$

Lemma 2.1 is mainly deterministic, whereas the following lemma is a weak convergence result on the aggregated process. Its proof is provided in [32]; a continuous-time counterpart can be found in [29, pp. 170–171].

LEMMA 2.3. *Under* (A1), *as* $\varepsilon \to 0$, $\overline{\alpha}^{\varepsilon}(\cdot)$ *converges weakly to* $\overline{\alpha}(\cdot)$, *which is a continuous-time Markov chain with state space* $\overline{\mathcal{M}} = \{1, \ldots, l\}$ *and generator* $\overline{Q}$ *given by* (2.7). *Moreover, for the occupation measures defined by*

$$o^{\varepsilon}_{n,ij} = \varepsilon \sum_{k=0}^{n} [I_{\{\alpha^{\varepsilon}_k = s_{ij}\}} - \nu^i_j I_{\{\alpha^{\varepsilon}_k \in \mathcal{M}_i\}}] \quad \textit{for each} \quad i = 1, \ldots, l, \ j = 1, \ldots, m_i,$$

the following mean square estimates hold:

$$\sup_{0 \le n \le T/\varepsilon} E|o^{\varepsilon}_{n,ij}|^2 = O(\varepsilon). \tag{2.10}$$

To proceed, we give additional conditions needed for the filtering problem.

(A2) $E|x_0|^2 < \infty$ and $E|y_0|^2 < \infty$. For each $\iota \in \mathcal{M}$, $A(\iota)$, $C(\iota)$, $\sigma_w(\iota)$, and $\sigma_v(\iota)$ are finite; $\sigma_w(\iota)\sigma'_w(\iota)$ and $\sigma_v(\iota)\sigma'_v(\iota)$ are positive definite matrices.

(A3) The sequences $\{w_n\}$ and $\{v_n\}$ are independent of $\{\alpha^{\varepsilon}_n\}$ and independent of each other. The $\{w_n\}$ and $\{v_n\}$ are stationary martingale difference sequences (with zero mean) such that

$$Ew_n w'_n = I, \ Ev_n v'_n = I,$$
$$E|w_n|^{2+\Delta} < \infty, \quad \text{and} \quad E|v_n|^{2+\Delta} < \infty \quad \text{for some} \quad \Delta > 0.$$

Remark 2.4. For simplicity and ease of presentation, we assume that the noises are stationary martingale difference sequences and that the covariance of w_n and v_n is the identity matrix. Even though no Gaussian assumption is used, as a result of the scaling, these noise processes will be asymptotically normal thanks to the functional central limit theorem.

In what follows, we use the weak convergence method to establish the desired results. Further details on the weak convergence method, which is an extension of convergence in distribution, can be found in, for example, [13, Chapter 3] or [19, Chapters 7 and 8].

3. Limit filtering problem. This section is devoted to the derivation of the limit filtering problem. In lieu of treating the discrete-time iterates, our analysis focuses on suitable continuous-time interpolations of piecewise constant processes.

For $0 \le n \le T/\varepsilon$, define the interpolations $x^{\varepsilon}(\cdot)$ and $y^{\varepsilon}(\cdot)$ as

$$x^{\varepsilon}(t) = x^{\varepsilon}_n, \quad y^{\varepsilon}(t) = y^{\varepsilon}_n, \quad t \in [n\varepsilon, n\varepsilon + \varepsilon), \tag{3.1}$$

where x^{ε}_n and y^{ε}_n are given in (2.2). Then $x^{\varepsilon}(\cdot)$ and $y^{\varepsilon}(\cdot) \in D^r[0,T]$, which is the space of $\mathbb{R}^r$-valued functions that are right continuous and have left limits, endowed with the Skorohod topology [13, p. 122]. Using weak convergence methods, we will show that the interpolated processes converge weakly to $x(\cdot)$ and $y(\cdot)$, which satisfy continuous-time hybrid Kalman filtering equations. Following the approach of weak convergence methods [13, 17], we first show that the sequences of interests are tight, and then we characterize the limit processes by using martingale averaging techniques.

Owing to the assumption on the system and observation noise and $\sqrt{\varepsilon}$ scaling, the following lemma, known as the functional central limit theorem or Donsker's invariance theorem, holds. Its proof is standard; see, for example, [13, Theorem 3.1, p. 351].

LEMMA 3.1. *Define*

$$w^\varepsilon(t) = \sqrt{\varepsilon} \sum_{j=0}^{t/\varepsilon-1} w_j \quad \text{and} \quad v^\varepsilon(t) = \sqrt{\varepsilon} \sum_{j=0}^{t/\varepsilon-1} v_j. \tag{3.2}$$

Under (A3), $w^\varepsilon(\cdot)$ *and* $v^\varepsilon(\cdot)$ *converge weakly to standard* r*-dimensional Brownian motions* $w(\cdot)$ *and* $v(\cdot)$*, respectively.*

In fact, correlated φ-mixing noises may be dealt with, and the corresponding central limit result can be obtained, but the notation will be much more complex for the subsequent averaging. Thus we decide to work with the martingale difference sequences $\{w_n\}$ and $\{v_n\}$. In the analysis to follow, we need the a priori bounds on $\{x_n^\varepsilon\}$ and $\{y_n^\varepsilon\}$, which are presented in the form of the following lemma. The proof is provided in the appendix.

LEMMA 3.2. *Assume* (A1)–(A3). *For* $\{x_n^\varepsilon\}$ *and* $\{y_n^\varepsilon\}$ *defined in* (2.2)*, the following bounds hold:*

$$\sup_{0\le n\le T/\varepsilon} E|x_n^\varepsilon|^2 < \infty \quad \text{and} \quad \sup_{0\le n\le T/\varepsilon} E|y_n^\varepsilon|^2 < \infty. \tag{3.3}$$

3.1. Tightness and weak convergence. To proceed, let $\mathcal{F}_n$ be the σ-algebra generated by $\{\alpha_j^\varepsilon, w_j, v_j : j \le n\}$, and let E_n be the conditional expectation with respect to $\mathcal{F}_n$; let $\mathcal{F}_t^\varepsilon$ be the σ-algebra generated by $\{\alpha^\varepsilon(s), w^\varepsilon(s), v^\varepsilon(s) : s \le t\}$, and let E_t^ε be the conditional expectation with respect to $\mathcal{F}_t^\varepsilon$. We are to derive the tightness of $\{x^\varepsilon(\cdot)\}$ and $\{y^\varepsilon(\cdot)\}$. This is a compactness result, which is established by verifying a tightness criterion; the proof is in the appendix.

THEOREM 3.3. *Assume* (A1)–(A3). *Then* $\{x^\varepsilon(\cdot)\}$ *is tight in* $D^r[0,T]$*, and so is* $\{y^\varepsilon(\cdot)\}$*, where* $D^r[0,T]$ *is the space of* $\mathbb{R}^r$*-valued functions that are right continuous and have left limits, endowed with the Skorohod topology.*

We are now in a position to obtain the weak convergence of the sequences $\{x^\varepsilon(\cdot)\}$ and $\{y^\varepsilon(\cdot)\}$. To prove the assertion, we use a martingale problem formulation. Thus our task becomes to figure out the limit by characterizing the operator of the limit martingale problem. The technique used is essentially an averaging approach. Different from the diffusion approximation in wideband noise systems [17], the limit $\overline{\alpha}^\varepsilon(\cdot)$ also contributes to the limit process and adds further complication. The result is recorded in the following theorem, whose proof is in the appendix as well.

THEOREM 3.4. *Suppose the conditions of Theorem* 3.3 *hold. Then* $x^\varepsilon(\cdot)$ *and* $y^\varepsilon(\cdot)$ *converge weakly to* $x(\cdot)$ *and* $y(\cdot)$*, respectively, such that* $x(\cdot)$ *and* $y(\cdot)$ *are solutions of the filtering equations*

$$\begin{aligned} dx &= \overline{A}(\overline{\alpha}(t))x\,dt + \overline{\sigma_w}(\overline{\alpha}(t))dw, \\ dy &= \overline{C}(\overline{\alpha}(t))x\,dt + \overline{\sigma_v}(\overline{\alpha}(t))dv, \end{aligned} \tag{3.4}$$

where $w(\cdot)$ *and* $v(\cdot)$ *are the independent* r*-dimensional standard Brownian motions given by Lemma* 3.1*,*

$$\overline{A}(i) = \sum_{j=1}^{m_i} \nu_j^i A(s_{ij}),\ \overline{B}(i) = \sum_{j=1}^{m_i} \nu_j^i B(s_{ij}) \quad \text{for each } i \in \overline{\mathcal{M}}, \tag{3.5}$$

and, for each $i \in \overline{\mathcal{M}}$, $\overline{\sigma}_w(i)$ *and* $\overline{\sigma}_v(i)$ *satisfy*

$$\overline{\sigma}_w(i)\overline{\sigma}'_w(i) = \sum_{j=1}^{m_i} \nu_j^i \sigma_w(s_{ij})\sigma'_w(s_{ij}), \tag{3.6}$$
$$\overline{\sigma}_v(i)\overline{\sigma}'_v(i) = \sum_{j=1}^{m_i} \nu_j^i \sigma_v(s_{ij})\sigma'_v(s_{ij}).$$

3.2. Markov chains with one ergodic class. The reduction of complexity is particularly pronounced if the transition matrix (2.3) consists of only one ergodic class (i.e., P in (2.3) consists of only one block). That is, $P^\varepsilon = P + \varepsilon Q$ such that P is irreducible and aperiodic. It is easily seen that, for sufficiently small $\varepsilon > 0$, P^ε is also irreducible. Consider the filtering problem (2.2). Similarly to the previous case, define $x^\varepsilon(\cdot)$ and $y^\varepsilon(\cdot)$ as the piecewise constant interpolations of x_k^ε and y_k^ε, respectively. Replace $\overline{A}(\cdot)$ and $\overline{\sigma}_w(\cdot)$ by

$$\overline{A}^0 = \sum_{j=1}^{m} A(j)\nu_j \quad \text{and} \quad \overline{\sigma_w}^0(\overline{\sigma_w}^0)' = \sum_{j=1}^{m} \nu_j \sigma_w(j)\sigma'_w(j), \tag{3.7}$$

with $\nu = (\nu_1, \ldots, \nu_m)$ denoting the stationary distribution of P. Similarly replace $\overline{C}(\cdot)$ and $\overline{\sigma}_v(\cdot)$ by $\overline{C}^0$ and $\overline{\sigma}_v^0$, respectively. The weak convergence of $(x^\varepsilon(\cdot), y^\varepsilon(\cdot))$ will still be obtained. The proofs are similar to the previous case. In fact, it is readily seen that Lemma 3.2 and Theorem 3.3 continue to hold. Lemma 2.1 still holds with obvious modifications, and (2.10) (in Lemma 2.3) is changed to

$$\sup_{0 \le n \le T/\varepsilon} E\left[\varepsilon \sum_{k=0}^{n} [I_{\{\alpha_k^\varepsilon = j\}} - \nu_j]\right]^2 = O(\varepsilon).$$

Using this mean square estimate and similar arguments as before, we can show that Theorem 3.4 continues to hold. It is interesting to note that the limit filtering problem becomes a standard Kalman filter, in which the jump process effect has been completely averaged out. We state this as the following result.

COROLLARY 3.5. *Consider the filtering problem* (2.2) *such that* P *is irreducible and aperiodic. Then* $(x^\varepsilon(\cdot), y^\varepsilon(\cdot))$ *converges weakly to* $(x(\cdot), y(\cdot))$*, that is, the solution of the filtering problem*

$$dx(t) = \overline{A}^0 x(t)dt + \overline{\sigma_w}^0 dw(t), \tag{3.8}$$
$$dy(t) = \overline{C}^0 x(t)dt + \overline{\sigma_v}^0 dv(t).$$

4. Simulation studies. In this section, we demonstrate the relationship between the full-order discrete-time system (2.2) and the reduced-order limit filtering equations (3.4) through simulation examples. All results are averaged over 50 trials.

For (2.2), we simulate a discrete-time Markov chain with four states (with two blocks, $m_1 = m_2 = 2$) for which the transition probability matrices are

$$\widetilde{P}^1 = \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix}, \ \widetilde{P}^2 = \begin{pmatrix} 0.25 & 0.75 \\ 0.68 & 0.32 \end{pmatrix}$$

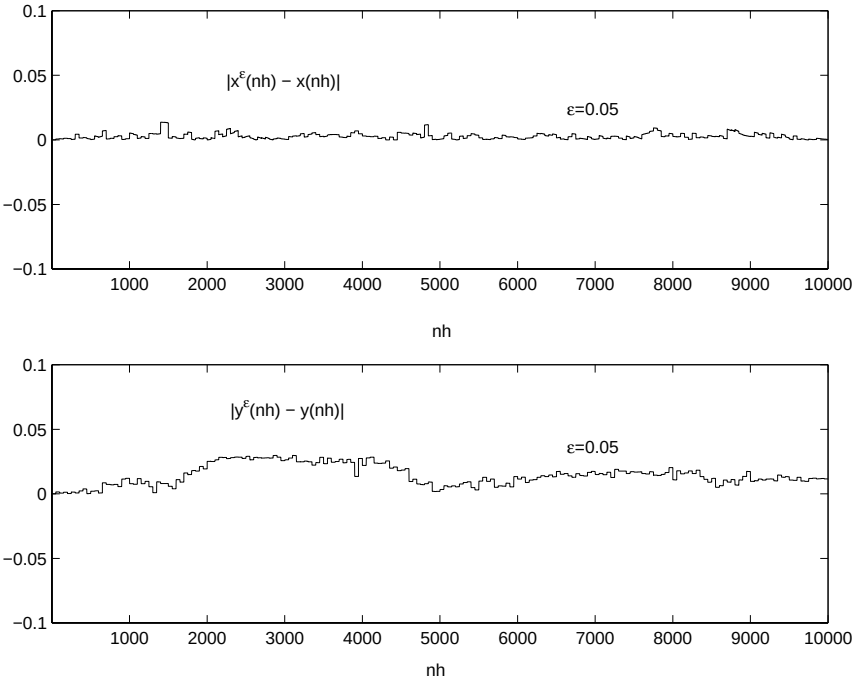


FIG. 4.1. *Absolute error between the piecewise constant interpolated full-order system and the reduced-order limit filtered system,* $\varepsilon = 0.05$.

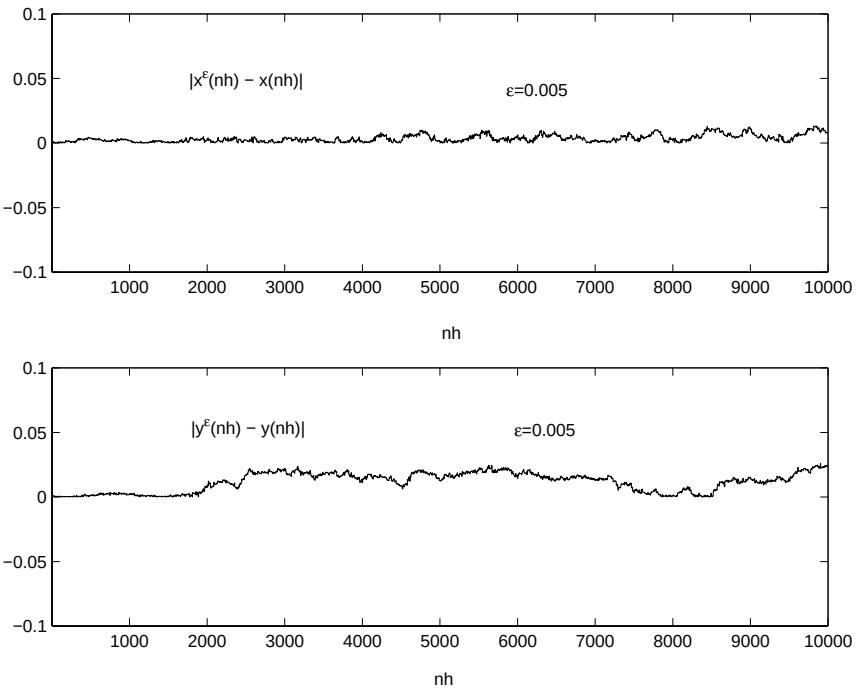


FIG. 4.2. *Absolute error between the piecewise constant interpolated full-order system and the reduced-order limit filtered system,* $\varepsilon = 0.005$.

and the generator is

$$Q = \begin{pmatrix} -0.6 & 0.4 & 0.1 & 0.1 \\ 0.05 & -0.4 & 0.05 & 0.3 \\ 0.1 & 0.2 & -0.7 & 0.4 \\ 0.15 & 0.05 & 0.1 & -0.3 \end{pmatrix}.$$

We take $(A(1)\ A(2)\ A(3)\ A(4)) = (-4.0\ -1.0\ -2.0\ -3.0)$, $(C(1)\ C(2)\ C(3)\ C(4)) = (0.2\ 0.5\ 0.1\ 1.0)$. Also, $(\sigma_w(1)\ \sigma_w(2)\ \sigma_w(3)\ \sigma_w(4)) = (0.2\ 0.5\ 0.1\ 1.0)$, $\sigma_v(i) = \sigma_w(i)$, for $i = 1, 2, 3, 4$. The noise sequences $\{w_n\}$ and $\{v_n\}$ are simulated as Gaussian random variables with zero mean and unity variance. The piecewise constant interpolated processes $x^\varepsilon(t)$ and $y^\varepsilon(t)$ are constructed from (3.1). The time horizon is taken to be $T = 10$. To simulate (3.4), a continuous-time Markov chain is used with a generator $\overline{Q}$. This is then discretized with a discretization interval $h = 0.001$. Figures 4.1 and 4.2 show the difference $|x^\varepsilon(\cdot) - x(\cdot)|$ and $|y^\varepsilon(\cdot) - y(\cdot)|$ for $\varepsilon = 0.05$ and $\varepsilon = 0.005$, respectively.

5. Remarks and extensions. This section is devoted to several remarks regarding the approximation issue. They include reduction of complexity as well as ramifications of the results we have obtained thus far.

Reduction of complexity. One of the main motivations of the current study is the effort of reduction of complexity. Regarding (2.2), note that the time horizon we are working with is $0 \le n \le \lfloor T/\varepsilon \rfloor$. As pointed out in [24], if we treat the discrete-time case directly, it can be reduced to an $m^{\lfloor T/\varepsilon \rfloor}$-dimensional recursive system of equations, where m is the total number of states of the Markov chain. For us, m is a fairly large number. As a result, the amount of computation becomes practically untrackable. One cannot complete the computation in polynomial time. By weak convergence methods, we have obtained a reduced or limit system of filtering equations. This limit system of equations allows us to find nearly optimal filtering, and the limit system has reduced complexity. In particular, if the transition matrix P given in (2.1) is irreducible, the limit becomes a Kalman filter (see Proposition 3.5).

For continuous-time Kalman filter problems with Markovian switching, it has been recognized (see [4, 11, 21]) that, in general, the problem is an infinite-dimensional one just as in the nonlinear filter case [20]. Nevertheless, Björk [4] proved that a finite-dimensional filter exists for a linear hybrid system if and only if the observation is independent of the state variable. For the filtering problem considered in this paper, this requires the observation process in the limit problem being independent of state. Corresponding to such a requirement, we can consider

$$\begin{aligned} x^\varepsilon_{n+1} &= x^\varepsilon_n + \varepsilon A(\alpha^\varepsilon_n) x^\varepsilon_n + \sqrt{\varepsilon}\sigma_w(\alpha^\varepsilon_n) w_n, \\ y^\varepsilon_{n+1} &= y^\varepsilon_n + \varepsilon C(\alpha^\varepsilon_n) + \sqrt{\varepsilon}\sigma_v v_n. \end{aligned} \tag{5.1}$$

Similar to the derivation of Theorem 3.4, we obtain the limit filtering equations

$$\begin{aligned} dx &= \overline{A}(\overline{\alpha}(t))x\,dt + \overline{\sigma_w}(\overline{\alpha}(t))dw, \\ dy &= \overline{C}(\overline{\alpha}(t))dt + \sigma_v dv. \end{aligned} \tag{5.2}$$

Note that the calculation of (5.1) leads to recursive filters of dimension $m^{\lfloor T/\varepsilon \rfloor}$, whereas (5.2) yields a finite-dimensional filtering problem.

Inclusion of transient states. In the previous sections, the main ingredient is the aggregation of states in each recurrent classes. The results obtained can be extended to the case in which the Markov chain has finite state space with inclusion of transient states. To be more specific, Let the transition probability be of the form (2.1). However, in lieu of (2.3), suppose the transition matrix $\widetilde{P}$ in (2.1) is given by

$$\widetilde{P} = \begin{pmatrix} \widetilde{P}^1 & & & \\ & \ddots & & \\ & & \widetilde{P}^l & \\ \widetilde{P}^1_* & \cdots & \widetilde{P}^l_* & \widetilde{P}_* \end{pmatrix}. \tag{5.3}$$

In lieu of (A1), assume (A1').

(A1') α^ε_n is a Markov chain with a transition probability matrix given by (2.1) and (5.3), and with state space

$$\begin{aligned} \mathcal{M} &= \mathcal{M}_1 \cup \mathcal{M}_2 \cap \cdots \cup \mathcal{M}_l \cup \mathcal{M}_* \\ &= \{s_{11}, \ldots, s_{1m_1}\} \cup \cdots \cup \{s_{l1}, \ldots, s_{lm_l}\} \cup \{s_{*1}, \ldots, s_{*m_*}\}, \end{aligned} \tag{5.4}$$

where, for each $i = 1, \ldots, l$, $\mathcal{M}_i = \{s_{i1}, \ldots, s_{im_i}\}$ is the state space corresponding to the transition matrix $\widetilde{P}^i$ and where the subspace $\mathcal{M}_* = \{s_{*1}, \ldots, s_{*m_*}\}$ collects the transient states. Moreover, $\widetilde{P}^i$ is irreducible for each $i = 1, \ldots, l$, and all eigenvalues of $\widetilde{P}_*$ are inside the unit disk.

To obtain the desired asymptotics, we still use aggregations. However, we aggregate only the states in each recurrent class. Partition the matrix Q as

$$Q = \begin{pmatrix} Q^{11} & Q^{12} \\ Q^{21} & Q^{22} \end{pmatrix}, \tag{5.5}$$

where

$$\begin{aligned} &Q^{11} \in \mathbb{R}^{(m-m_*)\times(m-m_*)},\ Q^{12} \in \mathbb{R}^{(m-m_*)\times m_*}, \\ &Q^{21} \in \mathbb{R}^{m_*\times(m-m_*)}, \text{ and } Q^{22} \in \mathbb{R}^{m_*\times m_*}. \end{aligned}$$

Set

$$\overline{Q}_* = \operatorname{diag}(\nu^1, \ldots, \nu^l)(Q^{11}\widetilde{\mathbf{1}} + Q^{12}A_*), \tag{5.6}$$

with

$$\begin{aligned} &A_* = (a_1, \ldots, a_l) \in \mathbb{R}^{m_*\times l} \quad \text{and} \\ &a_i = -(\widetilde{P}_* - I)^{-1}\widetilde{P}^i_*\mathbf{1}_{m_i} \text{ for } i = 1, \ldots, l. \end{aligned} \tag{5.7}$$

Let U be a random variable uniformly distributed over $[0,1]$. For each $j = 1, \ldots, m_*$, define an integer-valued random variable ξ_j by

$$\xi_j = I_{\{0\le U\le a_{m_1,j}\}} + 2I_{\{a_{1,j}<U\le a_{1,j}+a_{2,j}\}} + \cdots + lI_{\{a_{1,j}+\cdots+a_{l-1,j}<U\le 1\}}.$$

Define the aggregated process and its interpolation by

$$\begin{aligned} \overline{\alpha}^\varepsilon_n &= \begin{cases} i & \text{if } \alpha_n \in \mathcal{M}_i, \\ U_j & \text{if } \alpha^\varepsilon_n = s_{*j}, \end{cases} \\ \overline{\alpha}^\varepsilon(t) &= \overline{\alpha}^\varepsilon_n \quad \text{for} \quad t \in [n\varepsilon, n\varepsilon+\varepsilon). \end{aligned} \tag{5.8}$$

Then we can show that $\overline{\alpha}^{\varepsilon}(\cdot)$ converges weakly to $\overline{\alpha}(\cdot)$ and that the limit is still a Markov chain with state space $\overline{\mathcal{M}}$. Furthermore, we can obtain similar limit results for the filtering problems. The notation is more involved, but the main idea and the averaging techniques are as in the previous case. Loosely, the transient states are asymptoticly negligible. In the limit (reduced) system, only the states in the recurrent states are important. The limit is still an average with respect to the stationary measures of each recurrent class.

THEOREM 5.1. *Assume* (A1'), (A2), *and* (A3). *Then the conclusions of Theorems* 3.3 *and* 3.4 *continue to hold with* $\overline{Q}$ *replaced by* $\overline{Q}_*$ *defined in* (5.6).

Nonstationary Markov chains. Generally, nonstationary or time-inhomogeneous cases are much more difficult to deal with. However, for a class of problems, it can be worked out. The main setup is similar to that of [30]. In lieu of (2.1), assume that the transition probability matrix is nonstationary and given by

$$P^{\varepsilon}(\varepsilon n) = \widetilde{P}(\varepsilon n) + \varepsilon Q(\varepsilon n),$$

where $\widetilde{P}(\varepsilon n)$ is the dominating part of the transition matrix. In this case, we can carry out the analysis as in the previous case, although the details and notation are more involved.

Continuous-time problems. So far, we have considered discrete-time filtering problems exclusively. There is also a continuous-time analogue of the hybrid filtering problems. In place of (2.2), for $t \in [0, T]$, consider

$$\begin{aligned} dx^{\varepsilon}(t) &= A(\alpha^{\varepsilon}(t))x^{\varepsilon}(t)dt + \sigma_w(\alpha^{\varepsilon}(t))dw, \\ dy^{\varepsilon}(t) &= C(\alpha^{\varepsilon}(t))x^{\varepsilon}(t)dt + \sigma_v(\alpha^{\varepsilon}(t))dv, \end{aligned} \tag{5.9}$$

where $w(\cdot)$ and $v(\cdot)$ are independent standard Brownian motions, and where $\alpha^{\varepsilon}(\cdot)$ is a continuous-time singularly perturbed Markov chain with finite state space $\mathcal{M}$ and with generator

$$Q^{\varepsilon}(t) = \frac{\widetilde{Q}(t)}{\varepsilon} + \widehat{Q}(t), \tag{5.10}$$

where both $\widetilde{Q}(t)$ and $\widehat{Q}(t)$ are generators. The state space $\mathcal{M}$ can be of the form of either (2.4) (with recurrent states only) or (5.4) (inclusion of transient states). We can follow our approach of averaging and aggregation to reduce the complexity of the underlying system and obtain a limit system with much reduced state space. Various results on the asymptotic properties of $\alpha^{\varepsilon}(\cdot)$ can be found in [29, 31] among others. The proof of the following result is similar to the discrete-time case; we omit the details. Note that, since the problem is in continuous time, no interpolations are needed, however. For definiteness, we state the result for decomposition of the form (5.4). The matrix $\widetilde{Q}(t)$ has the form

$$\widetilde{Q}(t) = \begin{pmatrix} \widetilde{Q}^1(t) & & & \\ & \ddots & & \\ & & \widetilde{Q}^l(t) & \\ \widetilde{Q}^1_*(t) & \cdots & \widetilde{Q}^l_*(t) & \widetilde{Q}_*(t) \end{pmatrix}. \tag{5.11}$$

For each $i \in \{1, \ldots, l\}$, let $\widetilde{Q}^i_*(t) = B(t)\widetilde{Q}^i_{*,c}$, $\widetilde{Q}_*(t) = B(t)\widetilde{Q}_{*,c}$, where $B(t)$ is an $\mathbb{R}^{m_* \times m_*}$ matrix-valued function, and $\widetilde{Q}^i_{*,c} \in \mathbb{R}^{m_* \times m_i}$ and $\widetilde{Q}_{*,c} \in \mathbb{R}^{m_* \times m_*}$ are

constant matrices. It is readily seen that $B(t)$ is invertible for each $t \in [0,T]$, and, for each i,

$$a_i(t) \stackrel{\text{def}}{=} -\widetilde{Q}_*^{-1}(t)\widetilde{Q}_*^i(t)\mathbf{1}_{m_i} = -\widetilde{Q}_{*,c}^{-1}\widetilde{Q}_{*,c}^i\mathbf{1}_{m_i} = a_i \tag{5.12}$$

is a time-independent vector. Define $\overline{Q}_*(t)$ and $\overline{\alpha}^\varepsilon(t)$ as in (5.6) and (5.8) with $Q^{\iota\ell}$ (the partition of Q) replaced by $\widehat{Q}^{\iota\ell}(t)$ (the partition of $\widehat{Q}(t)$) and with a_i in (5.7) replaced by (5.12). Then we have the following theorem and its corollary.

THEOREM 5.2. *Suppose that, for each $i \in \{1,\dots,l\}$, $\widetilde{Q}^i(t)$ is weakly irreducible (see* [29, pp. 21–22] *for a definition of weak irreducibility and quasi-stationary distribution), that $\widetilde{Q}_*(t)$ has all of its eigenvalues on the left half of the complex plan, that $\widetilde{Q}(\cdot)$ and $\widehat{Q}(\cdot)$ are bounded and Borel measurable, and that $\widetilde{Q}(\cdot)$ is Lipschitz continuous on $[0,T]$. Then $(x^\varepsilon(\cdot), y^\varepsilon(\cdot))$ converges weakly to $(x(\cdot), y(\cdot))$ such that $(x(\cdot), y(\cdot))$ is a solution of the averaged filtering equations* (3.4), *where $\overline{A}(\cdot)$, $\overline{\sigma}_w(\cdot)$, $\overline{C}(\cdot)$, and $\overline{\sigma}_v(\cdot)$ are defined as before with time-dependent quasi-stationary distributions $\nu^i(t)$ used.*

COROLLARY 5.3. *Suppose that $Q^\varepsilon(t)$ is given by* (5.10) *such that $\widetilde{Q}(t)$ is weakly irreducible. Suppose that all other conditions in Theorem* 5.2 *are satisfied. Then $(x^\varepsilon(\cdot), y^\varepsilon(\cdot))$ converges weakly to $(x(\cdot), y(\cdot))$, satisfying*

$$\begin{aligned} dx(t) &= \overline{A}^0(t)x(t)dt + \overline{\sigma_w}^0(t)dw(t), \\ dy(t) &= \overline{C}^0(t)x(t)dt + \overline{\sigma_v}^0(t)dv(t), \end{aligned} \tag{5.13}$$

where $\overline{A}^0$, $\overline{C}^0$, $\overline{\sigma_w}^0$, and $\overline{\sigma_v}^0$ are defined as in (3.7) *with the time-dependent quasi-stationary distribution $\nu(t) = (\nu_1(t),\dots,\nu_m(t))$ used.*

Appendix. Proofs of results.

Proof of Lemma 3.2. We first work with x_n^ε. Iterating on the first equation in (2.2), for $0 \le n \le T/\varepsilon$,

$$x_{n+1}^\varepsilon = x_0^\varepsilon + \varepsilon\sum_{j=0}^n A(\alpha_j^\varepsilon)x_j^\varepsilon + \sqrt{\varepsilon}\sum_{j=0}^n \sigma_w(\alpha_j^\varepsilon)w_j.$$

Note that, for any $z \in \mathbb{R}^r$, $|z|^2 = \operatorname{tr}(zz')$, where $\operatorname{tr}(zz')$ denotes the trace of zz'. Consequently (recall that K is a generic positive constant),

$$\begin{aligned} E|x_{n+1}^\varepsilon|^2 &\le K\left(E|x_0^\varepsilon|^2 + \varepsilon^2 E\left|\sum_{j=0}^n A(\alpha_j^\varepsilon)x_j^\varepsilon\right|^2 + \varepsilon E\left|\sum_{j=0}^n \sigma_w(\alpha_j^\varepsilon)w_j\right|^2\right) \\ &\le KE|x_0^\varepsilon|^2 + K\varepsilon\sum_{j=0}^n E|x_j^\varepsilon|^2 + \varepsilon K\sum_{j=0}^n\sum_{k=0}^n E\operatorname{tr}\left(\sigma_w(\alpha_j^\varepsilon)w_j w_k'\sigma_w'(\alpha_k^\varepsilon)\right). \end{aligned} \tag{A.1}$$

Using the independence of $\{\alpha_n^\varepsilon\}$ and $\{w_n\}$ and the boundedness of $\sigma_w(\iota)$ for each

$\iota \in \mathcal{M}$ and noting that $Ew_j w_k' = 0$ if $j \neq k$,

$$\begin{aligned}
&\varepsilon \sum_{j=0}^{n}\sum_{k=0}^{n} \operatorname{tr}\left(E\sigma_w(\alpha_j^\varepsilon) w_j w_k' \sigma_w'(\alpha_k^\varepsilon)\right) \\
&\le \varepsilon K \left| \sum_{j=0}^{n}\sum_{k=0}^{n} E\{\sigma_w(\alpha_j^\varepsilon)[Ew_j w_k']\sigma_w'(\alpha_k^\varepsilon)\}\right| \\
&\le \varepsilon K \sum_{k=0}^{n} |Ew_k w_k'| \\
&\le \varepsilon K \frac{T}{\varepsilon} \le K < \infty.
\end{aligned} \tag{A.2}$$

Combining this with (A.1), an application of Gronwall's inequality yields

$$E|x_{n+1}^\varepsilon|^2 \le K + K\varepsilon \sum_{j=0}^{n} E|x_j^\varepsilon|^2 \le K\exp(K\varepsilon n) \le K < \infty.$$

Moreover, the bound holds uniformly in n for $0 \le n \le T/\varepsilon$.

As for y_n^ε, using the bound of $\sup_{0\le n\le T/\varepsilon} E|x_n^\varepsilon|^2$, we have

$$\begin{aligned}
E|y_{n+1}^\varepsilon|^2 &\le K\left(E|y_0^\varepsilon|^2 + \varepsilon^2 E\left|\sum_{j=0}^{n} C(\alpha_j^\varepsilon)x_j^\varepsilon\right|^2 + \varepsilon E\left|\sum_{j=0}^{n}\sigma_v(\alpha_j^\varepsilon)v_j\right|^2\right) \\
&\le KE|y_0^\varepsilon|^2 + K\varepsilon\sum_{j=0}^{n} E|x_j^\varepsilon|^2 + \varepsilon K\sum_{j=0}^{n}\sum_{k=0}^{n} E\operatorname{tr}\left(\sigma_v(\alpha_j^\varepsilon)v_j v_k'\sigma_v'(\alpha_k^\varepsilon)\right) \\
&\le K < \infty.
\end{aligned} \tag{A.3}$$

Moreover, the bound holds uniformly in $0 \le n \le T/\varepsilon$. □

Proof of Theorem 3.3. Let us first deal with the sequence $\{x^\varepsilon(\cdot)\}$. For any $\delta > 0$, $t > 0$, and $s > 0$ with $s \le \delta$, consider

$$\begin{aligned}
E_t^\varepsilon|x^\varepsilon(t+s) - x^\varepsilon(t)|^2 &= E_t^\varepsilon\left|\varepsilon\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1} A(\alpha_j^\varepsilon)x_j^\varepsilon + \sqrt{\varepsilon}\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1}\sigma_w(\alpha_j^\varepsilon)w_j\right|^2 \\
&= \varepsilon^2\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1}\sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} E_t^\varepsilon\operatorname{tr}[A(\alpha_j^\varepsilon)x_j^\varepsilon x_k^{\varepsilon\prime}A'(\alpha_k^\varepsilon)] \\
&\quad + 2\sqrt{\varepsilon^3}\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1}\sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} E_t^\varepsilon\operatorname{tr}[A(\alpha_j^\varepsilon)x_j^\varepsilon w_k'\sigma_w'(\alpha_k^\varepsilon)] \\
&\quad + \varepsilon\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1}\sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} E_t^\varepsilon\operatorname{tr}[\sigma_w(\alpha_j^\varepsilon)w_j w_k'\sigma_w'(\alpha_k^\varepsilon)]
\end{aligned} \tag{A.4}$$

$$\stackrel{\text{def}}{=} I_1^\varepsilon(t,s) + I_2^\varepsilon(t,s) + I_3^\varepsilon(t,s), \tag{A.5}$$

where $I_\ell^\varepsilon(t,s)$ for $\ell = 1,2,3$ are defined in an obvious manner.

Consider each of the terms on the right-hand side of (A.4) separately as follows. First, by the finiteness of $A(\iota)$ for each $\iota \in \mathcal{M}$,

$$I_1^\varepsilon(t,s) = \varepsilon^2 \sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1} \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} \operatorname{tr}\left(E_t^\varepsilon[A(\alpha_j^\varepsilon)x_j^\varepsilon x_k^{\varepsilon,\prime} A'(\alpha_k^\varepsilon)]\right)$$
$$\leq K\varepsilon^2 \sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1} \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} E_t^\varepsilon |x_j^\varepsilon||x_k^\varepsilon|.$$

By virtue of Lemma 3.2, an application of the Cauchy–Schwarz inequality then yields

$$EI_1^\varepsilon(t,s) \leq K\varepsilon^2 \sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1} \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} E^{1/2}|x_j^\varepsilon|^2 E^{1/2}|x_k^\varepsilon|^2$$
$$\leq K\varepsilon^2 \left(\frac{t+s}{\varepsilon} - \frac{t}{\varepsilon}\right)^2 \leq Ks^2 = O(\delta^2).$$

Thus

$$\lim_{\delta\to 0} \limsup_{\varepsilon\to 0} EI_1^\varepsilon(t,s) = \lim_{\delta\to 0} O(\delta^2) = 0. \tag{A.6}$$

As for the second term on the right-hand side of (A.4), note that x_j^ε and $A(\alpha_j^\varepsilon)$ are $\mathcal{F}_j$-measurable. Since, for $j < k$, $E_j w_k = 0$, the independence of $\{\alpha_n^\varepsilon\}$ and $\{w_n\}$ in (A3) and the finiteness of $A(\iota)$ and $\sigma_w(\iota)$ for each $\iota \in \mathcal{M}$ lead to

$$I_2^\varepsilon(t,s) = 2\sqrt{\varepsilon^3} \sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1} \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} \operatorname{tr} E_t^\varepsilon[A(\alpha_j^\varepsilon)x_j^\varepsilon w_k' \sigma_w'(\alpha_k^\varepsilon)]$$
$$\leq K\sqrt{\varepsilon^3} \sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1} \sum_{k\geq j} \left|\operatorname{tr}\left[E_t^\varepsilon A(\alpha_j^\varepsilon)x_j^\varepsilon (E_j w_k')(E_j \sigma_w'(\alpha_k^\varepsilon))\right]\right|$$
$$\leq K\sqrt{\varepsilon^3} \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} \sqrt{E_t^\varepsilon |x_k^\varepsilon|^2}\sqrt{E_t^\varepsilon |w_k|^2}.$$

Therefore, an application of the Cauchy–Schwarz inequality yields

$$\lim_{\delta\to 0} \limsup_{\varepsilon\to 0} EI_2^\varepsilon(t,s) = \lim_{\delta\to 0} \limsup_{\varepsilon\to 0} O(\sqrt{\varepsilon}) = 0. \tag{A.7}$$

Next, we consider the last term of (A.4). Using the martingale difference property, the independence of $\{\alpha_n^\varepsilon\}$ and $\{w_n\}$, and $E_j w_k = 0$ for $j < k$ and $E_k w_j = 0$ for $k < j$, we obtain

$$I_3^\varepsilon(t,s) = \varepsilon \sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1} \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} \operatorname{tr}\left(E_t^\varepsilon \sigma_w(\alpha_j^\varepsilon) w_j w_k' \sigma_w'(\alpha_k^\varepsilon)\}\right)$$
$$= \varepsilon \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} \left|\operatorname{tr}[E_t^\varepsilon \sigma_w(\alpha_k^\varepsilon) w_k w_k' \sigma_w'(\alpha_k^\varepsilon)]\right|,$$

and so

$$EI_3^\varepsilon(t,s) \le K\varepsilon\left(\frac{t+s}{\varepsilon} - \frac{t}{\varepsilon}\right) = O(\delta).$$

As a result,

$$\lim_{\delta\to 0}\limsup_{\varepsilon\to 0} EI_3^\varepsilon(t,s) = \lim_{\delta\to 0} O(\delta) = 0. \tag{A.8}$$

Combining (A.6), (A.7), and (A.8), we obtain

$$\lim_{\delta\to 0}\limsup_{\varepsilon\to 0} E|x^\varepsilon(t+s) - x^\varepsilon(t)|^2 = 0.$$

The criteria due to Kurtz [17, p. 47] then yields that $\{x^\varepsilon(\cdot)\}$ is tight in $D^r[0,T]$.

As far as the estimates of $y^\varepsilon(\cdot)$ are concerned, we merely note that

$$\begin{aligned} &E_t^\varepsilon|y^\varepsilon(t+s) - y^\varepsilon(t)|^2 \\ &\qquad = E_t^\varepsilon\left|\varepsilon\sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} C(\alpha_k^\varepsilon)x_k^\varepsilon + \sqrt{\varepsilon}\sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1}\sigma_v(\alpha_k^\varepsilon)v_k\right|^2 \\ &\qquad \le KE_t^\varepsilon\left|\varepsilon\sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} C(\alpha_k^\varepsilon)x_k^\varepsilon\right|^2 + KE_t^\varepsilon\left|\sqrt{\varepsilon}\sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1}\sigma_v(\alpha_k^\varepsilon)v_k\right|^2. \end{aligned}$$

The rest of the estimates are all similar to the previous case. Thus we also have that $\{y^\varepsilon(\cdot)\}$ is tight in $D^r[0,T]$. □

Proof of Theorem 3.4. Consider $\{x^\varepsilon(\cdot)\}$ first. In fact, we work with the pair $(x^\varepsilon(\cdot),\overline{\alpha}^\varepsilon(\cdot))$. Owing to the tightness of $\{x^\varepsilon(\cdot)\}$ and the weak convergence of $\{\overline{\alpha}^\varepsilon(\cdot)\}$, $\{(x^\varepsilon(\cdot),\overline{\alpha}^\varepsilon(\cdot))\}$ is tight. By virtue of the Prohorov theorem [13, p. 104], we can extract a weakly convergent subsequence. Select such a subsequence, and still denote it by $\{(x^\varepsilon(\cdot),\overline{\alpha}^\varepsilon(\cdot))\}$ for simplicity. Denote the limit of the sequence by $(x(\cdot),\overline{\alpha}(\cdot))$. By the Skorohod representation [13, p. 102], we may assume without loss of generality that $(x^\varepsilon(\cdot),\overline{\alpha}^\varepsilon(\cdot))$ converges to $(x(\cdot),\overline{\alpha}(\cdot))$ with probability one (w.p.1). Moreover, the convergence is uniform on each bounded time interval. We proceed to use martingale averaging techniques to figure out the limit.

To obtain the desired limit, it suffices to show that the limit $(x(\cdot),\overline{\alpha}(\cdot))$ is the solution of a martingale problem with operator $\mathcal{L}$ given by

$$\mathcal{L}f(x,i) = f_x'(x,i)\overline{A}(i)x + \frac{1}{2}\mathrm{tr}[f_{xx}(x,i)\overline{\sigma_w}(i)\overline{\sigma_w}'(i)] + \overline{Q}f(x,\cdot)(i), \quad i\in\overline{\mathcal{M}}, \tag{A.9}$$

where

$$\overline{Q}f(x,\cdot)(i) = \sum_{j\in\overline{\mathcal{M}}}\overline{q}_{ij}f(x,j) = \sum_{j\in\overline{\mathcal{M}},\ j\neq i}\overline{q}_{ij}(f(x,j)-f(x,i))$$

for each $i\in\overline{\mathcal{M}}$, and $f(\cdot,i)\in C_0^2$ (twice continuously differentiable function with compact support). Since the filtering equation is linear in the state variable, by using a similar argument to that in [29, Lemma 7.18], the corresponding martingale problem with operator $\mathcal{L}$ given in (A.9) has a unique solution.

To obtain the desired results, it suffices to show (see [19, Chapters 7 and 8]), for any positive integer k_0, any bounded and continuous function $h_\kappa(\cdot)$ with $\kappa \le k_0$, any $t, s > 0$, and $t_\kappa \le t \le t+s$, that the following equation holds:

$$E\prod_{\kappa=1}^{k_0} h_\kappa(x(t_\kappa),\overline{\alpha}(t_\kappa))\Bigg(f(x(t+s),\overline{\alpha}(t+s)) - f(x(t),\overline{\alpha}(t)) - \int_t^{t+s}\mathcal{L}f(x(u),\overline{\alpha}(u))du\Bigg) = 0. \tag{A.10}$$

To obtain (A.10), we begin with the pair $(x^\varepsilon(\cdot),\alpha^\varepsilon(\cdot))$. For each x, define $\check{f}$ by

$$\check{f}(x,\alpha) = \sum_{i=1}^{l} f(x,i)I_{\{\alpha\in\mathcal{M}_i\}} \quad \text{for each} \quad \alpha\in\mathcal{M}. \tag{A.11}$$

Note that, for each $\alpha = s_{ij} \in \mathcal{M}_i$, $\check{f}(x,\alpha)$ takes a constant value $f(x,i)$. Note also that, at any time instant t, $\alpha^\varepsilon(t) = \alpha^\varepsilon_{t/\varepsilon}$ takes on one of the m possible values from $\mathcal{M}$.

Note that $\check{f}(x^\varepsilon_k,\alpha^\varepsilon_k) = f(x^\varepsilon_k,\overline{\alpha}^\varepsilon_k)$ for each k. Choose a sequence of positive integers $\{n_\varepsilon\}$ such that $n_\varepsilon \to \infty$ but $\delta_\varepsilon = \varepsilon n_\varepsilon \to 0$ as $\varepsilon \to 0$. The piecewise constant interpolation implies that

$$\begin{aligned}
&\check{f}(x^\varepsilon(t+s),\alpha^\varepsilon(t+s)) - \check{f}(x^\varepsilon(t),\alpha^\varepsilon(t))\\
&= \sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon}[\check{f}(x^\varepsilon_{ln_\varepsilon+n_\varepsilon},\alpha^\varepsilon_{ln_\varepsilon+n_\varepsilon}) - \check{f}(x^\varepsilon_{ln_\varepsilon+n_\varepsilon},\alpha^\varepsilon_{ln_\varepsilon})]\\
&\quad + \sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon}[\check{f}(x^\varepsilon_{ln_\varepsilon+n_\varepsilon},\alpha^\varepsilon_{ln_\varepsilon}) - \check{f}(x^\varepsilon_{ln_\varepsilon},\alpha^\varepsilon_{ln_\varepsilon})],
\end{aligned} \tag{A.12}$$

and hence

$$\begin{aligned}
&\lim_{\varepsilon\to0} E\prod_{\kappa=1}^{k_0} h_\kappa(x^\varepsilon(t_\kappa),\overline{\alpha}^\varepsilon(t_\kappa))[\check{f}(x^\varepsilon(t+s),\alpha^\varepsilon(t+s)) - \check{f}(x^\varepsilon(t),\alpha^\varepsilon(t))]\\
&= \lim_{\varepsilon\to0} E\prod_{\kappa=1}^{k_0} h_\kappa(x^\varepsilon(t_\kappa),\overline{\alpha}^\varepsilon(t_\kappa)) \sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon}[\check{f}(x^\varepsilon_{ln_\varepsilon+n_\varepsilon},\alpha^\varepsilon_{ln_\varepsilon+n_\varepsilon}) - \check{f}(x^\varepsilon_{ln_\varepsilon+n_\varepsilon},\alpha^\varepsilon_{ln_\varepsilon})]\\
&\quad + \lim_{\varepsilon\to0} E\prod_{\kappa=1}^{k_0} h_\kappa(x^\varepsilon(t_\kappa),\overline{\alpha}^\varepsilon(t_\kappa)) \sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon}[\check{f}(x^\varepsilon_{ln_\varepsilon+n_\varepsilon},\alpha^\varepsilon_{ln_\varepsilon}) - \check{f}(x^\varepsilon_{ln_\varepsilon},\alpha^\varepsilon_{ln_\varepsilon})]\\
&\stackrel{\text{def}}{=} \lim_{\varepsilon\to0} E\prod_{\kappa=1}^{k_0} h_\kappa(x^\varepsilon(t_\kappa),\overline{\alpha}^\varepsilon(t_\kappa))[g^\varepsilon_1 + g^\varepsilon_2].
\end{aligned} \tag{A.13}$$

In the above, $\sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon}$ can also be written as $\sum_{ln_\varepsilon=t/\varepsilon}^{((t+s)/\varepsilon)-1}$. We proceed to obtain the desired limit by examining g^ε_i $(i=1,2)$ in (A.13).

By virtue of a Taylor expansion, rewrite g_2^ε as

$$\begin{aligned} g_2^\varepsilon &= \sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon} \check{f}_x'(x_{ln_\varepsilon},\alpha_{ln_\varepsilon}^\varepsilon)[x_{ln_\varepsilon+n_\varepsilon}^\varepsilon - x_{ln_\varepsilon}^\varepsilon] \\ &\quad + \frac{1}{2}\sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon} [x_{ln_\varepsilon+n_\varepsilon}^\varepsilon - x_{ln_\varepsilon}^\varepsilon]'\check{f}_{xx}(x_{ln_\varepsilon}^+,\alpha_{ln_\varepsilon}^\varepsilon)[x_{ln_\varepsilon+n_\varepsilon}^\varepsilon - x_{ln_\varepsilon}^\varepsilon] \\ &= \sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1} \check{f}_x'(x_{ln_\varepsilon}^\varepsilon,\alpha_{ln_\varepsilon}^\varepsilon)[\varepsilon A(\alpha_k^\varepsilon)x_k^\varepsilon + \sqrt{\varepsilon}\sigma_w(\alpha_k^\varepsilon)w_k] \\ &\quad + \frac{1}{2}\sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1} [\varepsilon A(\alpha_k^\varepsilon)x_k^\varepsilon + \sqrt{\varepsilon}\sigma_w(\alpha_k^\varepsilon)w_k]'\check{f}_{xx}(x_{ln_\varepsilon}^+,\alpha_{ln_\varepsilon}^\varepsilon) \\ &\quad \times \sum_{k_1=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1} [\varepsilon A(\alpha_{k_1}^\varepsilon)x_{k_1}^\varepsilon + \sqrt{\varepsilon}\sigma_w(\alpha_{k_1}^\varepsilon)w_{k_1}] \qquad \text{(A.14)} \\ &\stackrel{\text{def}}{=} \left[g_{2,1}^\varepsilon + \frac{1}{2}g_{2,2}^\varepsilon\right], \qquad \text{(A.15)} \end{aligned}$$

where $x_{ln_\varepsilon}^+$ is on the line segment joining $x_{ln_\varepsilon}^\varepsilon$ and $x_{ln_\varepsilon+n_\varepsilon}^\varepsilon$.

Then we have

$$\begin{aligned} &E\prod_{\kappa=1}^{k_0} h_\kappa(x^\varepsilon(t_\kappa),\overline{\alpha}^\varepsilon(t_\kappa))\left[\sqrt{\varepsilon}\sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon}\check{f}_x'(x_{ln_\varepsilon}^\varepsilon,\alpha_{ln_\varepsilon}^\varepsilon)\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}\sigma_w(\alpha_k^\varepsilon)w_k\right] \\ &= E\prod_{\kappa=1}^{k_0} h_\kappa(x^\varepsilon(t_\kappa),\overline{\alpha}^\varepsilon(t_\kappa))\left[\sqrt{\varepsilon}\sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon}\check{f}_x'(x_{ln_\varepsilon},\alpha_{ln_\varepsilon}^\varepsilon)\right. \\ &\qquad\left.\times\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}E_{ln_\varepsilon}\sigma_w(\alpha_k^\varepsilon)E_{ln_\varepsilon}w_k\right]. \end{aligned}$$

In the above, the second line is a consequence of the independence of $\{\alpha_n^\varepsilon\}$ and $\{w_n\}$ and the measurability of $x_{ln_\varepsilon}^\varepsilon$ and $\alpha_{ln_\varepsilon}^\varepsilon$ with respect to $\mathcal{F}_{ln_\varepsilon}$. In view of the boundedness of $h_\kappa(\cdot)$ and $\check{f}_x(\cdot)$ and the finiteness of $\sigma_w(\alpha_k^\varepsilon)$, we obtain

$$\begin{aligned} &E\prod_{\kappa=1}^{k_0} h_\kappa(x^\varepsilon(t_\kappa),\overline{\alpha}^\varepsilon(t_\kappa))\left[\sqrt{\varepsilon}\sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}\check{f}_x'(x_{ln_\varepsilon},\alpha_{ln_\varepsilon}^\varepsilon)E_{ln_\varepsilon}\sigma_w(\alpha_k^\varepsilon)E_{ln_\varepsilon}w_k\right] \\ &\to 0 \quad \text{as} \quad \varepsilon\to 0. \end{aligned} \qquad \text{(A.16)}$$

Next, let us treat the term on the second line of (A.14). First note that

$$\begin{aligned} &E\left|\varepsilon\sum_{i=1}^{l}\sum_{j=1}^{m_i}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}A(s_{ij})x_k^\varepsilon[I_{\{\alpha_k^\varepsilon=s_{ij}\}} - \nu_j^i I_{\{\alpha_k^\varepsilon\in\mathcal{M}_i\}}]\right| \\ &\le K\sum_{i=1}^{l}\sum_{j=1}^{m_i}E\left|\varepsilon\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}A(s_{ij})x_k^\varepsilon[I_{\{\alpha_k^\varepsilon=s_{ij}\}} - \nu_j^i I_{\{\alpha_k^\varepsilon\in\mathcal{M}_i\}}]\right|. \end{aligned}$$

Thus we need only examine the terms with fixed indices i and j. By a partial summation,

$$
\begin{aligned}
&E\left|\varepsilon \sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1} A(s_{ij})x_k^\varepsilon[I_{\{\alpha_k^\varepsilon=s_{ij}\}}-\nu_j^i I_{\{\alpha_k^\varepsilon\in\mathcal{M}_i\}}]\right| \\
&\le KE\left|\varepsilon A(s_{ij})x^\varepsilon_{ln_\varepsilon+n_\varepsilon-1}\sum_{k=0}^{ln_\varepsilon+n_\varepsilon-1}[I_{\{\alpha_k^\varepsilon=s_{ij}\}}-\nu_j^i I_{\{\alpha_k^\varepsilon\in\mathcal{M}_i\}}]\right| \\
&\quad + KE\left|\varepsilon A(s_{ij})x^\varepsilon_{ln_\varepsilon-1}\sum_{k=0}^{ln_\varepsilon-1}[I_{\{\alpha_k^\varepsilon=s_{ij}\}}-\nu_j^i I_{\{\alpha_k^\varepsilon\in\mathcal{M}_i\}}]\right| \\
&\quad + KE\left|\varepsilon \sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-2}(x_k^\varepsilon-x_{k+1}^\varepsilon)\sum_{k_0=0}^{k}[I_{\{\alpha_{k_0}^\varepsilon=s_{ij}\}}-\nu_j^i I_{\{\alpha_{k_0}^\varepsilon\in\mathcal{M}_i\}}]\right|.
\end{aligned}
$$

Using the mean square estimates on the occupation measures (2.10) and Lemma 3.2,

$$
\begin{aligned}
&E\left|\varepsilon A(s_{ij})x^\varepsilon_{ln_\varepsilon+n_\varepsilon-1}\sum_{k=0}^{ln_\varepsilon+n_\varepsilon-1}[I_{\{\alpha_k^\varepsilon=s_{ij}\}}-\nu_j^i I_{\{\alpha_k^\varepsilon\in\mathcal{M}_i\}}]\right| \\
&\le KE^{1/2}\left|x^\varepsilon_{ln_\varepsilon+n_\varepsilon-1}\right|^2 E^{1/2}\left|\varepsilon\sum_{k=0}^{ln_\varepsilon+n_\varepsilon-1}[I_{\{\alpha_k^\varepsilon=s_{ij}\}}-\nu_j^i I_{\{\alpha_k^\varepsilon\in\mathcal{M}_i\}}]\right|^2 \\
&= O(\sqrt{\varepsilon})\to 0 \quad \text{as} \quad \varepsilon\to 0.
\end{aligned}
$$

Similarly,

$$
E\left|\varepsilon A(s_{ij})x^\varepsilon_{ln_\varepsilon-1}\sum_{k=0}^{ln_\varepsilon-1}[I_{\{\alpha_k^\varepsilon=s_{ij}\}}-\nu_j^i I_{\{\alpha_k^\varepsilon\in\mathcal{M}_i\}}]\right|\to 0 \quad \text{as} \quad \varepsilon\to 0.
$$

Using (2.2), the Cauchy–Schwarz inequality, and the mean square estimates (2.10),

$$
\begin{aligned}
&E\left|\varepsilon \sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-2}(x_k^\varepsilon-x_{k+1}^\varepsilon)\sum_{k_0=0}^{k}[I_{\{\alpha_{k_0}^\varepsilon=s_{ij}\}}-\nu_j^i I_{\{\alpha_{k_0}^\varepsilon\in\mathcal{M}_i\}}]\right| \\
&\le \sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-2} E^{1/2}|x_k^\varepsilon-x_{k+1}^\varepsilon|^2 E^{1/2}\left|\varepsilon\sum_{k_0=0}^{k}[I_{\{\alpha_{k_0}^\varepsilon=s_{ij}\}}-\nu_j^i I_{\{\alpha_{k_0}^\varepsilon\in\mathcal{M}_i\}}]\right|^2 \\
&\le \sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-2} E^{1/2}|\varepsilon A(\alpha_k^\varepsilon)x_k^\varepsilon+\sqrt{\varepsilon}\sigma_w(\alpha_k^\varepsilon)w_k|^2 \\
&\times E^{1/2}\left|\varepsilon\sum_{k_0=0}^{k}[I_{\{\alpha_{k_0}^\varepsilon=s_{ij}\}}-\nu_j^i I_{\{\alpha_{k_0}^\varepsilon\in\mathcal{M}_i\}}]\right|^2 \\
&\to 0 \quad \text{as} \quad \varepsilon\to 0.
\end{aligned}
$$

Using the above estimates and the continuity of $\check{f}_x(\cdot,\alpha)$ for each $\alpha \in \mathcal{M}$,

$$
\begin{aligned}
&E\prod_{\kappa=1}^{k_0} h_\kappa(x^\varepsilon(t_\kappa),\overline{\alpha}^\varepsilon(t_\kappa))\left[\varepsilon \sum_{l:t\le l\delta_\varepsilon\le (t+s)-\varepsilon} \check{f}'_x(x^\varepsilon_{ln_\varepsilon},\alpha^\varepsilon_{ln_\varepsilon}) \sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1} A(\alpha^\varepsilon_k)x^\varepsilon_k\right]\\
&= E\prod_{\kappa=1}^{k_0} h_\kappa(x^\varepsilon(t_\kappa),\overline{\alpha}^\varepsilon(t_\kappa))\left[\sum_{l:t\le l\delta_\varepsilon\le (t+s)-\varepsilon}\sum_{i=1}^{l}\sum_{j=1}^{m_i} \check{f}'_x(x_{ln_\varepsilon},\alpha^\varepsilon_{ln_\varepsilon})\right.\\
&\qquad\left.\times \frac{\delta_\varepsilon}{n_\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1} A(s_{ij})x^\varepsilon_{ln_\varepsilon}\nu^i_j I_{\{\alpha^\varepsilon_k\in\mathcal{M}_i\}}\right]+o(1),
\end{aligned}
\tag{A.17}
$$

where $o(1)\to 0$ as $\varepsilon\to 0$. Then, as $\varepsilon\to 0$, letting $\varepsilon l n_\varepsilon \to u$, and using the techniques of [19, Chapter 8], (A.17) together with (A.16) leads to

$$
\begin{aligned}
&E\prod_{\kappa=1}^{k_0} h_\kappa(x^\varepsilon(t_\kappa),\overline{\alpha}^\varepsilon(t_\kappa))g^\varepsilon_{2,1}\\
&\to E\prod_{\kappa=1}^{k_0} h_\kappa(x(t_\kappa),\overline{\alpha}(t_\kappa))\left(\int_t^{t+s} f'_x(x(u),\overline{\alpha}(u))A(\overline{\alpha}(u))x(u)du\right)\quad\text{as}\quad \varepsilon\to 0.
\end{aligned}
\tag{A.18}
$$

As for $g^\varepsilon_{2,2}$, we have, by the continuity of $f_{xx}(\cdot,\alpha)$ for each $\alpha\in\mathcal{M}$, $x^+_{ln_\varepsilon}-x^\varepsilon_{ln_\varepsilon}\to 0$ in probability as $\varepsilon\to 0$. Consequently,

$$
E\prod_{\kappa=1}^{k_0} h_\kappa(x^\varepsilon(t_\kappa),\overline{\alpha}^\varepsilon(t_\kappa))g^\varepsilon_{2,2} \stackrel{\text{def}}{=} \widetilde{g}^\varepsilon_{2,2}+o(1),
$$

where $o(1)\to 0$ as $\varepsilon\to 0$ uniformly in t, and

$$
\begin{aligned}
\widetilde{g}^\varepsilon_{2,2} &= E\prod_{\kappa=1}^{k_0} h_\kappa(x^\varepsilon(t_\kappa),\overline{\alpha}^\varepsilon(t_\kappa))\\
&\times\left[\sum_{l:t\le l\delta_\varepsilon\le (t+s)-\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}[\varepsilon A(\alpha^\varepsilon_k)x^\varepsilon_k+\sqrt{\varepsilon}\sigma_w(\alpha^\varepsilon_k)w_k]'\check{f}_{xx}(x^\varepsilon_{ln_\varepsilon},\alpha^\varepsilon_{ln_\varepsilon})\right.\\
&\qquad\left.\times\sum_{k_1=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}[\varepsilon A(\alpha^\varepsilon_{k_1})x^\varepsilon_{k_1}+\sqrt{\varepsilon}\sigma_w(\alpha^\varepsilon_{k_1})w_{k_1}]\right].
\end{aligned}
$$

It then follows that

$$\begin{aligned}
\widetilde{g}^{\varepsilon}_{2,2} &= E\prod_{\kappa=1}^{k_0} h_\kappa(x^\varepsilon(t_\kappa),\overline{\alpha}^\varepsilon(t_\kappa))\\
&\quad\times\Bigg[\varepsilon^2\sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}(A(\alpha^\varepsilon_k)x^\varepsilon_k)'\check{f}_{xx}(x^\varepsilon_{ln_\varepsilon},\alpha^\varepsilon_{ln_\varepsilon})\sum_{k_1=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}A(\alpha^\varepsilon_{k_1})x^\varepsilon_{k_1}\\
&\quad+\sqrt{\varepsilon^3}\sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}(A(\alpha^\varepsilon_k)x^\varepsilon_k)'\check{f}_{xx}(x^\varepsilon_{ln_\varepsilon},\alpha^\varepsilon_{ln_\varepsilon})\sum_{k_1=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}\sigma_w(\alpha^\varepsilon_{k_1})w_{k_1}\\
&\quad+\sqrt{\varepsilon^3}\sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}(\sigma_w(\alpha^\varepsilon_k)w_k)'\check{f}_{xx}(x^\varepsilon_{ln_\varepsilon},\alpha^\varepsilon_{ln_\varepsilon})\sum_{k_1=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}A(\alpha^\varepsilon_{k_1})x^\varepsilon_{k_1}\\
&\quad+\varepsilon\sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}(\sigma_w(\alpha^\varepsilon_k)w_k)'\check{f}_{xx}(x^\varepsilon_{ln_\varepsilon},\alpha^\varepsilon_{ln_\varepsilon})\sum_{k_1=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}\sigma_w(\alpha^\varepsilon_{k_1})w_{k_1}\Bigg]\\
&= E\prod_{\kappa=1}^{k_0} h_\kappa(x^\varepsilon(t_\kappa),\overline{\alpha}^\varepsilon(t_\kappa))\\
&\quad\times\Bigg[\varepsilon\sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}(\sigma_w(\alpha^\varepsilon_k)w_k)'\check{f}_{xx}(x^\varepsilon_{ln_\varepsilon},\alpha^\varepsilon_{ln_\varepsilon})\sigma_w(\alpha^\varepsilon_k)w_k\Bigg]+o(1),
\end{aligned}$$

where $o(1)\to 0$ as $\varepsilon\to 0$. Furthermore, using the idea of the estimates leading to (A.17) and the mean square estimates (2.10), it can be shown that

$$\begin{aligned}
&\varepsilon\sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}(\sigma_w(\alpha^\varepsilon_k)w_k)'\check{f}_{xx}(x^\varepsilon_{ln_\varepsilon},\alpha^\varepsilon_{ln_\varepsilon})\sigma_w(\alpha^\varepsilon_k)w_k\\
&=\sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon}\sum_{i=1}^{l}\sum_{j=1}^{m_i}\frac{\delta_\varepsilon}{n_\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}\mathrm{tr}[\check{f}_{xx}(x^\varepsilon_{ln_\varepsilon},s_{ij})\sigma_w(s_{ij})w_kw_k'\sigma_w'(s_{ij})]I_{\{\alpha^\varepsilon_k=s_{ij}\}}\\
&=\sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon}\sum_{i=1}^{l}\sum_{j=1}^{m_i}\frac{\delta_\varepsilon}{n_\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}\mathrm{tr}[\check{f}_{xx}(x^\varepsilon_{ln_\varepsilon},\alpha^\varepsilon_{ln_\varepsilon})\sigma_w(s_{ij})w_kw_k'\sigma_w'(s_{ij})]\\
&\quad\times\nu^i_jI_{\{\alpha^\varepsilon_k\in\mathcal{M}_i\}}+o(1),
\end{aligned}$$

where $o(1)\to 0$ in probability as $\varepsilon\to 0$ uniformly in t. It then follows that

$$\begin{aligned}
\lim_{\varepsilon\to 0}&E\prod_{\kappa=1}^{k_0} h_\kappa(x^\varepsilon(t_\kappa),\overline{\alpha}^\varepsilon(t_\kappa))\\
&\times\Bigg[\varepsilon\sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}(\sigma_w(\alpha^\varepsilon_k)w_k)'\check{f}_{xx}(x^\varepsilon_{ln_\varepsilon},\alpha^\varepsilon_{ln_\varepsilon})\sigma_w(\alpha^\varepsilon_k)w_k\Bigg]\\
&=E\prod_{\kappa=1}^{k_0} h_\kappa(x(t_\kappa),\overline{\alpha}(t_\kappa))\left[\int_t^{t+s}\mathrm{tr}[f_{xx}(x(u),\overline{\alpha}(u))\sigma_w(\overline{\alpha}(u))\sigma_w'(\overline{\alpha}(u))]du\right].
\end{aligned}\tag{A.19}$$

Next, we consider the term g^ε_1. Using the continuity of $\check{f}(\cdot,\alpha)$ for each $\alpha\in\mathcal{M}$, the Markov property of α^ε_n, the mean square estimate (2.10) of the occupation measures,

(2.1), and Lemma 2.1, we have

$$
\begin{aligned}
(\text{A.20})\ & E\prod_{\kappa=1}^{k_0} h_\kappa(x^\varepsilon(t_\kappa),\overline{\alpha}^\varepsilon(t_\kappa))g_1^\varepsilon \\
&= E\prod_{\kappa=1}^{k_0} h_\kappa(x^\varepsilon(t_\kappa),\overline{\alpha}^\varepsilon(t_\kappa))\left[\sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon}(\check{f}(x^\varepsilon_{ln_\varepsilon},\alpha^\varepsilon_{ln_\varepsilon+n_\varepsilon})-\check{f}(x^\varepsilon_{ln_\varepsilon},\alpha^\varepsilon_{ln_\varepsilon}))\right] \\
&= E\prod_{\kappa=1}^{k_0} h_\kappa(x^\varepsilon(t_\kappa),\overline{\alpha}^\varepsilon(t_\kappa))\left[\sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}\sum_{i_1=1}^{l}\sum_{j_1=1}^{m_{i_1}}\left[\sum_{i=1}^{l}\sum_{j=1}^{m_i}\check{f}(x^\varepsilon_{ln_\varepsilon},s_{ij})\right.\right. \\
&\qquad \left.\left.\times P(\alpha^\varepsilon_{k+1}=s_{ij}|\alpha^\varepsilon_k=s_{i_1j_1})-\check{f}(x^\varepsilon_{ln_\varepsilon},s_{i_1j_1})\right]I_{\{\alpha^\varepsilon_k=s_{i_1j_1}\}}\right] \\
&= E\prod_{\kappa=1}^{k_0} h_\kappa(x^\varepsilon(t_\kappa),\overline{\alpha}^\varepsilon(t_\kappa))\left[\varepsilon\sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}(P-I+\varepsilon Q)\check{f}(x^\varepsilon_{ln_\varepsilon},\cdot)(\alpha^\varepsilon_k)\right] \\
&= E\prod_{\kappa=1}^{k_0} h_\kappa(x^\varepsilon(t_\kappa),\overline{\alpha}^\varepsilon(t_\kappa))\left[\varepsilon\sum_{l:t\le l\delta_\varepsilon\le(t+s)-\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}Q\check{f}(x^\varepsilon_{ln_\varepsilon},\cdot)(\alpha^\varepsilon_k)\right] \\
&\quad \to \int_t^{t+s}\overline{Q}f(x(u),\overline{\alpha}(u))du \ \text{ as } \ \varepsilon\to 0.
\end{aligned}
$$

Combining (A.18), (A.19), and (A.20),

$$
\begin{aligned}
(\text{A.21})\qquad & \lim_{\varepsilon\to 0} E\prod_{\kappa=1}^{k_0} h_\kappa(x^\varepsilon(t_\kappa),\overline{\alpha}^\varepsilon(t_\kappa))\left[\check{f}(x^\varepsilon(t+s),\overline{\alpha}^\varepsilon(t+s))-\check{f}(x^\varepsilon(t),\overline{\alpha}^\varepsilon(t))\right] \\
&\qquad = E\prod_{\kappa=1}^{k_0} h_\kappa(x(t_\kappa),\overline{\alpha}(t_\kappa))\left[\int_t^{t+s}\mathcal{L}f(x(u),\overline{\alpha}(u))du\right].
\end{aligned}
$$

On the other hand, by the weak convergence of $(x^\varepsilon(\cdot),\overline{\alpha}^\varepsilon(\cdot))$ to $(x(\cdot),\overline{\alpha}(\cdot))$, the Skorohod representation, and the definition of $\check{f}(\cdot)$, we have

$$
\begin{aligned}
(\text{A.22})\qquad & \lim_{\varepsilon\to 0} E\prod_{\kappa=1}^{k_0} h_\kappa(x^\varepsilon(t_\kappa),\overline{\alpha}^\varepsilon(t_\kappa))[\check{f}(x^\varepsilon(t+s),\alpha^\varepsilon(t+s))-\check{f}(x^\varepsilon(t),\alpha^\varepsilon(t))] \\
&\qquad = E\prod_{\kappa=1}^{k_0} h_\kappa(x(t_\kappa),\overline{\alpha}(t_\kappa))[f(x(t+s),\overline{\alpha}(t+s))-f(x(t),\overline{\alpha}(t))].
\end{aligned}
$$

By (A.21) and (A.22), (A.10) holds. Using the same techniques, detailed estimates yield the second equation in (3.4). Thus the desired results follow. □

REFERENCES

[1] Y. Bar-Shalom and X. R. Li, *Estimation and Tracking: Principles, Techniques, and Software*, Artech House Publishers, Norwood, MA, 1996.

[2] D. P. Bertsekas, *Dynamic Programming and Stochastic Control*, Academic Press, New York, 1976.

[3] T. R. Bielecki and L. Stettner, *Ergodic control of a singularly perturbed Markov process in discrete time with general state and compact action spaces*, Appl. Math. Optim., 38 (1998), pp. 261–281.
[4] T. Björk, *Finite-dimensional optimal filters for a class of Itô processes with jumping parameters*, Stochastics, 4 (1980), pp. 167–183.
[5] G. Blankenship, *Singularly perturbed difference equations in optimal control problems*, IEEE Trans. Automat. Control, 26 (1981), pp. 911–917.
[6] O. L. V. Costa, *Linear minimum mean square error estimation for discrete-time Markov jump linear systems*, IEEE Trans. Automat. Control, 39 (1994), pp. 1685–1689.
[7] P. J. Courtois, *Decomposability: Queueing and Computer System Applications*, Academic Press, New York, 1977.
[8] D. P. de Farias, J. C. Geromel, J. B. R. do Val, and O. L. V. Costa, *Output feedback control of Markov jump linear systems in continuous time*, IEEE Trans. Automat. Control, 45 (2000), pp. 944–949.
[9] S. Dey, *Reduced-complexity filtering for partially observed nearly completely decomposable Markov chains*, IEEE Trans. Signal Process., 48, (2000), pp. 3334–3344.
[10] A. Doucet, N. J. Gordon, and V. Krishnamurthy, *Particle filtering for state estimation for jump Markov linear systems*, IEEE Trans. Signal Process., 49 (2001), pp. 613–624.
[11] F. Dufour and P. Bertrand, *The filtering problem for continuous-time linear systems with Markovian switching coefficients*, Systems Control Lett., 23 (1994), pp. 453–461.
[12] F. Dufour and R. J. Elliott, *Adaptive control of linear systems with Markov perturbations*, IEEE Trans. Automat. Control, 43 (1997), pp. 351–372.
[13] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*, John Wiley, New York, 1986.
[14] A. Il'in, R. Z. Khasminskii, and G. Yin, *Singularly perturbed switching diffusions: Rapid switchings and fast diffusions*, J. Optim. Theory Appl., 102 (1999), pp. 555–591.
[15] M. Iosifescu, *Finite Markov Processes and Their Applications*, John Wiley, Chichester, UK, 1980.
[16] R. Z. Khasminskii, G. Yin, and Q. Zhang, *Asymptotic expansions of singularly perturbed systems involving rapidly fluctuating Markov chains*, SIAM J. Appl. Math., 56 (1996), pp. 277–293.
[17] H. J. Kushner, *Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory*, MIT Press, Cambridge, MA, 1984.
[18] H. J. Kushner, *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, Birkhäuser Boston, Boston, 1990.
[19] H. J. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, New York, 1997.
[20] R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes* I & II, Springer-Verlag, New York, 2001.
[21] B. M. Miller and W. J. Runggaldier, *Kalman filtering for linear systems with coefficients driven by a hidden Markov jump process*, Systems Control Lett., 31 (1997), pp. 93–102.
[22] A. A. Pervozvanskii and V. G. Gaitsgori, *Theory of Suboptimal Decisions: Decomposition and Aggregation*, Kluwer, Dordrecht, The Netherlands, 1988.
[23] R. G. Phillips and P. V. Kokotovic, *A singular perturbation approach to modelling and control of Markov chains*, IEEE Trans. Automat. Control, 26 (1981), pp. 1087–1094.
[24] W. J. Runggaldier and C. Visentin, *Combined filtering and parameter estimation: Approximation and robustness*, Automatica J. IFAC, 26 (1990), pp. 401–404.
[25] S. P. Sethi and Q. Zhang, *Hierarchical Decision Making in Stochastic Manufacturing Systems*, Birkhäuser Boston, Boston, 1994.
[26] H. A. Simon and A. Ando, *Aggregation of variables in dynamic systems*, Econometrica, 29 (1961), pp. 111–138.
[27] D. N. C. Tse, R. G. Gallager, and J. N. Tsitsiklis, *Statistical multiplexing of multiple time-scale Markov streams*, IEEE J. Selected Areas Comm., 13 (1995), pp. 1028–1038.
[28] C. Yang, Y. Bar-Shalom, and C.-F. Lin, *Discrete-time point process filter for mode estimation*, IEEE Trans. Automat. Control, 37 (1992), pp. 1812–1816.
[29] G. Yin and Q. Zhang, *Continuous-Time Markov Chains and Applications: A Singular Perturbation Approach*, Springer-Verlag, New York, 1998.
[30] G. Yin and Q. Zhang, *Singularly perturbed discrete-time Markov chains*, SIAM J. Appl. Math., 61 (2000), pp. 834–854.
[31] G. Yin, Q. Zhang, and G. Badowski, *Asymptotic properties of a singularly perturbed Markov chain with inclusion of transient states*, Ann. Appl. Probab., 10 (2000), pp. 549–572.

[32] G. Yin, Q. Zhang, and G. Badowski, *Decomposition and aggregation of large-dimensional Markov chains in discrete time*, in Proceedings of the 40th IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 2001, pp. 1687–1692.

[33] Q. Zhang, *Hybrid filtering for linear systems with non-Gaussian disturbances*, IEEE Trans. Automat. Control, 45 (2000), pp. 50–61.

[34] Q. Zhang and G. Yin, *On nearly optimal controls of hybrid LQG problems*, IEEE Trans. Automat. Control, 44 (1999), pp. 2271–2282.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 6, pp. 1843–1867

ON STABILITY OF BANG-BANG TYPE CONTROLS*

URSULA FELGENHAUER†

Abstract. From the theory of nonlinear optimal control problems it is known that the solution stability w.r.t. data perturbations and conditions for strict local optimality are closely related facts. For important classes of control problems, sufficient optimality conditions can be formulated as a combination of the independence of active constraints' gradients and certain coercivity criteria. In the case of discontinuous controls, however, common pointwise coercivity approaches may fail.

In the paper, we consider sufficient optimality conditions for strong local minimizers which make use of an integrated Hamilton–Jacobi inequality. In the case of linear system dynamics, we show that the solution stability (including the switching points localization) is ensured under relatively mild regularity assumptions on the switching function zeros. For the objective functional, local quadratic growth estimates in L_1 sense are provided. An example illustrates stability as well as instability effects in case the regularity condition is violated.

Key words. optimal control, optimality conditions, strong local minimality, control stability, solution structure stability

AMS subject classifications. 49K40, 49N15, 49N10

PII. S0363012901399271

1. Introduction.

The paper is devoted to optimality conditions and the solution stability for certain linear, linear-quadratic, or linear time-optimal control problems with typically discontinuous optimal control functions. At the present time, optimal controls with bang-bang structure are intensively studied, which is reflected in the growing number of publications on the topic. Second-order sufficient conditions are derived, e.g., by Noble and Schaettler [28], Schättler [36], or Ledzewicz and Schaettler [15], by using the method of characteristics in adapted form for analyzing the flow of extremals. In his paper [35], Sarychev introduced special first- and second-order variations including measure type additives at the switching points in order to handle control discontinuities. Very recently, Agrachev, Stefani, and Zezza [3] proposed an approach treating the switching times as the main unknown. In their work, ideas of the so-called symplectic geometry going back to [1] (also [2]) are basically involved. A comprehensive theory of second-order conditions in optimal control is given by Milyutin and Osmolovskii in [27] and Osmolovskii [29]. The central feature in their book is the analysis of Pontryagin minima. The criteria derived are of high generality and strong in their closeness to necessary optimality conditions but partly difficult to check in applications. It was shown, e.g., in [30], [31] how they can be applied to bang-bang control regimes.

Our paper starts from weak duality relations going back to the work of Klötzler and his group ([13], [32], and [14], e.g.). It has been modified and extensively used in the recent past for deriving weak local optimality conditions, cf. [34], [26], [33]. The criteria obtained are often given in terms of the independence of the gradients of the active constraints w.r.t. the control, together with certain coercivity properties of the Hamilton function. For the class of problems considered in sections 3–5, however, in

*Received by the editors December 5, 2001; accepted for publication (in revised form) August 6, 2002; published electronically February 6, 2003.
http://www.siam.org/journals/sicon/41-6/39927.html
†Institut für Mathematik, Brandenburgische Technische Universität Cottbus, Germany, PF 101344, 03013 Cottbus (felgenh@math.tu-cottbus.de).

particular the classical Legendre–Clebsch condition ([4], e.g.) is not fulfilled, so that the optimality test requires alternative estimation techniques.

In [11], [12] we have shown that the optimality criteria in their integrated version from [8] are, in principle, applicable to problems with discontinuous optimal control regimes. In section 2 this approach is shortly described, and the basic Hamilton–Jacobi inequality in integrated form is given (Theorem 2.2). Subsequently, local quadratic growth estimates for the objective functionals are obtained in the bang-bang case for soft termination control problems with linear systems dynamics (section 3), and also for a related time-optimal problem (section 5).

The investigation is restricted to nonsingular extremals having only finitely many switching points (Assumption 1) with regular zeros of the switching function components (Assumption 2). Under these conditions, the solution is proved to be a strict strong local minimizer such that the objective functional satisfies a local quadratic growth estimation in $L_2 \times L_1$ sense (Theorem 3.4). The proof uses a direct estimation technique for admissible variations without involving approximation cones. Although the obtained characterization of the optimum could be strengthened by means of the theory developed in [27] and [29] (see also section 7), the method presented here is of interest as a self-contained and easy to prove alternative way of analyzing strong local minimizers.

In section 4 it is proved that, under the given assumptions, the switching structure of the optimal control behaves stably w.r.t. small data perturbations (Theorem 4.1). The result is independent of the previous section. As a supplement, in section 5 the time-optimal problem is analyzed. Introducing an additional state variable for the free final time, one can find an extended dual formulation and derive appropriate optimality conditions. Although the estimates obtained (see Lemma 5.3) are slightly weaker than the requirements in Theorem 2.2, a local quadratic growth estimation results whenever (x, T) is close to the reference state and end-time solution (cf. Theorem 5.4). In addition, a sensitivity result for the optimal time (Lemma 5.5) is derived.

Finally, section 6 is devoted to the example of a two-dimensional chain problem with fixed end-time T. If T is smaller than the optimal termination time, the solution is locally strict and stable. When the optimal end-time T^* is reached, the solution becomes singular and instable: for $T > T^*$, the solution is no longer unique, and the switching structure shows serious bifurcations.

2. Local optimality criteria in integrated form. In this section, it will be shortly explained how sufficient optimality criteria and related growth estimates can be obtained from an abstract dualization for optimal control problems. The approach goes back to Klötzler [13], and it has been repeatedly used and described in detail, e.g., in [14], [32], [34], or [26]. In [8] or [11], variants of the basically parametric criteria have been considered in integrated form.

First, let be given a general nonlinear constrained optimal control problem (*primal* problem formulation):

$$\text{(P)} \qquad \min\ J(x,u) \ =\ k(x(0),x(T)) \ +\ \int_0^T r(t,x(t),u(t))\,dt$$

$$\text{(2.1)} \qquad \text{subject to (s.t.)} \quad \dot{x} \ = f(t,x(t),u(t)) \qquad \text{a.e. in } [0,T],$$

$$\text{(2.2)} \qquad \beta(x(0),x(T)) \ =\ 0,$$

$$\text{(2.3)} \qquad g(t,x(t),u(t)) \ \le\ 0 \qquad \text{a.e. in } [0,T]\,,$$

where $T > 0$ is given.

The pair $(x,u) \in W^1_\infty(0,T;R^n) \times L_\infty(0,T;R^k)$ is called *admissible* for (P) if the state equation (2.1) together with the boundary condition (2.2) and the inequality constraints (2.3) (where $g : [0,T] \times R^n \times R^k \to R^m$, $\beta : R^n \times R^n \to R^s$) is fulfilled. All data functions are assumed to be sufficiently smooth. An admissible pair (x_0,u_0) is called a (*global*) minimizer for (P) if $J(x_0,u_0) \le J(x,u)$ for all admissible (x,u). For characterizing *local* minimizers, the following definitions are used (see, e.g., [27], [31]).

DEFINITION 2.1. *The pair (x_0,u_0) is called a weak local minimizer of* (P) *if it is admissible and if a constant $\epsilon > 0$ exists such that $J(x_0,u_0) \le J(x,u)$ for any admissible pair (x,u) with $\| x - x_0\|_\infty + \| u - u_0\|_\infty < \epsilon$.*

An admissible pair (x_0,u_0) is called a strong local minimizer if $J(x_0,u_0) \le J(x,u)$ for any admissible (x,u) with $\| x - x_0\|_\infty < \epsilon$. If, for each positive M, an $\epsilon > 0$ exists such that $J(x_0,u_0) \le J(x,u)$ is satisfied for any admissible pair (x,u) with $\| x - x_0\|_\infty < \epsilon$ and $\|u\|_\infty \le M$, then (x_0,u_0) is a bounded-strong minimizer.

If, in addition, for $(x,u) \ne (x_0,u_0)$ (resp., $x \ne x_0$) the inequality $J(x_0,u_0) < J(x,u)$ holds true, (x_0,u_0) is a strict weak (resp., strict strong or strict bounded-strong) local minimizer.

Denote by H the Hamiltonian and by $\hat{H}$ the *augmented* Hamiltonian related to (P):

$$H(t,x,u,p) \;=\; r(t,x,u) \;+\; p^T f(t,x,u)\;,$$

$$\hat{H}(t,x,u,p,\mu) \;=\; H(t,x,u,p) \;+\; \mu^T g(t,x,u)\;, \qquad \mu \ge 0.$$

Further, let W stand for the set

$$W(t) \;=\; \{\,(x,u) :\; g(t,x,u) \le 0\,\}.$$

Second, let us consider a dual variable S given as a function $S : [0,T] \times R^n \to R$ and the auxiliary functional θ (for $\xi_{1,2} \in R^n$, resp.):

$$\theta(\xi_1,\xi_2;\,S) \;=\; k(\xi_1,\xi_2) \;+\; S(0,\xi_1) - S(T,\xi_2).$$

We assume that S is continuously differentiable w.r.t. x and at least piecewise continuously differentiable w.r.t. t. Define

$$T(S) \;=\; \inf_{(\xi_1,\xi_2)} \{\theta(\xi_1,\xi_2;\,S) \;:\; \beta(\xi_1,\xi_2) = 0\,\}.$$

Then the following problem is a dual to the original control problem (P):

$$\begin{aligned} \text{(D)} \quad & \max\; T(S) \\ & \text{s.t. } H(t,x,u,S_x(t,x)) \;+\; S_t(t,x) \;\ge\; 0 \qquad \forall\,(x,u) \in W(t) \;\text{ a.e. on }\; [0,T]\,. \end{aligned}$$

In the case of control problems without inequality constraints, the dual function S satisfies the Hamilton–Jacobi equation and can be interpreted as a *verification function* (see [5], e.g.). The dual problem formulation by Klötzler [13], or Pickenhain [32], thus allows for generalizing this variational approach to constrained control problems.

It is easy to see that $J(x,u) \geq T(S)$ whenever (x,u) is admissible for problem (P) and S is feasible for (D): denoting $x(0) = x_1, x(T) = x_2$ for admissible (x,u), we have

$$\begin{aligned} J(x,u) - T(S) &= \int_0^T r(t,x(t),u(t))\,dt + k(x_1,x_2) - T(S) \\ &= \int_0^T \left[H(t,x(t),u(t),S_x(t,x(t))) + S_t(t,x(t))\right]\,dt \\ &\quad - \int_0^T \left[S_x(t,x(t))^T \dot{x}(t) + S_t(t,x(t))\right]\,dt + k(x_1,x_2) - T(S) \\ &= \int_0^T \left[H(t,x(t),u(t),S_x(t,x(t))) + S_t(t,x(t))\right]\,dt \\ &\qquad + \theta(x_1,x_2;\,S) - \inf_{(\xi_1,\xi_2)} \theta(\xi_1,\xi_2;\,S) \geq 0\,. \end{aligned} \tag{2.4}$$

Thus, a weak duality relation holds for the problem pair (P), (D) (see [13], also [32], [8]). Let us introduce

$$\begin{aligned} \Psi(x,u;\,S) &= \int_0^T \left[\, H(t,x(t),u(t),S_x(t,x(t))) + S_t(t,x(t))\,\right]\,dt\,, \\ \psi(\xi_1,\xi_2;\,S) &= \theta(\xi_1,\xi_2;\,S) - T(S)\,. \end{aligned} \tag{2.5}$$

Then, the duality gap $(J(x,u) - T(S))$ equals zero if and only if for some admissible (x_0,u_0) and feasible dual S, $\Psi(x_0,u_0;\,S) = 0$ and $\psi(x_0(0),x_0(T);\,S) = 0$. In this case, the pair (x_0,u_0) is a solution of (P).

The analysis of the behavior of Ψ and ψ can be used to verify local minimality of a solution including estimates for local growth terms if available (cf., e.g., [12]). It can be applied to *weak* as well as to *strong* local optima in dependence of the reference sets choice. For example, consider the sets

$$\begin{aligned} W_\epsilon(t) = W(t) \cap B_\epsilon(x_0(t),u_0(t))\,, \qquad \hat{W}_\epsilon(t) &= W(t) \cap \left(B_\epsilon(x_0(t)) \times R^k\right), \\ \tilde{W}_{\epsilon,M}(t) &= W(t) \cap \left(B_\epsilon(x_0(t)) \times B_M(0)\right), \end{aligned}$$

where $B_r(z)$ denotes closed balls of radius r in the related Euclidean spaces.

The following result then characterizes general strict local optima.

THEOREM 2.2 (see [26], [12]). *Let (x_0,u_0) be admissible for* (P). *Suppose that a function $S : [0,T] \times R^n \to R$ exists which is Lipschitz continuous w.r.t. x and piecewise continuously differentiable w.r.t. t such that for suitably chosen positive constants c and ϵ the following relations hold with $1 \leq p < \infty$, $\gamma = 1$, $D(t) = W_\epsilon(t)$, and $D_\pi = B_\epsilon(x_0(0),x_0(T)) \cap \{\xi = (\xi_1,\xi_2) : \beta(\xi) = 0\}$:*

(R1) $$\Psi(x,u;S) \geq c\left(\|x - x_0\|_2^2 + \gamma\,\|u - u_0\|_p^2\right)$$
$$\forall \text{ admissible } (x,u) \text{ with } (x(t),u(t)) \in D(t) \quad \text{a.e. in } [0,T];$$

(R2) $$\Psi(x_0,u_0;S) = 0\,, \qquad \psi(x_0(0),x_0(T);\,S) = 0;$$

(R3) $$\psi(\xi_1,\xi_2;\,S) \geq 0 \qquad \forall\,\xi \in D_\pi\,.$$

Then (x_0,u_0) is a strict weak local minimizer of (P).

If (R1)–(R3) *hold true for a certain constant $\epsilon > 0$ with $\gamma = 0$, $D(t) = \hat{W}_\epsilon(t)$, and D_π given above, then (x_0,u_0) is a (strict) strong local minimizer.*

If, for arbitrary $M > 0$, a positive ϵ exists such that (R1)–(R3) *are satisfied with $D(t) = \tilde{W}_{\epsilon,M}(t)$, the point (x_0, u_0) is a (strict) bounded-strong local optimum.*

Furthermore,

$$J(x,u) - J(x_0,u_0) \;\geq\; c'\left(\|x - x_0\|_2^2 \;+\; \gamma\,\|u - u_0\|_p^2\right). \tag{2.6}$$

holds true for all admissible (x,u) on the related reference set, i.e., with $(x(t),u(t)) \in D(t)$ a.e. on $[0,T]$, and $(x(0),x(T)) \in D_\pi$, respectively.

The proof of the theorem follows from (2.4) together with the description of the functionals in (2.5). Under condition (R2), in particular, $J(x_0,u_0) = T(S)$ holds true. The above theorem differs from formulations given in [32], [26], e.g., mainly by the consequent usage of the Hamilton–Jacobi inequality (cf. (D)) in its *integrated* instead of its *parametric* form. This relaxation was already used in [8] for analyzing Ritz type discretization methods in optimal control.

Let us mention the fact that the local growth estimate for J is formulated in terms of L_2 (resp., L_p) topology whereas the reference sets are L_∞-neighborhoods w.r.t. x in particular. This effect is well known as part of the *two-norm discrepancy* in optimal control problems (see, e.g., [16]).

Starting from abstract results like Theorem 2.2, one can establish sufficient optimality conditions for general control problems in a form which in principle is suitable for numerical tests (see [26], [20], also [18] and [21]). The known criteria include independence of the (nearly) active constraints as well as certain coercivity conditions.

The independence conditions are usually given in terms of the *invertibility* w.r.t. control for the (nearly) active inner constraints and a certain *controllability* assumption. Using Pontryagin's maximum principle, the following system of first-order necessary optimality conditions is obtained for a weak local minimizer:

Canonical system (complementarity formulation):

$$\begin{aligned}
\dot{x} \;&= \hat{H}_p \;=\; f\,, \qquad \beta(x(0),x(T)) = 0\,;\\
\dot{p} \;&= -\hat{H}_x;\\
p(0) &= -\nabla_1 k - \nabla_1\beta\cdot\rho, \quad p(T) = \nabla_2 k + \nabla_2\beta\cdot\rho\,, \qquad \rho \in R^s;\\
\hat{H}_u \;&= 0\,;\\
\mu^T g &= 0\,, \quad \mu \geq 0\,, \quad g \leq 0.
\end{aligned} \tag{2.7}$$

Sufficient conditions are usually given as second-order criteria including, first, the so-called Legendre–Clebsch condition, and, second, the solvability of a certain Riccati matrix differential (in)equality on a parameter function Q. As a rule, these characterizations are strongly local w.r.t. (x,u) and thus are mainly related to weak local optima (cf. [26], also [22] and [23].

The matrix function Q and the terms in (2.7) are connected with the optimal S as follows (see [8]):

$$\dot{S}(t,x_0) = -r_0[t]\,, \quad S_x(t,x_0) = p_0(t)\,, \quad S_{xx}(t,x_0) = Q(t).$$

(Here r_0 stands for r evaluated along $(x_0(t),u_0(t))$, and p_0 is the related costate trajectory.)

It has been shown that the independence and the coercivity conditions are not only sufficient for the solution optimality: they are also sufficient for the stability of the solutions in L_∞ sense under reasonable small data perturbations [17], [19], [21]. The conditions are further proved to be also necessary for Lipschitz stability of solutions over a certain class of perturbations [7]. In the case of continuous controls, local

convergence of Euler's and related discretization methods [20], [8], [6] can be guaranteed. Furthermore, estimates of the form (2.6) are of practical interest, e.g., for analyzing the convergence of minimizing sequences [9], [11].

Consider the auxiliary functional Ψ from (2.5) with its integrand

$$R[t] \;=\; (\, H(t,x,u,S_x) \;+\; S_t \,)\;[t]\;. \tag{2.8}$$

As has been shown in [12] using $\Psi(x_0,u_0;\,S)=0$ together with the ansatz

$$S(t,x) \;=\; S_0(t) \;+\; p_0(t)^T(x-x_0(t)) \;+\; 0.5\,(x-x_0(t))^T Q(t)\,(x-x_0(t))$$

and $S_x = p_0 + Q(x-x_0)$, the function R can be expressed by

$$\begin{aligned} R[t] = \hat H(x,u,p_0,\mu_0) - \hat H(x_0,u_0,p_0,\mu_0) - \hat H_x(x_0,u_0,p_0,\mu_0)^T(x-x_0) \\ +\, 0.5\,(x-x_0)^T\dot Q(x-x_0) \;+\; (x-x_0)^T Q\,(\,f(x,u)-f(x_0,u_0)\,) \\ -\,\mu_0^T\,(\,g(x,u)-g(x_0,u_0))\,. \end{aligned} \tag{2.9}$$

For $z=(x,u)$ near $z_0=(x_0,u_0)$, the Taylor expansion yields $R = R^{(2)} + \mathrm{o}(|z-z_0|^2)$,

$$R^{(2)}[t] = 0.5 \begin{pmatrix} x-x_0 \\ u-u_0 \end{pmatrix}^T \begin{pmatrix} \hat H_{xx} + Qf_x + f_x^T Q + \dot Q & \hat H_{xu} + Qf_u \\ \hat H_{ux} + f_u^T Q & \hat H_{uu} \end{pmatrix} \begin{pmatrix} x-x_0 \\ u-u_0 \end{pmatrix}.$$

If the Hessian herein is positive definite, an estimate of type (R1) holds true with $p=2$, $\gamma=1$. In particular, the Legendre–Clebsch condition $\hat H_{uu} \succ 0$ together with a Riccati equation for Q resulting from a Schur complement approach are sufficient [32], [26] for weak local optimality then.

If we allow for more general variations in u under the restriction $x \approx x_0$, then the following decomposition is useful: $R[t] \;=\; R_1[t] \;+\; R_2[t]$, where

$$\begin{aligned} R_1[t] = \hat H(x,u_0,p_0,\mu_0) - \hat H(x_0,u_0,p_0,\mu_0) - \hat H_x(x_0,u_0,p_0,\mu_0)^T(x-x_0) \\ +\, 0.5\,(x-x_0)^T\dot Q(x-x_0) \;+\; (x-x_0)^T Q\,(f(x,u_0)-f(x_0,u_0)) \\ -\;\mu_0^T\,(g(x,u_0)-g(x_0,u_0))\,; \end{aligned} \tag{2.10}$$

$$\begin{aligned} R_2[t] = \hat H(x,u,p_0,\mu_0) - \hat H(x,u_0,p_0,\mu_0) \;+\; (x-x_0)^T Q(f(x,u)-f(x,u_0)) \\ -\;\mu_0^T\,(\,g(x,u)-g(x,u_0)\,)\,. \end{aligned}$$

The term R_1 is the variation in R w.r.t. x, and for x sufficiently close to x_0 we have

$$R_1[t] \;=\; 0.5\,(x-x_0)^T\left(\dot Q \;+\; Q\,f_x \;+\; f_x^T Q \;+\; \hat H_{xx}\right)(x-x_0) \;+\mathrm{o}(|x-x_0|^2). \tag{2.11}$$

Thus, R_1 is positive in the case that Q satisfies a certain *reduced* Riccati differential inequality. The analysis of R_2, however, essentially depends of the kind of minimum achieved in $\hat H$ at $u=u_0(t)$. In [12] and [11], typical cases with inner and boundary optima (related to the actual control set) are discussed in detail.

The problem class and examples considered in [12] and partly in [10] in general satisfy weak local optimality criteria including the Legendre–Clebsch condition. Instead, the problems considered in the present paper are linear in the control with a zero Hessian $\hat H_{uu}$ so that pointwise coercivity conditions necessarily fail. In the next section we will show how to overcome this difficulty.

3. Soft termination control problem. Bang-bang type controls are typical optimal regimes in problems, where the system dynamics are linear in control, and the control set is a box or polyhedron in R^k. Our aim is to consider problems with linear, in both state and control, state equation and given initial position. We will ask for a control which in a prescribed time gains the system as close as possible to a given final state (mostly zero). The control function herein is subject to componentwise upper and lower bounds, which for simplicity are given in normalized form.

Not always, the terminal state can be reached, for a given end-time. Since some deviation from the desired final position is allowed, this problem class is also called *soft termination control*. The model problem can be written in the form

$$(\mathrm{P}_S) \qquad \min\ J(x,u)\ =\ 0.5\,\|\,x(T)-b\,\|^2$$

$$\text{s.t. } \dot{x}(t)\ = A(t)x(t)\ +\ B(t)u(t) \quad \text{a.e. in } [0,T]; \tag{3.1}$$

$$x(0)\ =\ a\,; \tag{3.2}$$

$$|\,u_i(t)\,|\ \le 1, \quad i=1,\dots,k, \qquad \text{a.e. in } [0,T]\,. \tag{3.3}$$

It will be assumed that $a \neq b$ and that the matrix functions A and B are continuously differentiable on $[0,T]$.

The Hamilton function related to (P_S) is given by

$$H(t,x,u,p)\ =\ p^T A(t)\,x\ +\ p^T B(t)\,u\ ,$$

whereas the *augmented* Hamiltonian for $\mu_{1,2} \ge 0$ and $e=(1,1,\dots,1)^T$ reads as

$$\hat{H}(t,x,u,p,\mu)\ =\ H\ +\ \mu_1^T(u-e)\ -\ \mu_2^T(u+e)\ .$$

Notice that for the above problem the so-called independence condition for (nearly) active constraints always holds true. Thus, from Pontryagin's maximum principle we obtain the *switching function*

$$\sigma(t)\ =\ B(t)^T p(t)\ , \tag{3.4}$$

where the costate p satisfies the adjoint equation

$$\dot{p}(t)\ =\ -A(t)^T p(t), \qquad p(T)=x(T)-b,$$

and the optimal control is given by

$$u_{0,i}(t)\ \in\ \begin{cases} \{\,-sign\,\sigma_i(t)\,\} & \text{if} \quad \sigma_i(t)\ \neq\ 0, \\ [-1,+1] & \text{if} \quad \sigma_i(t)\ =\ 0, \end{cases} \qquad i=1,\dots,k\ . \tag{3.5}$$

Further, the multiplier functions μ_j suffice the relations

$$\mu_1(t)\ =\ \big(\,B(t)^T p(t)\,\big)_-\ , \qquad \mu_2(t)\ =\ \big(\,B(t)^T p(t)\,\big)_+$$

with right-hand sides denoting componentwise positive, respectively, negative, parts.

The Hesse matrix of $\hat{H}$ w.r.t. u is zero everywhere on $[0,T]$ so that the classical Legendre–Clebsch condition is not fulfilled (and this is true even when we consider its stable subspace formulation, e.g., [20], [8], or [7]). Thus, in general the usual stable weak optimality criteria are not fulfilled—a property corresponding to the obvious fact that, as a rule, the optimal controls are unstable in L_∞ under shifts of the switching

points caused by data perturbations. However, under rather mild conditions on the solution structure, quadratic estimates of type (2.6) can be proved, e.g., in $L_2 \times L_1$-norm. To this aim suppose the following.

Assumption 1. The optimal control has no singular arcs. The set of switching points $\Sigma = \{ t \in [0,T] : \exists\, i \in \{1,\dots,m\}$ with $\sigma_i(t) = 0 \}$ is finite, and $0, T$ do not belong to Σ. In particular, $\Sigma = \{ t_s : 1 \le s \le l \}$ for some $l \in N$.

Second, a restriction on the switching functions zeros is required.

Assumption 2. On $[0,T]$, the functions $\sigma_i,\ i = 1,\dots,k$, are continuously differentiable. For $I(s) = \{i : \sigma_i(t_s) = 0\}$, the term $\min_{1\le s\le l}\ \min_{i\in I(s)} |\,\dot\sigma_i(t_s)\,| = m_0$ is positive.

A first direct conclusion from Assumption 2 consists of the following invertibility result for σ.

LEMMA 3.1. *Let $\sigma(t) = B(t)^T p(t)$ satisfy Assumptions* 1 *and* 2. *If $t_s \in \Sigma$ is a point such that $\sigma_i(t_s) = 0$, then for arbitrary continuously differentiable $\hat B, \hat p$ close to B and p in C^0 sense, in every sufficiently small neighborhood of t_s the function $\hat\sigma_i(t) = (\hat B(t)^T \hat p(t))_i$ has a unique zero $\hat t_s$. Further, with the supremum norm $\|\cdot\|_\infty$ on C^0, we have*

$$\left|\,\hat t_s - t_s\right| \;=\; O(\,\|\hat B - B\|_\infty + \|\hat p - p\|_\infty)\,.$$

Proof. Consider the mapping $F_i : \ C^1(0,T;R^{n\times k}) \times C^1(0,T;R^n) \times [0,T] \ \to R$ defined by $F_i(M,\eta,t) = (M(t)^T\eta(t))_i$. This mapping is differentiable w.r.t. all components in $(M,\eta,t) = (B,p,t_s)$, and the derivative $(\partial/\partial t)\,F_i$ is continuous near this point. Further, $(\partial F_i/\partial t)(B,p,t_s) : \ R \to R$ is surjective due to Assumption 2. By the implicit function theorem in Banach spaces, the equation $F_i = 0$ near (B,p,t_s) is invertible w.r.t. t, and we end up with a first-order approximation for $\hat t_s$:

$$\hat t_s - t_s \doteq -\,(\dot\sigma_i(t_s))^{-1}\,\Big(\,(\partial F_i/\partial B)\,(\hat B - B)\;+\;(\partial F_i/\partial p)\,(\hat p - p)\Big)\;. \tag{3.6}$$

Thus, the assertion of the lemma holds true. □

The above lemma will be used in the next section for further investigation of the switching structure stability.

Now, let us reconsider the auxiliary functional Ψ from (2.5) and the related integrand $R[t]$ in (2.8). We refer to the representation of R in the form of a sum $R_1 + R_2$ given in the previous section. The approximation (2.11) for R_1 motivates consideration of the reduced Riccati equation for $Q = S_{xx}$ together with appropriate boundary restrictions (cf. [26]), which in case of problem (P_S) take the form

$$\begin{aligned} \dot Q \;+\; A^T Q \;+\; Q\,A \succeq \gamma\, I \qquad &\text{a.e.},\\ I \;-\; Q(T) \succeq 0\;. \end{aligned} \tag{3.7}$$

LEMMA 3.2. *For arbitrary $\gamma \in (0, 0.5)$, the system* (3.7) *has a bounded on $[0,T]$ solution Q satisfying $\|Q\|_\infty = O(\gamma)$.*

Proof. Consider the equality case in (3.7) with $\gamma = 0.5$ and $Q(T) = 0.5\,I$. Since the differential equation in this special case is linear w.r.t. Q, a solution $Q = Q_1 \in C^1(0,T;R^{n\times n})$ exists. Now, setting $Q = Q_\gamma = 2\,\gamma\,Q_1$, it is easy to see that Q_γ solves the inequalities (3.7) for arbitrary $\gamma \in (0, 0.5)$.
In addition, $\|Q_\gamma\|_\infty = 2\,\gamma\|Q_1\|_\infty = \mathrm{O}(\gamma)$. □

Our main result in preparing a local quadratic growth estimation in the spirit of Theorem 2.2 consists of the following statement.

LEMMA 3.3. *Let Assumptions* 1 *and* 2 *hold true. Then a matrix function* Q *and positive constants* ϵ, c *exist such that for all admisssible* (x,u) *satisfying* $(x(t),u(t)) \in \tilde{W}_{\epsilon,1}(t)$ *a.e. on* $[0,T]$,

$$\int_0^T R[t]\,dt \;\geq\; c\,\big(\,\|\,x-x_0\|_2^2 + \|\,u-u_0\|_1^2\,\big)\,. \tag{3.8}$$

Proof. Starting with $R[t] = R_1[t] + R_2[t]$ from (2.10) and the abbreviations $y = x - x_0$, $v = u - u_0$, for (P_S) we have

$$R_1 \;=\; 0.5\,y^T(\dot{Q} + A^TQ + Q\,A)\,y\,, \qquad R_2 = (p_0 + Qy)^TBv.$$

Remember that

$$\begin{aligned}
\big(B^Tp_0\big)_i > 0 \quad &\Rightarrow \quad (u_0)_i = -1 \quad \Rightarrow \qquad v_i \geq 0,\\
\big(B^Tp_0\big)_i < 0 \quad &\Rightarrow \quad (u_0)_i = +1 \quad \Rightarrow \qquad v_i \leq 0.
\end{aligned}$$

Choose $Q \equiv 0$ first. Then,

$$\int_0^T R[t]\,dt \;=\; \int_0^T p_0^TBv\,dt \;=\; \int_0^T \left(\sum_{i=1}^n |\,\sigma_i|\,\cdot\,|\,v_i|\right) dt \;\geq\; 0\,.$$

Under Assumption 1, it follows in particular that

$$\int_0^T R[t]\,dt \;=\; \Psi(x,u;S) \;>\; 0 \qquad \forall\,(x,u) \text{ such that } u \neq u_0\,. \tag{3.9}$$

Although this relation is weaker than (R1), together with appropriate boundary inequalities it already allows us to deduce strict optimality of (x_0,u_0). In order to obtain the estimate (3.8) of type (R1), let us take $Q = Q_\gamma$ from Lemma 3.2. Then we get

$$\int_0^T R_1[t]\,dt \;\geq\; \frac{\gamma}{2}\,\|\,x-x_0\|_2^2\,. \tag{3.10}$$

Further, the part R_2 can be expressed by

$$R_2[t] \;=\; \big(B^Tp_0\big)^T\,v \;+\; y^TQ\,B\,v \;\geq\; \sum_{i=1}^m \big|\,B^Tp_0\big|_i\cdot\,|\,v_i| \;-\; \big|\,y^TQ\,B\,v\,\big|\,, \tag{3.11}$$

so that for the integral over R_2 we obtain $\int_0^T R_2[t]\,dt \;\geq\; J_2 \;-\; J_1$ with

$$J_1 = \int_0^T |\,y^TQ\,B\,v\,|\,dt, \qquad J_2 = \int_0^T \sum_{i=1}^m |\,B^Tp_0|_i\,|\,v_i|\,dt. \tag{3.12}$$

From the state equation for admissible x and x_0 we know that $\dot{y} \;-\; A\,y \;=\; B\,v$ and $y(0) = 0$. If $\Phi(t)$ denotes the fundamental solution for this linear system, then

$$y(t) \;=\; \Phi(t)\int_0^t \Phi^{-1}(s)\,B(s)v(s)\,ds.$$

Consequently, y can be estimated by

$$\|y\|_\infty \le \|\Phi\|_\infty \int_0^T \|\Phi^{-1}\|_\infty \|B\|_\infty |v| \, dt = c(A,B)\|v\|_1. \tag{3.13}$$

Thus, the integral J_1 from (3.12) satisfies the inequality

$$J_1 \le c(A,B)\|Q\|_\infty \|B\|_\infty \|v\|_1^2 \le \gamma c_1 \|v\|_1^2. \tag{3.14}$$

For the second part J_2, from Assumption 2 the following property of $B^T p_0 = \sigma$ follows: For given $\delta > 0$ denote $\omega_\delta = \bigcup_{1\le s\le l}(t_s - \delta, t_s + \delta)$. Then a constant $\bar\delta$ exists such that for all $\delta \in (0, \bar\delta)$,

$$\min_i \left| \left(B^T p(t) \right)_i \right| \ge 0.5\, m_0 \delta \qquad \forall\, t \in [0,T]\backslash\omega_\delta.$$

Inserting this estimate into the formula (3.12) for J_2, we arrive at

$$J_2 \ge 0.5\, m_0 \delta \int_{[0,T]\backslash\omega_\delta} |v(t)| \, dt\,. \tag{3.15}$$

But the variation terms v are bounded in L_∞-norm by $2M = 2$, so that

$$\begin{aligned} \|v\|_1 = \int_0^T |v(t)|\, dt &= \int_{[0,T]\backslash\omega_\delta} |v(t)|\, dt + \int_{\omega_\delta} |v(t)|\, dt \\ &\le \int_{[0,T]\backslash\omega_\delta} |v(t)|\, dt + 4l\,\delta \end{aligned}$$

follows from Assumption 1. Using this relation together with (3.15), we obtain

$$J_2 \ge 0.5\, m_0 \delta \, (\|v\|_1 - 4l\delta)\,. \tag{3.16}$$

Combining now the inequalities (3.14), (3.16) for J_1 and J_2, the following estimate for $\int R_2\, dt$ results:

$$\int_0^T R_2[t]\, dt \ge 0.5\, m_0 \delta \, (\|v\|_1 - c_2\delta) - c_1\gamma \|v\|_1^2 \tag{3.17}$$

with $c_2 = 4l$ and c_1 from (3.14). We will choose δ such that

$$\delta = \min\left\{ \frac{1}{2\,c_2}, \frac{\bar\delta}{2T} \right\} \|v\|_1 =: c_3 \|v\|_1,$$

and $\gamma < \bar c = (m_0 c_3)/(8c_2)$. Then, from (3.10) and (3.17) we get

$$\int_0^T R[t]\, dt \ge \bar c\, \|v\|_1^2\,.$$

Taking into account (3.13), the desired estimate (3.8) follows immediately, e.g., with $c = 0.5\bar c \, \min\{1, c(A,B)^{-2}\}$. □

Finally, as a direct consequence of Theorem 2.2, Lemma 3.3, and (3.9), we obtain the following.

THEOREM 3.4. *Let (x_0, u_0) be an extremal point of (P_S) with the related adjoint function p, and $\sigma = B^T p$. Suppose Assumption 1 holds true. Then (x_0, u_0) is a strict strong local minimizer.*

If, in addition, Assumption 2 is fulfilled for the switching function, then for each $M > 0$, positive constants ϵ and c' exist such that

$$J(x,u) - J(x_0,u_0) \;\geq\; c' \left(\|x - x_0\|_2^2 \;+\; \|u - u_0\|_1^2 \right)$$

for all admissible (x,u) satisfying $(x(t),u(t)) \in \tilde{W}_{\epsilon,M}$ a.e. on $[0,T]$.

Remark. Since the control set in the problem under consideration is compact, the definitions of strong and of bounded-strong local optimality coincide. For the above result, however, the boundedness of u is essential.

4. Structural stability of the control. In this section, under which conditions the structure of the optimal control function is stable w.r.t. perturbations in the problem data will be analyzed. Notice that the growth condition of Theorem 3.4 can be read as a first stability result, which in particular yields the local convergence of minimizing sequences (cf. [11], [10]) for (P_S) when the approximations are taken from sufficiently small neighborhoods of the solution in $L_\infty \times L_1$, e.g. In particular, such variations may differ in the switching points and thus violate closeness conditions w.r.t. the L_∞ topology in the control component.

Denote the reference data for (P_S) by $(\bar{a}, \bar{b}, \bar{A}, \bar{B})$, and consider problems with the same type of linear dynamics but for $(a,b,A,B) \approx (\bar{a}, \bar{b}, \bar{A}, \bar{B})$ in the sense of $R^n \times R^n \times C^0(0,T;R^{n\times n}) \times C^1(0,T;R^{n\times k})$. For the perturbation analysis, the classical implicit function theorem will be used.

Let us start with the *canonical system* related to (P_S), i.e.,

$$\begin{array}{lll} \dot{x}(t) = A(t)x(t) \;+\; B(t)u(t)\,, & x(0) = a\,, & \\ \dot{p}(t) = -A(t)^T p(t)\,, & p(T) = x(T) - b\,, & \\ \sigma(t) = B(t)^T p(t)\,, & u_i(t) = -sign\,(\sigma_i(t)) & \text{for } t \notin \Sigma. \end{array} \tag{4.1}$$

Assume for the moment that besides $x(0) = a$ the initial condition for p is known, e.g., $p(0) = z \in R^n$. Then the above system can be interpreted as follows: The data set (z,a,A,B) defines $p = p(\cdot\,; z, a, A, B)$ from the adjoint equation and, subsequently, u via the switching condition. Thus, the function $x = x(\cdot\,; z, a, A, B)$ is uniquely determined from the state equation provided the control does not include singular arcs. The system (4.1) is solved if and only if

$$F(z,a,b,A,B) \;=\; x(T;z,a,A,B) \;-\; p(T;z,a,A,B) \;-\; b \;=\; 0\,. \tag{4.2}$$

If we can show that small changes in (a,b,A,B) cause only small perturbations in the solution z of (4.2) and the resulting switching functions, then the set Σ of the points of control discontinuity, in accordance to Lemma 3.1, will be stable w.r.t. (a,b,A,B). Let us start with defining fundamental solutions for the canonical equations by

$$\begin{array}{ll} \dot{\Phi} \;+\; A^T\Phi = 0\,, & \Phi(0) \;=\; I\,, \\ \dot{\Psi} \;-\; A\Psi = 0\,, & \Psi(0) \;=\; I\,. \end{array} \tag{4.3}$$

Formally, these matrix functions can be given as $\Phi(t) \;=\; \exp\{-\int_0^t A^T(s)\,ds\}$, $\Psi(t) \;=\; \exp\{\int_0^t A(s)\,ds\}$, and in particular we have $\Psi(t)^T\Phi(t) \equiv I$ (I—the unit matrix). Then

we can express $p = p(\cdot; z, a, A, B)$ and $x = x(\cdot; z, a, A, B)$ as

$$
\begin{aligned}
p(t) &= \Phi(t)\, z\,, \\
x(t) &= \Psi(t)\, a \,+\, \Psi(t) \int_0^t \Psi^{-1}(s) B(s) u(s)\, ds \\
&= \Psi(t)\, a \,-\, \Psi(t) \int_0^t \Phi^T(s) B(s)\, sign\left(B(s)^T \Phi(s)\, z \right)\, ds\,.
\end{aligned} \tag{4.4}
$$

Consequently, with the abbreviation $B(t)^T \Phi(t) = \Gamma(t)$,

$$
F(z, a, b, A, B) \,=\, \Psi(T)\, a \,-\, \Phi(T)\, z \,-\, b \,-\, \Psi(T) \int_0^T \Gamma^T(t)\, sign\left(\Gamma(t)\, z \right)\, dt\,. \tag{4.5}
$$

Notice that the dependency of F on (A, B) herein is "hidden" in the construction of the fundamental solution functions Φ, Ψ and the term $\Gamma = B^T \Phi$, which all smoothly depend on the input data. The same holds true for F then.

The smoothness of F as a function of z is less obvious. Starting from (4.2), let us rewrite F as $F_1 + F_2$, where

$$
F_1(z, a, b, A) \,=\, \Psi(T)\, a \,-\, \Phi(T)\, z \,-\, b \tag{4.6}
$$

is continuously differentiable in all variables. The second part $F_2 = F - F_1$ is given by

$$
F_2(z, A, B) \,=\, -\Psi(T) \int_0^T \Gamma^T(t)\, sign\left(\Gamma(t)\, z \right)\, dt\,.
$$

In order to analyze the differentiability of F w.r.t. z, consider $\Delta F_2 = F_2(z, \bar{A}, \bar{B}) - F_2(\bar{z}, \bar{A}, \bar{B})$:

$$
\Delta F_2 \,=\, -\,\bar{\Psi}(T) \int_0^T \bar{\Gamma}^T(t)\, \left[\, sign\left(\bar{\Gamma}(t)\, z \right) - sign\left(\bar{\Gamma}(t)\, \bar{z} \right) \right]\, dt, \tag{4.7}
$$

where $\bar{\Psi}$ and $\bar{\Gamma} = \bar{B}^T \bar{\Phi}$ are constructed in accordance to (4.3) with $A = \bar{A}$ and $B = \bar{B}$.

The difference ΔF_2 is essentially determined by the change of the switching function $\sigma = B^T p = \Gamma\, z$ for different z. Under Assumptions 1 and 2, it follows from Lemma 3.1 that for z near $\bar{z}$ the signs of $\sigma_i(t) = \left(\bar{\Gamma}(t)\, z \right)_i$ and $\bar{\sigma}_i(t) = \left(\bar{\Gamma}(t)\, \bar{z} \right)_i$ differ each from the other only near t_s and for $i \in I(s)$. In detail, if, e.g., t_i' is a zero of σ_i close to t_s and such that $t_i' < t_s$, then for $\dot{\bar{\sigma}}_i(t_s) > 0$ the difference of signs is equal to 2 on (t_i', t_s) and -2 in the case $\dot{\bar{\sigma}}_i < 0$. In general,

$$
sign\left(\sigma_i(s) \right) \,-\, sign\left(\bar{\sigma}_i(s) \right) \,=\, 2\, sign\left(\dot{\bar{\sigma}}_i(t_s) \right) \,\cdot\, sign(t_s - t_i')
$$

for all s between t_s and t_i'. Outside this interval, σ_i and $\bar{\sigma}_i$ are of the same sign. With this information, from (4.7) we obtain

$$
\begin{aligned}
\Delta F_2 &= -\bar{\Psi}(T) \int_0^T \sum_{i=1}^{n} \bar{\Gamma}_i^T(\tau)\, \left[\, sign\left(\sigma_i(\tau) \right) \,-\, sign\left(\bar{\sigma}_i(\tau) \right) \right]\, d\tau \\
&= -2\, \bar{\Psi}(T) \sum_{s=1}^{l} \sum_{i \in I(s)} sign(\dot{\bar{\sigma}}_i(t_s)) \int_{t_i'}^{t_s} \bar{\Gamma}_i^T(\tau)\, d\tau \\
&= \,2\, \bar{\Psi}(T) \sum_{s=1}^{l} \sum_{i \in I(s)} sign(\dot{\bar{\sigma}}_i(t_s))\, \bar{\Gamma}_i^T(t_s)\, \left(t_i' - \, t_s \right) \,+\, \mathrm{o}(\, |t_i' - t_s|\,)
\end{aligned}
$$

(where $\bar{\Gamma}_i$ denotes the ith row of $\bar{\Gamma} = \bar{B}^T \bar{\Phi}$, and thus $\bar{\sigma}_i(t)$ is equal to $\bar{\Gamma}_i(t) z$.)

From Lemma 3.1 it follows that the difference $(t_i' - t_s)$ can be approximated by

$$t_i' - t_s \;=\; -(\dot{\bar\sigma}_i(t_s))^{-1}\bar B^i(t_s)^T(p(t_s) - \bar p(t_s))$$

($\bar B^i$—the ith column of $\bar B$, or equal to $\partial F_i/\partial p$). Using $p = \bar\Phi\, z$, we see that

$$\begin{aligned}\Delta F_2 &= -2\,\bar\Psi(T)\sum_{s=1}^{l}\sum_{i\in I(s)} |\,\dot{\bar\sigma}_i(t_s)|^{-1}\bar\Gamma_i(t_s)^T\bar B^i(t_s)^T(p(t_s)-\bar p(t_s)) \;+\; \mathrm{o}(\,\|p-\bar p\|_\infty)\,,\\ &= -2\,\bar\Psi(T)\sum_{s=1}^{l}\sum_{i\in I(s)} |\,\dot{\bar\sigma}_i(t_s)|^{-1}\bar\Gamma_i(t_s)^T\bar\Gamma_i(t_s)(z-\bar z) \;+\; \mathrm{o}(\,|z-\bar z|\,)\,.\end{aligned}$$

Finally, with $F = F_1 + F_2$, from (4.6) and the last expression we obtain

$$\frac{\partial F}{\partial z} \;=\; -\bar\Phi(T) - 2\;\bar\Psi(T)\sum_{s=1}^{l}\sum_{i\in I(s)} (\,|\,\dot{\bar\sigma}_i(t_s)\,|\,)^{-1}\,\bar\Gamma_i(t_s)^T\bar\Gamma_i(t_s)\,. \tag{4.8}$$

Therefore, the mapping F is differentiable in all variables at $(\bar z, \bar a, \bar b, \bar A, \bar B)$, and the derivative $\partial F/\partial z$ is continuous near the given point.

After these preliminaries, we are able to formulate a structural stability result for bang-bang type optimal controls in (P_S).

THEOREM 4.1. *Let (x_0, u_0) be a solution of (P_S) without singular arcs and the related set of switching points given by $\Sigma = \{t_s :\; 1 \le s \le l\}$. Suppose that A, B are continuously differentiable in t and let Assumptions 1 and 2 of section 3 hold true. Then the switching structure of u_0 is stable in the following sense: For arbitrary positive ϵ, one can find a constant $\delta > 0$ such that for all $(\hat A, \hat B, \hat a, \hat b)$ with*

$$\|\,\hat A - A\,\|_0 + \|\,\hat B - B\,\|_{C^1} \;+\; |\hat a - a| \;+\; |\hat b - b| \;\le\; \delta,$$

the problem has a unique solution $(\hat x, \hat u)$ in a certain $L_\infty \times L_1$-neighborhood of (x_0, u_0) with $\hat u$ having the same number of switching points as u_0, and dist$\{\hat\Sigma, \Sigma\} < \epsilon$.

Remark. The norms figuring in the theorem and the proof below are the maximal over $[0,T]$ matrix norm $\|\cdot\|_0$ in C^0, and for R^n the Euclidean norm $|\cdot|$.

Proof. Consider (4.2) as an equation for z depending on the parameters $a, b \in R^n$, $A \in C^0(0,T;R^{n\times n})$, and $B \in C^1(0,T;R^{n\times k})$. The mapping F is continuously differentiable w.r.t. all variables. Moreover, by the construction of Φ and Ψ from (4.3) and from (4.8) one can see that

$$\frac{\partial F}{\partial z} \;=\; -2\,\Phi^{-T}(T)\,M,$$

where

$$M \;=\; 1/2\,\Phi(T)^T\Phi(T) + \sum_{s=1}^{l}\sum_{i\in I(s)} (\,|\,\dot\sigma_i(t_s)\,|\,)^{-1}\,\Gamma_i(t_s)^T\Gamma_i(t_s)$$

is positive definite. Thus, it represents a surjective map from R^n to R^n. From the implicit function theorem it can be deduced, first, that the equation $F(z, \hat a, \hat b, \hat A, \hat B) = 0$ for arbitrary $(\hat a, \hat b, \hat A, \hat B)$ near (a, b, A, B) has a solution $\hat z \in R^n$, which is unique in a certain neighborhood of $z\,(= p(0))$. Second, a constant $c_F > 0$ exists such that

$$|\,\hat z - z\,| \;\le\; c_F\,\Big(\,\|\,\hat A - A\,\|_0 + \|\,\hat B - B\,\|_0 \;+\; |\hat a - a| \;+\; |\hat b - b|\Big)\,.$$

Further, by $p(t) = \Phi(t)\, z$, $\sigma(t) = B(t)^T p(t) = \Gamma(t)\, z$ we obtain the C^1 error estimate

$$|\,\hat{p} - p\,\|_{C^1} + \|\,\hat{\sigma} - \sigma\,\|_{C^1} \le c_{sw} \left(\|\,\hat{A} - A\,\|_0 + \|\,\hat{B} - B\,\|_{C^1} + |\hat{a} - a| + |\hat{b} - b| \right). \tag{4.9}$$

Using Lemma 3.1 we may deduce that for every switching point t_s of u_0 there exists a locally unique switching point $\hat{t}_s$ of the perturbed solution $\hat{u}$ with

$$|\,\hat{t}_s - t_s| \le c_t \left(\|\,\hat{A} - A\,\|_0 + \|\,\hat{B} - B\,\|_{C^1} + |\hat{a} - a| + |\hat{b} - b| \right), \tag{4.10}$$

so that for sufficiently small δ the conclusion of the theorem follows. □

5. Linear time-optimal termination problem. Let us reconsider the steering problem from section 3 but now with the requirement to finish the process at $x = b$ in minimal time T:

$$(\mathrm{P}_T) \qquad \min \ T$$

$$\text{s.t. } \dot{x}(t) = A(t)x(t) + B(t)u(t) \quad \text{a.e. in } [0,T]; \tag{5.1}$$

$$x(0) = a\,, \ x(T) = b\,, \quad T > 0; \tag{5.2}$$

$$|\,u_i(t)\,| \le 1, \ i = 1, \dots, k, \qquad \text{a.e. in } [0,T]\,. \tag{5.3}$$

Consider the case that $a \neq b$, and A, B are continuously differentiable for $t \ge 0$. Further, we will always assume existence of admissible state-control pairs (i.e., the *controllability* of the system).

If $b = 0$, (P_T) is also called *hard termination control* problem.

By setting $T = \int_0^T dt$, the objective functional can be rewritten in Lagrange form. The Hamilton function then turns into

$$H(t,x,u,p) = 1 + p^T A(t)\, x + p^T B(t)\, u. \tag{5.4}$$

Consequently, the adjoint equation and transversality conditions are

$$\dot{p}(t) = -A(t)^T p(t), \qquad H[T] = 0,$$

whereas $p(0)$, $p(T)$, and the final time T are free. The switching function for u is given as before by (3.4), i.e.,

$$\sigma(t) = B(t)^T p(t), \qquad u(t) = -sign\ (\,\sigma(t)\,).$$

There are several ways to reduce (P_T) to a problem on a fixed time interval. Following Hestenes in [24] and [25], a new time variable was introduced by

$$t = T\,\tau\,, \qquad 0 \le \tau \le 1,$$

together with the extended state vector

$$y(\tau) = (y'(\tau),\, y_{n+1}) := (x(T\tau),\, T)$$

and the transformed control v with $v(\tau) := u(T\tau)$, respectively. The state equation after this transformation turns into

$$\dot{y} = \frac{dy}{d\tau} = \tilde{f}(\tau, y, v) = \begin{pmatrix} T\, f(T\tau, x, u) \\ 0 \end{pmatrix}$$

with $f(t,x,u) = A(t)x + B(t)u$, whereas the objective functional is rewritten as

$$\tilde{J}(y,v) \;=\; \int_0^1 T\, d\tau.$$

If we consider the related Hamiltonian functions, for $\tilde{p} = (p\,,\, p_{n+1})$ we get

$$\begin{aligned} H^+(\tau,y,v,\tilde{p}) &= T\, H(T\tau,x,u,p), \\ \hat{H}^+(\tau,y,v,\tilde{p},\tilde{\mu}) &= T\, \hat{H}(T\tau,x,u,p,\mu)\,, \quad \mu \;=\; \tilde{\mu}/T\,. \end{aligned} \tag{5.5}$$

From these relations one can easily reformulate the optimality conditions for a problem with free final time now (cf. [24], e.g.). In the maximum principle (PMP), the additional transversality condition w.r.t. to the free final time is usually given by $\hat{H}[T] \;=\; 0$. Equivalently, with the above state transformation we obtain a boundary value problem for the additional adjoint component, i.e.,

$$\dot{p}_{n+1} = \; -\hat{H} \;-\; T\tau\, \hat{H}_t\,, \qquad p_{n+1}(0) = p_{n+1}(1) = 0\,.$$

The (first-order) *independence* conditions for the transformed problem are equivalent to the formulations related to (P_S). For the example problem (P_T), under Assumption 1 they are automatically fulfilled. Second-order *coercivity* type conditions, however, have to be suitably modified.

Notice that in the case of linear systems dynamics as in (P_T), the Hessian w.r.t. u of $\hat{H}$ is not positive definite. In order to derive optimality conditions and obtain second-order local growth estimates for $\tilde{J} = T$, we will use again the duality approach explained in section 1 but with an appropriate *extended* dual formulation now.

Let (x_0, u_0, T) be an extremal of (P_T) for which Assumptions 1 and 2 are fulfilled, and denote by p the corresponding adjoint function. We will assume that (x, u, T') is admissible for (P_T), and that at least (x, T') are close to (x_0, T), e.g., in $C^0 \times R$ sense. Let us introduce the dual variable in the form

$$\tilde{S} \;= \tilde{S}(\tau,y) = \tilde{S}_0(\tau) \;+\; \tilde{p}_0^T(y - y_0) \;+\; 0.5(y-y_0)^T \tilde{Q}(y-y_0),$$

where $\tilde{p}_0 = \tilde{p}_0(\tau) = (p^T(T\tau), \lambda(\tau))^T$, $y_0 = y_0(\tau) = (x_0^T(T\tau), T)^T$, and

$$\tilde{Q} = \tilde{Q}(\tau) = \begin{pmatrix} Q(T\tau) & \eta(\tau) \\ \eta^T(\tau) & q(\tau) \end{pmatrix}. \tag{5.6}$$

These expressions are related to the original data as follows:

$$\begin{aligned} \tilde{S}(\tau,y) &= S_0(T\tau) + p(T\tau)^T(x - x_0) \;+\; \lambda(\tau)(T' - T) \;+\; (x-x_0)^T \eta(\tau)(T'-T) \\ &\quad + 0.5(x-x_0)^T Q(T\tau)(x-x_0) \;+\; 0.5(T'-T)^2 q(\tau), \\ \tilde{S}_y \;&= \tilde{p}_0 + \tilde{Q}(y-y_0) \;=\; \begin{pmatrix} p + Q(x-x_0) + (T'-T)\eta \\ \lambda + (x-x_0)^T\eta + (T'-T)\, q \end{pmatrix}, \\ \tilde{S}_\tau \;&= \dot{\tilde{S}}_0 + \dot{\tilde{p}}_0^T(y-y_0) \;+\; 0.5(y-y_0)^T \dot{\tilde{Q}}(y-y_0) \\ &\qquad\qquad - (\,\tilde{p}_0 + \tilde{Q}(y-y_0)\,)^T (dy_0/d\tau)\,. \end{aligned}$$

(Here and in the following the symbol "dot" ($\dot{}$) is used for derivatives w.r.t. $\tau = t/T$.) In analogy to section 2, next consider the auxiliary functional $\tilde{\Psi}$ from the *duality gap*,

$$\tilde{\Psi}(y,v;\tilde{S}) \;=\; \int_0^1 \Big[\, H^+(\tau,y,v,\tilde{S}_y) \;+\; \tilde{S}_\tau(\tau,y) \Big]\, d\tau \;=:\; \int_0^1 \tilde{R}[\tau]\, d\tau\,, \tag{5.7}$$

and the boundary term related to (R3), i.e., $\tilde{\psi}(\tilde{\xi}_1, \tilde{\xi}_2; \tilde{S}) = \tilde{S}(0, \tilde{\xi}_1) - \tilde{S}(1, \tilde{\xi}_2) - T$ with

$$\begin{aligned} \tilde{\psi}(y(0), y(1); \tilde{S}) &= \tilde{S}(0; a, T') - \tilde{S}(1; b, T') - T \\ &= 0.5 \ (\, q(0) - q(1) \,) \ (T' - T)^2, \end{aligned} \tag{5.8}$$

due to $\dot{S}_0 = -r_0 = -1$. For $\tilde{z} = (x, u, T')$ near $\tilde{z}_0 = (x_0, u_0, T)$, the Taylor formula for the integrand in (5.7) says that $\tilde{R} = \tilde{R}^{(2)} + \mathrm{o}(|\tilde{z} - \tilde{z}_0|^2)$, (cf. (2.9)), where

$$\tilde{R}^{(2)}[t] = 0.5 \begin{pmatrix} y - y_0 \\ v - v_0 \end{pmatrix}^T \begin{pmatrix} \hat{H}^+_{yy} + \tilde{Q}\tilde{f}_y + \tilde{f}_y^T\tilde{Q} + \dot{\tilde{Q}} & \hat{H}^+_{yv} + \tilde{Q}\tilde{f}_v \\ \hat{H}^+_{vy} + \tilde{f}_v^T\tilde{Q} & \hat{H}^+_{vv} \end{pmatrix} \begin{pmatrix} y - y_0 \\ v - v_0 \end{pmatrix}.$$

If the Hessian herein is positive definite for a certain matrix function $\tilde{Q}$ (satisfying $q(0) - q(1) > 0$ in addition), then the triple (x_0, u_0, T) in accordance to (5.7), (5.8), and Theorem 2.2 is a *weak* local minimizer to (P_T). Sufficient conditions of coercivity type using an *extended* Riccati differential system have been derived and discussed for the general (i.e., possibly nonlinear) case in [25]. They include, in particular, the assumption $\hat{H}^+_{vv} = T\hat{H}_{uu} \succeq 0$, which is violated in the linear case. Therefore, we will return to the integrand $\tilde{R}$ in (5.7) and analyze it for (P_T) in its extended form.

For simplicity, the following considerations are restricted to the *autonomous case*, i.e., to problems with constant matrices $A \in R^{n\times n}$ and $B \in R^{n\times m}$.

Consider the decomposition $\tilde{R}[t] = \tilde{R}_1[t] + \tilde{R}_2[t]$ induced by (2.10). In the linear autonomous case with free final time we have

$$\begin{aligned} \tilde{R}_2[\tau] &= \hat{H}^+(\tau, y, v, \tilde{p}_0, \tilde{\mu}_0) - \hat{H}^+(\tau, y, v_0, \tilde{p}_0, \tilde{\mu}_0) - \tilde{\mu}_0^T \, (g(y, v) - g(y, v_0)) \\ &\qquad + (y - y_0)^T \tilde{Q} (\tilde{f}(y, v) - \tilde{f}(y, v_0)) \\ &= \ T' \, p^T B(u - u_0) \ + \ T' \, (\, Q(x - x_0) + (T' - T)\eta \,)^T \, B(u - u_0), \end{aligned} \tag{5.9}$$

and, for (x, T') sufficiently close to the reference data (x_0, T),

$$\tilde{R}_1[\tau] \ = \ 0.5 \, (y - y_0)^T \left(\hat{H}^+_{yy} + \tilde{Q} \, \tilde{f}_y + \tilde{f}_y^T \tilde{Q} \ + \dot{\tilde{Q}} \right) (y - y_0) \ + \ \mathrm{o}(|y - y_0|^2) \tag{5.10}$$

with

$$\tilde{f}_{y'} \ = \ \begin{pmatrix} T f_x \\ 0 \end{pmatrix} \ = \ T \begin{pmatrix} A \\ 0 \end{pmatrix}, \qquad \tilde{f}_{y_{n+1}} \ = \ \begin{pmatrix} f \\ 0 \end{pmatrix} \ = \ \begin{pmatrix} Ax + Bu \\ 0 \end{pmatrix}$$

and

$$\begin{aligned} \hat{H}^+_{y'y'} = T \, \hat{H}_{xx} \ = \ 0 \, , \qquad \hat{H}^+_{y'y_{n+1}} \ = \ \hat{H}_x \ = \ A^T p \, , \\ \hat{H}^+_{y_{n+1}y_{n+1}} = 0 \, . \end{aligned}$$

Inserting these expressions together with (5.6) into (5.10), we obtain

$$\begin{aligned} \tilde{R}_1 &= 0.5 \, (y - y_0)^T \, M \, (y - y_0) \ + \ \mathrm{o}(|y - y_0|^2) \\ &= 0.5 \begin{pmatrix} x - x_0 \\ T' - T \end{pmatrix}^T \begin{pmatrix} M_1 & M_2 \\ M_2^T & M_3 \end{pmatrix} \begin{pmatrix} x - x_0 \\ T' - T \end{pmatrix} \\ &\qquad + \mathrm{o}(|x - x_0|^2 + |T' - T|^2), \end{aligned} \tag{5.11}$$

where the matrix blocs M_i of the Hessian M are given by

$$\begin{aligned} M_1 &= \dot{Q} \,+\, T\,\left(Q\,A \,+\, A^T Q\right), \\ M_2 &= \dot{\eta} \,+\, T\,A^T\eta \,+\, A^T p \,+\, Q(Ax_0 + Bu_0)\,, \\ M_3 &= \dot{q} \,+\, 2\,\eta^T(Ax_0 + Bu_0)\,. \end{aligned}$$

In order to ensure the positive definiteness of M it will be sufficient to solve appropriate linear differential equations for the components of $\tilde{Q}$, so that the only crucial point would be the boundary restriction on q (cf. (5.8)). However, as will be shown in the following, for the time-optimal termination problem (P_T) a relaxation becomes possible which can be trivially fulfilled for linear state equations.

LEMMA 5.1. *For every $\gamma \in (0,1)$, the matrix differential equation $M = \gamma I$, i.e.,*

$$\dot{Q} \,+\, T\,\left(Q\,A \,+\, A^T Q\right) = \gamma\, I\,, \tag{5.12}$$

$$\dot{\eta} \,+\, T\,A^T\eta = -A^T p \,-\, Q(Ax_0 + Bu_0)\,, \tag{5.13}$$

$$\dot{q} = \gamma \,-\, 2\,\eta^T(Ax_0 + Bu_0)\,, \tag{5.14}$$

has a bounded on $[0,T]$ solution $\tilde{Q} = \left(\begin{smallmatrix} Q & \eta \\ \eta^T & q \end{smallmatrix}\right)$. In particular, the solution can be chosen such that for $\gamma \to 0$ the components η and q are uniformly bounded, and Q satisfies $\|Q\|_\infty = O(\gamma)$.

Proof. Denote by Q_1 the solution of the initial value problem for the linear equation (5.12) with $\gamma = 1$ and $Q(0) = I$. Setting next $Q = \gamma\, Q_1$, we can solve the (linear) equations for η and q, e.g., with $\eta(0) = 0$ and $q(0) = 0$ then. The solutions satisfy

$$\begin{aligned} \|Q\|_\infty &= \gamma\,\|Q_1\|_\infty \;=\; \mathrm{O}(\gamma)\,, \\ \|\eta\|_\infty &\le c_1\|p\|_1 + \mathrm{O}(\gamma) \;\le\; M_\eta\,, \\ \|q\|_\infty &\le c_2\|\eta\|_1 + \gamma \;\le\; M_q \end{aligned}$$

for positive constants $c_{1,2}$ and M_η, M_q not depending of γ for $\gamma \in [0,1]$. Thus, the conclusion follows. □

LEMMA 5.2. *Let the solution (x_0, u_0, T) of (P_T) satisfy Assumption 1. Then, for every $\gamma \in (0,1)$, positive constants ϵ, $\tilde{c}_1$ exist such that*

$$\tilde{R}_1[\tau] \;\ge\; 0.25\,\gamma\,\left(\,|\,x(T'\tau) - \,x_0(T\,\tau)\,|^2 \,+\, |T' - T|^2\right) \tag{5.15}$$

holds together with $|\psi| \;\le\; \tilde{c}_1|T' - T|^2$ for arbitrary admissible (x, u, T') satisfying $\|x' - x_0\|_\infty + |T' - T| \,<\, \epsilon$ (where x' stands for the transformed state $x'(t) = x(T't/T)$ on $[0,T]$).

Proof. Let the matrix parameter $\tilde{S}_{yy} = \tilde{Q}$ in (5.6) be chosen in accordance to Lemma 5.1. Then the proof of the first part of this lemma follows immediately from (5.11) and Lemma 5.1. The second part is a direct consequence of relation (5.8) and the estimate for q from the above lemma. □

LEMMA 5.3. *Suppose that (x_0, u_0, T) suffices the Assumptions 1 and 2 of section 3. Further, let $\tilde{Q}$ be determined from Lemma 5.1, where γ is chosen sufficiently small. Then there exist positive constants ϵ', $\tilde{c}_2$, and $\tilde{c}_3$ such that*

$$\int_0^1 \tilde{R}_2[\tau]\,d\tau \;\ge\; \tilde{c}_2\|u' - \,u_0\|_1^2 \,-\, \tilde{c}_3|T' - T|^2 \tag{5.16}$$

for arbitrary admissible (x, u, T') *with* $\|x' - x_0\|_\infty + |T' - T| < \epsilon'$ *(where* $x'(t) = x(T't/T)$ *and* $u'(t) = u(T't/T)$ *are the transformed state and control, resp.).*

Proof. The proof follows the line from Lemma 3.3 but uses the extended dual formulation with $\tilde{S}$, respectively, $\tilde{Q}$.

Under Assumption 2, from the expression (5.9) for $\tilde{R}_2[t]$ in analogy to (3.17) we obtain the following estimate for sufficiently small $\delta > 0$:

$$(T')^{-1} \int_0^1 \tilde{R}_2[\tau]\, d\tau \geq 0.5\, m_0\, \delta\ (\, \|v - v_0\|_1 - c_1\, \delta\,) \\ - c_2 \|Q\|_\infty \|B\| \cdot \|v - v_0\|_1^2 \ - \ \|\eta^T B\|_\infty \|v - v_0\|_1\, |T' - T|\,.$$

Since $B = \text{const}$ and $\|\eta\|_\infty \leq M_\eta$ for arbitrary $\gamma \in (0, 1)$ (cf. Lemma 5.1), the last term can be estimated by

$$\|\eta^T B\|_\infty \|v - v_0\|_1\, |T' - T| \ \leq\ c_3 \left(\, \rho\, \|v - v_0\|_1^2 + (4\rho)^{-1} |T' - T|^2 \,\right)$$

with a certain constant c_3 and arbitrary positive ρ. Moreover,

$$\|Q\|_\infty \|B\| \cdot \|v - v_0\|_1^2 \ \leq\ c_4 \gamma \|v - v_0\|_1^2$$

for small positive γ. Thus one can choose, first, $\delta = c' \|v - v_0\|_1$ with sufficiently small c', and subsequently γ and ρ small enough so that the estimate (5.16) follows. ∎

Our final result is given by the next theorem.

THEOREM 5.4. *Let* (x_0, u_0, T) *be an extremal point of* (P_T) *with constant* A *and* B*, and assume Assumptions* 1 *and* 2 *hold true. Then the triple* (x_0, u_0, T) *is a strict local minimizer, and for some positive constants* $\tilde{c}$ *and* $\tilde{\epsilon}$ *the estimate*

$$T' - T \ =\ J(x, u) - J(x_0, u_0) \ \geq\ \tilde{c}\, \left(\, \|x' - x_0\|_2^2 + \|u' - u_0\|_1^2 \,\right)$$

is valid for all admissible (x, u, T') *with* $\|x' - x_0\|_\infty + |T' - T| < \tilde{\epsilon}$ *(with the notations* $x'(t) = x(T't/T)$ *and* $u'(t) = u(T't/T)$ *on* $[0, T]$*, resp.).*

Proof. Let (x, u) be an arbitrary admissible state-control pair corresponding to the final time $T' > 0$, and denote the related extended state and control functions by y, respectively, v. First, consider the functional $\tilde{J}(y, v) = \int_0^1 T' d\tau$ reformulated in terms of $\tilde{S}$ and H^+ from (5.5) (see also (2.4)):

$$\begin{aligned} \tilde{J}(y, v) &= \int_0^1 H^+(\tau, y, v, \tilde{S}_y(\tau, y))\, d\tau \ - \ \int_0^1 \tilde{S}_y(\tau, y)^T \tilde{f}(\tau, y, v)\, d\tau \\ &= \int_0^1 \left(\, H^+(\tau, y, v, \tilde{S}_y) \ + \ \tilde{S}_\tau \,\right)\, d\tau \ + \ \tilde{S}(0, y(0)) - \tilde{S}(1, y(1)). \end{aligned}$$

Denoting by (y_0, v_0) the extended trajectories related to (x_0, u_0, T), from (5.7), (5.8) we deduce

$$T' - T \ =\ \tilde{J}(y, v) \ - \ \tilde{J}(y_0, v_0) \ =\ \tilde{\Psi}(y, v; \tilde{S}) \ + \ \tilde{\psi}(y(0), y(1); \tilde{S})\,.$$

Using Lemma 5.2 and 5.3, for sufficiently small $\gamma > 0$ we obtain the estimate

$$\begin{aligned} T' - T &\geq \int_0^1 \tilde{R}[\tau]\, d\tau \ + \ \tilde{\psi}(y(0), y(1), \tilde{S}) \\ &\geq\ \frac{\gamma}{4}\, \|x' - x_0\|_2^2 + \tilde{c}_2 \|u' - u_0\|_1^2 \ - \ (\tilde{c}_1 + \tilde{c}_3)\, |T' - T|^2 \end{aligned} \tag{5.17}$$

(with the notations x', u' used in the statement of the theorem).

Assume for the moment that $T' < T$: From (5.17) it follows that

$$(T' - T)\ (\,1 + (\tilde{c}_1 + \tilde{c}_3)(T' - T)\,)\ \geq\ 0\,.$$

In the case that $|T' - T| < \tilde{\epsilon} \leq 0.5(\tilde{c}_1 + \tilde{c}_3)^{-1}$, this leads to a contradiction. Consequently, Assumptions 1 and 2 are sufficient for the local optimality of (x_0, u_0) and the related final time T.

With $T' \geq T$, from $|T' - T| < \tilde{\epsilon}$ chosen above we now arrive at

$$(T' - T)\ +\ (\tilde{c}_1 + \tilde{c}_3)(T' - T)^2\ \leq\ (3/2)\,(T' - T)\,,$$

so that finally for some positive $\tilde{c}$ we end up with

$$T' - T\ \geq\ \frac{\gamma}{6}\,\|x' - x_0\|_2^2 + (2\tilde{c}_2/3)\,\|u' - u_0\|_1^2\ \geq\ \tilde{c}\,\left(\,\|x' - x_0\|_2^2 +\ \|u' - u_0\|_1^2\,\right),$$

i.e., the desired result. □

We will accomplish Theorem 5.4 by a sensitivity result concerning the optimal time.

LEMMA 5.5. *Let $(x_0, u_0; T)$ be a solution of (P_T) satisfying Assumptions 1 and 2. Further, let (x, u) be an admissible pair corresponding to the final time T' (with $x(T') = b$ in particular). Then positive constants ρ and c exist such that*

$$|\,T' - T\,|\ \leq\ c\,\|u - u_0\|_1 \qquad \forall\, T' \ \text{ with } \ \ |\,T' - T\,| < \rho\,.$$

Remark. The L_1-norm on the right-hand side is related to the interval $[0, T']$, where for $T' > T$ the function u_0 is continued to $[T, T']$ as a constant. The function x_0 then denotes the corresponding solution of the state equation on the larger interval $[0, T']$.

Proof. Notice first that from the transversality condition $H[T] = 1 + p^T \dot{x}_0 = 0$ it follows that $\dot{x}_0(T) \neq 0$.

Let i be an index with $|\dot{x}_{0,i}(T)| = \max_j\{|\dot{x}_{0,j}(T)|\} = m > 0$. Then, in a sufficiently small neighborhood $(T - \rho, T + \rho)$ of T, $\dot{x}_{0,i}$ does not change its sign, and the estimate $|\dot{x}_{0,i}(t)| \geq m/2$ holds true. Consequently, for $0 < |t - T| < \rho$,

$$|\,x_{0,i}(t)\ - x_{0,i}(T)| \geq\ (m/2)\,|t - T|\,.$$

Taking in particular $t = T'$, with $x_{0,i}(T) = x_i(T') = b_i$ and $c_m = 2/m$ we get

$$|T' - T|\ \leq\ c_m\ |\,x_{0,i}(T') - x_{0,i}(T)\,|\ =\ c_m\ |\,x_{0,i}(T') - x_i(T')\,|\,.$$

Since the state equation is linear, we arrive at the assertion

$$|T' - T|\ \leq\ c_m\,\|\,x - x_0\,\|_\infty \leq\ c(A, B, m)\|\,u - u_0\,\|_1\,. \qquad \square$$

6. Test example. This section is devoted to the stability analysis of *hard* and of *soft* termination control problems for a simple two-dimensional chain (cf. [30]). It will be shown that failures in the stability assumptions may cause serious instabilities of the solution structure even in elementary model cases.

Consider a particular problem of type (P),

$$(\mathrm{P}^0) \qquad \min\ 0.5\,\|x(T) - b\|^2 \quad \text{s.t.} \quad \dot{x}_1(t)\ =\ x_2(t)\,,\ \ \dot{x}_2(t)\ =\ u(t) \quad \text{a.e. in } [0, T],$$
$$x(0)\ =\ a\,; \qquad |\,u(t)\,|\ \leq 1 \qquad \text{a.e. in } [0, T].$$

(This problem is also known as the rocket car problem, where x_1 stands for the position, x_2 for the velocity, and u for the acceleration of the vehicle.) We will restrict ourselves to the case that a_2, b_2 are positive, and $b_1 > a_1$. According to the results of section 4, the adjoint equation together with the transversality condition take the form

$$\dot{p}_1 \equiv 0\,, \quad \dot{p}_2 = -p_1, \quad p(T) \,=\, x(T) - b,$$

so that with $p_1(0) = z_1,\ p_2(0) = z_2$ the *switching function* reads as

$$\sigma(t) \,= p_2(t) \,=\, -z_1 t \,+\, z_2\,.$$

Therefore we have $t_s = z_2/z_1$ whenever $z_1 \neq 0$, i.e., the extremal trajectories cannot have more than one switching point on $[0, T]$.

Consider the solution representation (4.4) from section 4: The fundamental solutions to the equations (4.3) are given by

$$\Psi(t) \,=\, \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}, \qquad \Phi(t) \,=\, \begin{pmatrix} 1 & 0 \\ -t & 1 \end{pmatrix}$$

so that $\Gamma(t) = (\,-t\ \ 1\,)$. Consequently,

$$\begin{aligned} p_1(t) &= z_1\,, \qquad p_2(t) \,=\, z_2 - z_1 t\,, \\ x_1(t) &= a_1 + a_2 t \,-\, \int_0^t sign\,(\,z_2 - z_1 s\,)\,\cdot\,(t-s)\,ds\,, \\ x_2(t) &= a_2 \,-\, \int_0^t sign\,(\,z_2 - z_1 s\,)\,ds\,. \end{aligned}$$

Thus, any discontinuous nonsingular extremal belongs to one of the following types:

Type 1: $\quad z_1 < 0\,,\ \ z_2 < 0 \quad$ or $\quad u_0 = \begin{cases} +1 & \text{for } 0 \le t < t_s, \\ -1 & \text{for } t_s \le t \le T. \end{cases}$

In the first part, the object moves with maximal acceleration, and the trajectory is part of a parabola $P_1 :\ x_1 = 0.5x_2^2 + \alpha$ in the phase plane, whereas $x \in P_2 :\ x_1 = -0.5x_2^2 + \beta$ (with appropriately chosen α, β) in the second (retardation) part. Denoting the switching point by x^S, the arcs may be written as

$$\begin{aligned} P_1 :\quad & x_1(t) = x_1^S - x_2^S(t_s - t) + 0.5(t - t_s)^2, \\ & x_2(t) = x_2^S - (t_s - t), \qquad 0 \le t \le t_s\,; \\ P_2 :\quad & x_1(t) = x_1^S + x_2^S(t - t_s) - 0.5(t - t_s)^2, \\ & x_2(t) = x_2^S - (t - t_s), \qquad t_s \le t \le T\,. \end{aligned} \tag{6.1}$$

Notice that $z_1 \equiv p_1(t)$ with $p_1(T) = x_1(T) - b_1$ so that $x_1(T) < b_1$ for this situation.

Type 2: $\quad z_1 > 0\,,\ \ z_2 > 0 \quad$ or $\quad u_0 = \begin{cases} -1 & \text{for } 0 \le t < t_s, \\ +1 & \text{for } t_s \le t \le T. \end{cases}$

The switching structure says that, in the phase plane, the points move with retardation along $x_1 = -0.5x_2^2 + \beta'$ in the first part and accelerating along $x_1 = +0.5x_2^2 + \alpha'$ in the second part. The condition $0 < z_1 \equiv p_1(t)$ says that $x_1(T) > b_1$ holds true for the second type.

In both cases, one can find explicit formulas for $p,\ x$, and the function F in (4.5). Indeed, for type 2, e.g., we have for $t > t_s = z_2/z_1$ that

$$
\begin{aligned}
x_1(t) &= a_1 + a_2 t + \frac{t^2}{2} - \frac{2tz_2}{z_1} + \left(\frac{z_2}{z_1}\right)^2, \qquad p_1(t) \;=\; z_1,\\
x_2(t) &= a_2 + t - \frac{2z_2}{z_1}, \qquad\qquad\qquad\quad p_2(t) \;=\; z_2 - tz_1\,.
\end{aligned}
\tag{6.2}
$$

From the first equation taken at $t = T$ we see that $t_s = z_2/z_1$ satisfies

$$
T(a_2 - t_s) \;=\; x_1(T) - a_1 - \frac{t_s^2}{2} - \frac{(T - t_s)^2}{2} \;\le\; 0
$$

due to $|u| \le 1$, so that $t_s \ge a_2$ follows. Taking into account (6.2), we see that $x_2(t_s) \le 0$. Notice that thus the type 2 trajectories switch to the second parabola not in the first point of intersection (which is in the upper half plane) but only when these curves intersect in the lower part of the phase plane. This effect had been earlier discussed for the time-optimal case; cf., e.g., [30].

Using (6.2), the Jacobian of $F = x(T) - p(T) - b$ now can be expressed as

$$
\frac{\partial F}{\partial z} \;=\; -\begin{pmatrix} 1 & 0 \\ -T & 1 \end{pmatrix} - \frac{2}{z_1^3}\begin{pmatrix} z_2^2 - Tz_1z_2 & Tz_1^2 - z_1z_2 \\ -z_1z_2 & z_1^2 \end{pmatrix}
$$

(and the same result may be obtained from (4.8) with $|\dot\sigma(t_s)| = |\,z_1| = -z_1$ and $t_s = z_2/z_1$).

It follows from the proof of Theorem 4.1 that $\partial F/\partial z$ is a regular matrix. In the example case, it is easy to confirm this fact by calculating the determinant:

$$
\left|\frac{\partial F}{\partial z}\right| \;=\; 1 \;+\; \frac{2}{|z_1|}\left(1 \;+\; (T - z_2/z_1)^2\right) \;>\; 0\,.
$$

Therefore, any solution of the structural type 2 is a stable solution in the sense of section 4. Moreover, Assumptions 1 and 2 hold true for such trajectories so that Theorem 3.4 can be applied to show the strong local optimality.

Analogous arguments show the stability of extremal trajectories of type 1.

In order to find locally optimal solutions of (P^0), let us first consider the related time-optimal problem

$$
\begin{aligned}
(\mathrm{P}^0_T) \qquad \min\ T \qquad \text{s.t.}\quad & \dot x_1(t) \;=\; x_2(t)\,,\ \ \dot x_2(t) \;=\; u(t) \quad \text{a.e. in } [0,T];\\
& x(0) \;=\; a\,,\ \ x(T) = b, \quad |\,u(t)\,| \le 1 \quad \text{a.e. in } [0,T]\,.
\end{aligned}
$$

As before, we will consider extremals without singular arcs, i.e., trajectories which are synthesized from parabolas of the type P_1 or P_2, respectively. Depending on the end points localization we have the following.

If $b_1 - a_1 > l = 0.5\,|\,a_2^2 - b_2^2|$, then the point b is attainable with trajectories of the structural type 1. If, in addition, $b_1 - a_1 < L = 0.5\,(\,a_2^2 + b_2^2)$, then the target can be reached by type 2 curves too, so that two extremals exist. We denote the corresponding time values by T^* and T^+. The type 1 extremal with the higher velocity components $x_2(t)$ is the global minimizer of (P^0_T). (For a detailed analysis see [30].) Due to $x_1^s \in (a_1, b_1)$, $x_2^s \ge \max\{a_2, b_2\} > 0$ we have $p_1 \equiv z_1 = -1/x_2(t_s^*) < 0$ and $\dot\sigma(t_s^*) = z_1 < 0$, so that this solution satisfies all assumptions of Theorems 5.4 and 4.1 and the minimum is locally strong and stable.

For the second extremal, where, in the phase plane, the points move under retardation in the first part and with acceleration in the second part, we have $p_1 \equiv z_1 = -1/x_2(t_s^+) > 0$, i.e., $\dot\sigma(t_s^+) = z_1 > 0$ in particular. Therefore, again we can apply the theory of section 5, respectively, 4 to see that this suboptimal extremal also gives a strict locally strong minimizer with stable switching structure.

Let us return now to problem (P^0) with given end-time T. We will mainly consider T near the local optimizers T^* or T^+ and, without loss of generality, assume that $T > |\,b_2 - a_2|$.

If the time parameter T is smaller than T^*, the target point b cannot be reached. The solution can be constructed as a trajectory of type 1 by finding the switching time t_s. Using for the trajectories analogous formulas as in (6.2), and abbreviating $r = T - t_s$, we get

$$J(x,u) \;=\; \frac{1}{2}\,\|\,x(T) - b\,\|^2 \;=\; \frac{1}{2}\,\big(\,(c_1 - r^2)^2 + (c_2 - 2r)^2\big) \;=:\; 2\,\phi(r),$$

with $c_1 = T^2/2 + a_2T + a_1 - b_1$, $c_2 = T + a_2 - b_2 > 0$. Since $\lim_{r\to\infty}\phi(r) = +\infty$ and $\phi'(0) = -c_2 < 0$, a positive minimum point r_s exists. From $\phi'(r_s) = 0$ the representation $r_s = -(c_2 - 2r_s)/(c_1 - r_s^2)$ follows. Using that $(c_1 - r_s^2) = x_1(T) - b_1 = z_1 < 0$ holds true for type 1 solutions, we get

$$r_s \;<\; 0.5c_2 = \;0.5(T \;+\; (b_2 - a_2)) \;<\; T\,.$$

Furthermore, the relation $t_s \;>\; 0.5\,(T \;+\; (b_2 - a_2))$ is characteristic for the solution trajectory in case $T < T^*$.

Consider next the situation $T^* \ll T < T^+$: naturally, we will ask for a solution of type 2 now. The approach used above leads to the auxiliary problem of minimizing $\phi(r)$ from

$$J(x,u) \;=\; \frac{1}{2}\,\big(\,(\hat c_1 + r^2)^2 + (2r - \hat c_2)^2\big) \;\;=:\; 2\hat\phi(r)$$

with the notations $\hat c_1 = \,-0.5\,T^2 + a_2T + a_1 - b_1$, $\hat c_2 = \,T - a_2 + b_2$, and $r = T - t_s$. For a minimizer t_s from $(0,T)$ in analogy to case 1 it can be proved that

$$t_s \;>\; 0.5\,(T \;-\; (b_2 - a_2)).$$

In both cases discussed so far, due to $T < \bar T \in \{T^*, T^+\}$, the condition $\dot\sigma(t_s) = z_1 \neq 0$ is fulfilled so that the solutions are stable in their structure in the sense of Theorem 4.1. Moreover, they are strict strong local minimizers of (P^0) satisfying the local quadratic growth estimate from Theorem 3.4. The situation changes when we consider $T = T^*$ or $T = T^+$: In this case, the solution of (P^0) terminates in b, and we have $z_1 = p_1 \equiv p_1(\bar T) = x_1(\bar T) - b_1 = 0$ so that in fact we have to do with a (completely) singular optimal control solution: $\sigma(t) \equiv 0$. Due to the coincidence with the time-optimal case, however, the solution is uniquely determined and represents the limit of solutions for $T \uparrow \bar T$.

The singularity mentioned for (P^0) with $T = \bar T$ causes instabilities in the behavior of the solutions when the problem data are changed. To see this, consider, e.g., $T = T^* + \delta$ with small positive δ: obviously, the optimal value of the objective function in the soft termination problem is zero (which means that the final state position $x(T) = b$ is attainable). One way for constructing optimal trajectories in the case $T = T^* + \delta$ consists of the following approach.

Consider, first, the time-optimal path along the parabolas P_1 on $[0, t_s]$ and along P_2 for $(t_s, T^*]$; cf. (6.1). On P_1, let a point $x^H = x(t_H)$ be chosen such that $v_H =$

$x_2^H > \max\{a_2, b_2\}$, and $t_H < t_s$. The point with the same x_2-component on P_2 denote by x^R. With $d = t_s - t_H = x_2^S - x_2^H$, its distance to x^H can be calculated from (6.1) as

$$L \;=\; x_1^R \;-\; x_1^H \;=\; 2x_2^S d \;-\; d^2\;.$$

Assume that the point moves for $0 \le t \le t_H$ with maximal acceleration along P_1, but then with constant velocity v_H over a distance $l \in (0, L)$. When the motion is then continued with $u = 1$, the point follows a parabola P_1' determined by

$$x_1 \;=\; 0.5\,(x_2)^2 + \; x_1^S - 0.5\,(x_2^S)^2 \;+\; l, \tag{6.3}$$

which meets P_2 at a point x^B with $x_2^H < x_2^B = v_B < x_2^S$. From this point on we set $u = -1$ and, following the parabola P_2, achieve b.

The time delay caused by breaking the acceleration and holding the velocity v_H on a certain interval of length l is given by

$$\Delta_H(l) \;=\; l/v_H \;-\; 2\,(x_2^S - x_2^B).$$

Denoting $d_B = x_2^S - x_2^B$ $(= t_B - t_s$ from the time-optimal trajectory), we see from (6.1) and (6.3) that

$$x_1^B - x_1^S \;= 0.5\,l \;=\; x_2^S d_B - 0.5\,d_B^2\,,$$

consequently,

$$\Delta_H(l) \ge\; l/v_B \;-\; 2\,d_B \;=\; \frac{l - 2x_2^S d_B + 2d_B^2}{x_2^S - d_B} \;=\; \frac{d_B^2}{v_B} \;>\; 0.$$

On the other hand, for $d = x_2^S - x_2^H$ it follows analogously that

$$\Delta_H(l) \le \frac{L}{v_H} - 2\,d \;=\; \frac{d^2}{v_H} \;=:\; D_H.$$

Summing up, we see that in dependence of the choice of x^H and the length l of the singular arc with $u \equiv 0$, a small but arbitrary time delay $\Delta_H(l) \in (0, D_H)$ can be realized. Moreover, this technique can be repeatedly applied for finitely or even countably many points on the remaining acceleration parabolas of type P_1, respectively, P_1', before their switching point to the parabola P_2 is reached. We have only to make sure that the sum of the individual time delays equals $\delta = T - T^*$.

The arguments show that near the time parameter $T = T^*$, where the stability condition contained in Assumption 2 fails, for $T^* < T \ll T^+$ one can observe bifurcations of the optimal solution including essential changes in the structure of the optimal control behavior. The same qualitative result can be obtained (by a slightly modified construction) for T near the second local optimizer T^+ satisfying $T > T^+$.

7. Conclusion. The stability analysis of bang-bang control regimes is of practical as well as of theoretical interest. A thorough comparison of the different approaches appearing in the literature could lead to better insights and should be a field for further investigations. The benefit for constructing robust and efficient numerical algorithms is another, widely open question.

In addition to the remarks in the introduction, some comments have to be added concerning the local growth results in Theorem 3.4 and Theorem 5.4: as it was pointed out by one of the referees, the estimate in Theorem 3.4 is a consequence of Theorem 5.2, [27]. Further, the statement of Theorem 5.4 could be also obtained from Theorem 13.1, [27], and the results on linear time-optimal problems from [29].

Acknowledgments. The author is very grateful to the anonymous referees for their detailed and helpful remarks. In particular, they have led to substantial improvements in the presentation of section 2 and of some of the proofs in sections 3 and 5.

REFERENCES

[1] A. A. Agrachev and R. V. Gamkrelidze, *Symplectic geometry for optimal control*, in Nonlinear Controllability and Optimal Control, H. J. Sussmann, ed., Marcel Dekker, New York, 1990, pp. 727–785.

[2] A. A. Agrachev, G. Stefani, and P. L. Zezza, *Strong minima in optimal control*, Proc. Steklov Inst. Math., 220 (1998), pp. 4–22.

[3] A. A. Agrachev, G. Stefani, and P. L. Zezza, *Symplectic methods for strong local optimality in the bang-bang case*, in Contemporary Trends in Nonlinear Geometric Control Theory and Its Applications, Mexico City, 2000, World Scientific, River Edge, NJ, 2002, pp. 169–181.

[4] A. E. Bryson, Jr., and Y. Ho, *Applied Optimal Control. Optimization, Estimation, and Control*, Hemisphere, New York, 1975.

[5] F. H. Clarke and V. Zeidan, *Sufficiency and the Jacobi condition in the calculus of variation*, Canad. J. Math., 38 (1986), pp. 1199–1209.

[6] A. L. Dontchev and W. W. Hager, *The Euler approximation in state constrained optimal control*, Math. Comp., 70 (2001), pp. 173–203.

[7] A. L. Dontchev and K. Malanowski, *A characterization of Lipschitzian stability in optimal control*, in Calculus of Variations and Optimal Control, Haifa, 1998, Chapman and Hall/CRC Press, Boca Raton, FL, 2001, pp. 62–76.

[8] U. Felgenhauer, *Diskretisierung von Steuerungsproblemen unter stabilen Optimalitätsbedingungen*, Habilitation thesis, Brandenburgische Technische Universität Cottbus, 1999.

[9] U. Felgenhauer, *On smoothness properties and approximability of optimal control functions*, in Optimization with Data Perturbation II, Ann. Oper. Res. 101, D. Ward, D. Klatte, and J. Rückmann, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001, pp. 23–42.

[10] U. Felgenhauer, *Stability and local growth near bounded-strong local optimal controls*, in System Modelling and Optimization XX, Proceedings of the 20th IFIP TC7 Conference, Trier 2001, Kluwer Academic Publishers, Dordrecht, The Netherlands, to appear.

[11] U. Felgenhauer, *Structural properties and approximation of optimal controls*, Nonlinear Anal., 47 (2001), pp. 1869–1880.

[12] U. Felgenhauer, *Weak and strong optimality conditions for constrained control problems with discontinuous control*, J. Optim. Theory Appl., 110 (2001), pp. 361–387.

[13] R. Klötzler, *On a general conception of duality in optimal control*, in Lecture Notes in Math. 703, Springer–Verlag, New York, 1979, pp. 189–196.

[14] R. Klötzler and S. Pickenhain, *Pontryagin's maximum principle for multidimensional control problems*, in Internat. Ser. Numer. Math. 111, Birkhäuser, Basel, 1993, pp. 21–30.

[15] U. Ledzewicz and H. Schaettler, *High-order approximations for abnormal bang-bang extremals*, in Systems Modelling and Optimization, Detroit, 1997, Chapman & Hall/CRC Res. Notes Math. 396, Chapman and Hall/CRC Press, Boca Raton, FL, 1999, pp. 126–134.

[16] K. Malanowski, *Two-norm approach in stability and sensitivity analysis of optimization and optimal control problems*, Adv. Math. Sci. Appl., 2 (1993), pp. 397–443.

[17] K. Malanowski, *Stability and sensitivity analysis of solutions to infinite-dimensional optimization problems*, in Proceedings of the 16th IFIP-TC7 Conference on System Modelling and Optimization, Lecture Notes in Control and Inform. Sci. 197, J. Henry and J.-P. Yvon, eds., Springer–Verlag, London, 1994, pp. 109–127.

[18] K. Malanowski, *Stability and sensitivity analysis of solutions to nonlinear optimal control problems*, Appl. Math. Optim., 32 (1995), pp. 111–141.

[19] K. Malanowski, *Stability analysis of solutions to parametric optimal control problems*, in Proceedings of the IV. Conference on Parametric Optimization and Related Topics, Enschede, 1995, Approx. Optim., J. Guddat, H. T. Jongen, F. Nožička, G. Still, and F. Twilt, eds., Lang, Frankfurt, 1996, pp. 227–244.

[20] K. Malanowski, C. Büskens, and H. Maurer, *Convergence of approximations to nonlinear control problems*, in Mathematical Programming with Data Perturbation, Lecture Notes in Pure and Appl. Math. 195, A. V. Fiacco, ed., Marcel Dekker, New York, 1997, pp. 253–284.

[21] K. Malanowski and H. Maurer, *Sensitivity analysis for parametric optimal control problems with control-state constraints*, Comput. Optim. Appl., 5 (1996), pp. 253–283.

[22] K. Malanowski and H. Maurer, *Sensitivity analysis for optimal control problems subject to higher order state constraints*, in Optimization with Data Perturbation II, Ann. Oper. Res 101, D. Ward, D. Klatte, and J. Rückmann, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001, pp. 43–73.

[23] K. Malanowski and H. Maurer, *Sensitivity analysis for state constrained optimal control problems*, Discrete Contin. Dynam. Systems, 4 (1998), pp. 241–272.

[24] H. Maurer, *Second order sufficient conditions for control problems with free final time*, in Proceedings of the 3rd European Control Conference, Rome, 1995, A. Isidori et al., eds., 1995, pp. 3602–3606.

[25] H. Maurer and H. J. Oberle, *Second order sufficient conditions for optimal control problems with free final time: The Riccati approach*, SIAM J. Control Optim., 41 (2002), pp. 380–403.

[26] H. Maurer and S. Pickenhain, *Second order sufficient conditions for optimal control problems with mixed control-state constraints*, J. Optim. Theory Appl., 86 (1995), pp. 649–667.

[27] A. A. Milyutin and N. P. Osmolovskii, *Calculus of Variations and Optimal Control*, Amer. Math. Soc., Providence, RI, 1998.

[28] J. Noble and H. Schaettler, *Sufficient conditions for relative minima of broken extremals in optimal control theory*, J. Math. Anal. Appl., 269 (2002), pp. 98–128.

[29] N. P. Osmolovskii, *Quadratic conditions for nonsingular extremals in optimal control (a theoretical treatment)*, Russian J. Math. Phys., 2 (1995), pp. 487–512.

[30] N. P. Osmolovskii, *Quadratic conditions for nonsingular extremals in optimal control (examples)*, Russian J. Math. Phys., 5 (1998), pp. 373–388.

[31] N. P. Osmolovskii, *Second-order conditions for broken extremals*, in Calculus of Variations and Optimal Control, Haifa, 1998, Chapman and Hall/CRC Press, Boca Raton, FL, 2001, pp. 198–216.

[32] S. Pickenhain, *Sufficiency conditions for weak local minima in multidimensional optimal control problems with mixed control-state restrictions*, Z. Anal. Anwendungen, 11 (1992), pp. 559–568.

[33] S. Pickenhain, *Duality in optimal control with first order differential equations*, in Encyclopedia of Optimization, Vol. I, C. A. Floudas and P. M. Pardalos, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001, pp. 472–477.

[34] S. Pickenhain and K. Tammer, *Sufficient conditions for local optimality in multidimensional control problems with state restrictions*, Z. Anal. Anwendungen, 10 (1991), pp. 397–405.

[35] A. V. Sarychev, *First- and second-order sufficient optimlaity conditions for bang-bang controls*, SIAM J. Control Optim., 35 (1997), pp. 315–340.

[36] H. Schättler, *On the local structure of time-optimal bang-bang trajectories in* $\mathbb{R}^3$, SIAM J. Control Optim., 26 (1988), pp. 186–204.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 6, pp. 1868–1885

GLOBAL WEAK SHARP MINIMA ON BANACH SPACES*

KUNG FU NG† AND XI YIN ZHENG†

Abstract. We consider a proper lower semicontinuous function f on a Banach space X with $\lambda = \inf\{f(x) : x \in X\} > -\infty$. Let $\alpha \geq \lambda$ and $S_\alpha = \{x \in X : f(x) \leq \alpha\}$. We define the lower derivative of f at the set S_α by

$$\underline{D}(f, S_\alpha) = \liminf_{x \to S_\alpha} \frac{f(x) - \alpha}{dist(x, S_\alpha)},$$

where $x \to S_\alpha$ can be interpreted in various ways. We show that, when f is convex and $\alpha = \lambda$, it is equal to the largest weak sharp minima constant. In terms of these derivatives and subdifferentials, we present several characterizations for convex f to have global weak sharp minima. Some of these results are also shown to be valid for nonconvex f. As applications, we give error bound results for abstract linear inequality systems.

Key words. weak sharp minima, error bound, Banach spaces, Asplund space

AMS subject classifications. 90C31, 90C25, 49J52

PII. S0363012901389469

1. Introduction.

Throughout this paper, let X be a Banach space and $f : X \to R\bigcup\{+\infty\}$ a proper lower semicontinuous function; we always assume that f is bounded below and denote by λ the infimum of f on X. For each $\alpha \in R$, let

$$S_\alpha := \{x \in X : f(x) \leq \alpha\};$$

in particular,

$$S_\lambda = \{x \in X : f(x) \leq \lambda\} = \{x \in X : f(x) = \lambda\}.$$

We say that f has global weak sharp minima if $S_\lambda \neq \emptyset$ and there exists $\tau > 0$ such that

$$\tau dist(x, S_\lambda) \leq f(x) - \lambda \qquad \forall x \in X. \tag{1.1}$$

In this case, τ is called a weak sharp minima constant; and $\tau_f := \sup\{\tau : \tau \text{ satisfies } (1.1)\}$ is called the largest weak sharp minima constant of f.

Weak sharp minima occur in many optimization problems; in particular, they are related to the convergence analysis of iterative procedures. Several authors [1, 2, 4, 6, 13, 18, 19, 20] studied such minima (but only local ones). Among these, [1, 2, 6, 13, 18] considered the unique minimizer solution set case. Burke and Ferris [4] extended it to the nonunique minimizer solution set case; Studniarski and Ward [19] and Ward [20] studied local weak sharp minima with a nonunique minimizer solution set. Most previous works add assumptions on the minimizer solution set S_λ and

*Received by the editors May 18, 2001; accepted for publication (in revised form) August 6, 2002; published electronically February 27, 2003. This research was supported by a direct grant (CUHK) and an earmarked grant from the Research Grant Council of Hong Kong.
http://www.siam.org/journals/sicon/41-6/38946.html

†Department of Mathematics, The Chinese University of Hong Kong, Shatin, NT, Hong Kong (kfng@math.cuhk.edu.hk, xyzheng@math.cuhk.edu.hk). The research of the second author was supported by the National Natural Science Foundation of People's Republic of China and ABSF of Yunnan Province, People's Republic of China.

require the space X to be finite dimensional. In this paper, we consider global weak sharp minima on a Banach space and add no assumptions on S_λ. In section 2, we first prove a characterization for a proper lower semicontinuous function on a Banach space to have global weak sharp minima. Using this characterization and a recent significant result (local Lipschitz fuzzy sum rule of Fréchet subdifferential) on Asplund spaces, we give a sufficient condition in terms of Fréchet subdifferential for a proper lower semicontinuous function on an Asplund space to have global weak sharp minima. Moreover, we consider weak sharp minima with constraint for a locally Lipschitz function. In section 3, we discuss weak sharp minima of a lower semicontinuous convex function and give several equivalent conditions for such a function to have weak sharp minima. In section 4, as applications of results in sections 2 and 3, we establish some error bound results for abstract linear inequality systems, which are more general than previous linear inequality systems considered by other authors.

We conclude this section with a compilation of some notations which will be used throughout the paper. For $x \in \mathrm{dom}(f) := \{x \in X : \ f(x) < +\infty\}$ and $\varepsilon \geq 0$, let $\tilde{\partial}_\varepsilon f(x)$ denote the set

$$\left\{x^* \in X^* : \ \liminf_{y \to x} \frac{f(y) - f(x) - \langle x^*, y - x\rangle}{\|y - x\|} \geq -\varepsilon\right\}.$$

$\tilde{\partial} f(x)$ stands for $\tilde{\partial}_0 f(x)$ and is called the Fréchet subdifferential of f at x. Let

$$\begin{aligned}\partial f(x) &:= \limsup_{y \overset{f}{\to} x} \tilde{\partial} f(y)\\ &:= \{x^* \in X^* : \ x_n^* \overset{w^*}{\to} x^* \text{ with } x_n^* \in \tilde{\partial} f(x_n),\ x_n \to x \text{ and } f(x_n) \to f(x)\}.\end{aligned}$$

$\partial f(x)$ is called the limiting Fréchet subdifferential of f at x. If X is an Asplund space, $\partial f(x) = \limsup_{y \overset{f}{\to} x, \varepsilon \downarrow 0} \tilde{\partial}_\varepsilon f(y)$ (see [14, Theorem 2.9]). Clearly, $\tilde{\partial} f(x) \subset \partial f(x)$. It is known that if f is convex, then

$$\tilde{\partial} f(x) = \partial f(x) = \{x^* \in X^* : \ \langle x^*, y - x\rangle \leq f(y) - f(x) \qquad \forall y \in X\}.$$

For more discussions on generalized differentials, see [5] and [12]. For a closed subset K of X and $x \in K$, define $N(K, x) := \partial \delta_K(x)$, where δ_K is the indicator function of K. It is known [14] that $N(K, x) = \bigcup_{\lambda > 0} \lambda \partial dist(x, K)$. Recall that X is called an Asplund space if every continuous convex function on an open convex subset D of X is Fréchet differentiable on a dense G_δ subset of D. It is well known [17] that each Banach space with an equivalent Fréchet smooth norm, each Banach space with separable dual, and each reflexive Banach space are examples of Asplund spaces.

2. Weak sharp minima for lower semicontinuous functions. We first present a general characterization for f to have global weak sharp minima, which is also a tool to prove other results in this section.

THEOREM 2.1. *f has global weak sharp minima with a constant $\tau > 0$ if and only if there exists a sequence $\{\lambda_n\}$ in R such that $\lambda_n \to \lambda^+$ and for each $x \in X$ with $f(x) > \lambda$,*

$$\tau \liminf_{n \to \infty} dist(x, S_{\lambda_n}) \leq f(x) - \lambda. \tag{2.1}$$

Proof. It is clear that (2.1) holds if f has global weak sharp minima with a constant $\tau > 0$. Conversely, suppose that (2.1) holds. We need only show that $S_\lambda \neq \emptyset$

and that for each $\gamma \in (0,\ \tau)$ and each $x \in X \setminus S_\lambda$,

$$\gamma dist(x, S_\lambda) \le f(x) - \lambda.$$

Fixing $\gamma \in (0,\ \tau)$, for each $y \in \mathrm{dom}(f)$, let

$$F(y) = \{y' \in X :\ \gamma\|y - y'\| \le f(y) - f(y')\}.$$

Then $F(y)$ is a nonempty closed subset of X for each $y \in \mathrm{dom}(f)$. Since $\lambda = \inf\{f(z) : z \in X\} > -\infty$, one can inductively construct a sequence $\{x_n\}$ such that

(i) $x_0 = x$,
(ii) $x_n \in F(x_{n-1})$,
(iii) $f(x_n) \le \inf\{f(y) :\ y \in F(x_{n-1})\} + \frac{1}{n}$.

By (ii) and the definition of F, it is easy to verify that $F(x_n) \subset F(x_{n-1})$. This and (iii) imply that $F(x_n) \subset B(x_n, \frac{1}{\gamma n})$, where $B(x_n, \frac{1}{\gamma n})$ denotes the ball with center x_n and radius $\frac{1}{\gamma n}$. It follows from the completeness of X that there exists $z \in X$ such that $\{z\} = \bigcap_{n=0}^{\infty} F(x_n) \subset F(x)$. Thus $F(z) = \{z\}$ and $\gamma\|x - z\| \le f(x) - f(z)$. It remains to show that $f(z) = \lambda$. If this is not the case, then $f(z) > \lambda$. By (2.1) and passing to a subsequence if necessary, we can assume that the lower limit on the left-hand side of (2.1) is the full limit. By lower semicontinuity of f and $\lambda_n \to \lambda^+$, one has $\lim_{n\to\infty} dist(z, S_{\lambda_n}) > 0$. Since $\gamma \in (0,\ \tau)$, it follows that

$$\gamma dist(z, S_{\lambda_n}) < f(z) - \lambda \qquad \forall \text{ large enough } n.$$

Therefore, for each large enough n there exists $z_n \in S_{\lambda_n} \setminus \{z\}$ such that

$$\gamma\|z - z_n\| < f(z) - \lambda_n \le f(z) - f(z_n),$$

and so $z_n \in F(z) = \{z\}$; thus $z_n = z$, contradicting our choice of z_n.

THEOREM 2.2. *Let X be an Asplund space and $\tau > 0$ be such that*

$$\inf\{\|x^*\| :\ x^* \in \tilde{\partial} f(x),\ x \in X \text{ and } f(x) > \lambda\} \ge \tau. \tag{2.2}$$

Then f has global weak sharp minima with a constant τ.

Proof. Suppose not; then by Theorem 2.1 there exist $x_0 \in X$ and $\lambda_0 > \lambda$ such that

$$\tau dist(x_0, S_{\lambda_0}) > f(x_0) - \lambda,$$

that is,

$$f(x_0) < \inf\{f(x) :\ x \in X\} + \tau dist(x_0, S_{\lambda_0}).$$

Pick $\alpha > 0$ such that

$$f(x_0) < \inf\{f(x) :\ x \in X\} + (\tau - \alpha)(dist(x_0, S_{\lambda_0}) - \alpha).$$

By the Ekeland variational principle (cf. [5, Theorem 7.5.1]), there exists $v \in X$ such that

(i) $\|v - x_0\| < dist(x_0, S_{\lambda_0}) - \alpha$,
(ii) $f(v) < f(x) + (\tau - \alpha)\|x - v\|$ for each $x \in X \setminus \{v\}$.

Since X is an Asplund space, it follows from (ii) and Theorem 2.12 in [22] that there exist $u_i \in X$ and $x_i^* \in X^*$ $(i = 1, 2)$ such that

$$\|u_i - v\| < \alpha \quad (i = 1, 2), \tag{2.3}$$

$$x_1^* \in \tilde{\partial} f(u_1), \quad x_2^* \in (\tau - \alpha)\partial(\|\cdot - v\|)(u_2), \quad \text{and} \quad \|x_1^* + x_2^*\| < \alpha.$$

It follows from the fact that $\partial(\|\cdot - v\|)(u_2) \subset \{x^* \in X^* : \|x^*\| \le 1\}$ that

$$\|x_1^*\| < \|x_2^*\| + \alpha \le \tau.$$

Therefore,

$$\inf\{\|x^*\| : x^* \in \tilde{\partial} f(u_1)\} < \tau. \tag{2.4}$$

On the other hand, by (2.3) and (i), one has that $u_1 \notin S_{\lambda_0}$, and so $f(u_1) > \lambda_0 > \lambda$. Thus (2.4) contradicts the given assumption (2.2).

The following result concerns the constrained case.

THEOREM 2.3. *Let X be an Asplund space, K a closed nonempty subset of X, and f a locally Lipschitz function on X. Let $\lambda_K := \inf\{f(x) : x \in K\}$ be finite. Assume that there exists $\tau > 0$ such that for each $x \in K$ with $f(x) > \lambda_K$,*

$$\inf\{\|x^* + y^*\| : x^* \in \partial f(x) \text{ and } y^* \in N(K, x)\} \ge \tau. \tag{2.5}$$

Then $S(\lambda_K) := \{x \in K : f(x) = \lambda_K\} \ne \emptyset$ and

$$\tau dist(x, S(\lambda_K)) \le f(x) - \lambda_K \qquad \forall x \in K.$$

Proof. Let $\phi = f + \delta_K$. Then ϕ is a proper lower semicontinuous function on X, $\inf\{\phi(x) : x \in X\} = \lambda_K$, and $S(\lambda_K) = \{x \in X : \phi(x) = \lambda_K\}$. It suffices to show that ϕ has global weak sharp minima with constant τ. By Theorem 4.1 in [14], for each $x \in K$

$$\partial\phi(x) = \partial(f + \delta_K)(x) \subset \partial f(x) + \partial\delta_K(x) = \partial f(x) + N(K, x).$$

Thus (2.5) reads $\inf\{\|z^*\| : z^* \in \partial\phi(x)\} \ge \tau$ for each $x \in K$ with $f(x) > \lambda_K$. Since $\tilde{\partial}\phi(x) = \emptyset$ for each $x \notin K$ and since $\tilde{\partial}\phi(x) \subset \partial\phi(x)$ for each $x \in K$ with $\phi(x) > \lambda_K$, it follows that

$$\inf\{\|z^*\| : z^* \in \tilde{\partial}\phi(x), x \in X \text{ and } \phi(x) > \lambda_K\} \ge \tau.$$

Thus Theorem 2.2 imples that ϕ has global weak sharp minima with the constant τ.

Remark. From the proofs of Theorems 2.2 and 2.3, it is clear that we can generalize these two theorems to any triple $(X, \mathcal{F}, \partial_a)$, where X is a Banach space, $\mathcal{F}$ a function space on X, and ∂_a is an abstract subdifferential operator with appropriate properties. For example, we can take $\mathcal{F}$ to be the set of all proper lower semicontinuous functions from X to $R \cup \{\infty\}$ and $\partial_a : \mathcal{F} \times X \to 2^{X^*}$ with the following properties:

(i) For each equivalent norm $\|\cdot\|_e$ of X and each $x \in X$,

$$\partial_a\|\cdot\|_e(x) = \{x^* : \langle x^*, h\rangle \le \|x + h\|_e - \|x\|_e, \ h \in X\}.$$

(ii) For any $g \in \mathcal{F}$ and any equivalent norm $\|\cdot\|_e$ of X, if $v \in X$ is a minimum point of $g + \|\cdot\|_e$ on X, then for any $\varepsilon > 0$ there exist $v_1, v_2 \in X$ and $x_1^*, x_2^* \in X^*$ such that

$$\|v_i - v\| < \varepsilon \ \ (i = 1, 2), \ \ x_1^* \in \partial_a g(v_1), \ \ x_2^* \in \partial_a\|\cdot\|_e(v_2), \text{ and } \|x_1^* + x_2^*\| < \varepsilon.$$

In this setup, following the proof of Theorem 2.2 one can obtain the following result: $f \in \mathcal{F}$ has weak sharp minimum with a constant $\tau > 0$ if

$$\inf\{\|x^*\| : x^* \in \partial_a f(x),\ x \in X \text{ with } f(x) > \lambda\} \geq \tau.$$

After the completion of our first draft, we learnt that Wu and Ye [21] have obtained an error bound result similar to Theorem 2.2 in terms of the abstract subdifferential. Jourani [11] also proved an error bound result similar to Theorem 2.2, but some stronger assumptions are added on the abstract subdifferential. In the approach of both [21] and [11], the solution set S is required to be nonempty as an assumption (the proof of [11] requires this assumption though it is not explicitly mentioned). Theorem 2.2 allows us to show that this assumption is automatically satisfied.

The following two propositions tell us that the scenario of weak sharp minima only happens for nonsmooth functions when either f is convex or X is finite dimensional.

PROPOSITION 2.4. *Let X be a finite dimensional space. Suppose that f is a differentiable function on X and that $f(\cdot)$ is not constant on X. Then f has no weak sharp minina.*

Proof. Suppose to the contrary that f has global weak sharp minima with constant $\tau > 0$. Then

$$\tau dist(x, S_\lambda) \leq f(x) - \lambda \qquad \forall x \in X. \tag{2.6}$$

Since f is not constant, one can pick $z \notin S_\lambda$. Since S_λ is a nonempty closed subset of the finite dimensional space X, there exists $x_0 \in S_\lambda$ such that $\|z - x_0\| = dist(z, S_\lambda)$. It follows that for each $t \in (0,\ 1]$,

$$\tau t\|z-x_0\| = \tau dist(x_0+t(z-x_0), S_\lambda) \leq f(x_0+t(z-x_0))-\lambda = f(x_0+t(z-x_0))-f(x_0).$$

Thus $0 < \tau\|z - x_0\| \leq df(x_0)(z - x_0) = \bigtriangledown f(x_0)(z - x_0)$, and so $\bigtriangledown f(x_0) \neq 0$, contradicting the fact that x_0 is a minimizer of f.

PROPOSITION 2.5. *Let f be a differentiable convex function on a Banach space X. Suppose that $f(\cdot)$ is not constant on X. Then f has no weak sharp minima.*

Proof. Suppose to the contrary that (2.6) holds. By the convexity and continuity of f, S_λ is a closed convex subset of X. Noting that ∂S_λ is nonempty, it follows from the Bishop–Phelps theorem that S_λ has support points, that is, there exist $x_0 \in S_\lambda$ and $x^* \in X^*$ with $\|x^*\| = 1$ such that

$$\langle x^*, x_0\rangle = \sup\{\langle x^*, x\rangle : x \in S_\lambda\}. \tag{2.7}$$

Pick $h \in X$ such that $\|h\| = 1$ and $\langle x^*, h\rangle > \frac{1}{2}$. This implies that $dist(th, \ker(x^*)) > \frac{1}{2}t$ for each $t \in (0,\ 1]$, where $\ker(x^*) = \{x \in X : \langle x^*, x\rangle = 0\}$. It is easy to verify from (2.7) that for each $t \in (0,\ 1]$,

$$\frac{1}{2}t \leq dist(th, \ker(x^*)) = dist(x_0 + th, x_0 + \ker(x^*)) \leq dist(x_0 + th, S_\lambda).$$

This and (2.6) imply that for each $t \in (0,\ 1)$,

$$\frac{1}{2}t\tau \leq f(x_0 + th) - \lambda = f(x_0 + th) - f(x_0).$$

It follows that $\frac{1}{2}\tau \leq df(x_0)(h) = \bigtriangledown f(x_0)(h)$, and so $\bigtriangledown f(x_0) \neq 0$. This contradicts the fact that x_0 is a minimizer of f.

3. Characterizations of weak sharp minima for convex functions. Throughout this section, we assume that X is a Banach space, $f : X \to R \cup \{\infty\}$ is a proper lower semicontinuous convex function and that $\lambda = \inf\{f(x) : \ x \in X\} > -\infty$. Let $\mathrm{dom}(f) := \{x \in X : \ f(x) < \infty\}$. Thus f is continuous on each line segment contained in $\mathrm{dom}(f)$. For $x_1, x_2 \in X$, $(x_1, \ x_2]$ denotes the line segment $\{tx_1 + (1-t)x_2 : \ 0 < t \leq 1\}$. Similar notations $(x_1, \ x_2)$ and $[x_1, \ x_2]$ are self-explanatory. For any $\alpha \geq \lambda$, let $S_\alpha := \{x \in X : \ f(x) \leq \alpha\}$. We write

$$\begin{aligned} x \to S_\alpha \ \text{ if } x \notin S_\alpha \text{ and } dist(x, S_\alpha) \to 0, \\ x_f \to S_\alpha \ \text{ if } x \notin S_\alpha \text{ and } f(x) \to \alpha, \\ \text{and } \ x \xrightarrow{f} S_\alpha \ \text{ if } x \to S_\alpha \text{ and } x_f \to S_\alpha. \end{aligned}$$

For an extended real-valued function ϕ on X, $\delta > 0$ and $\alpha \geq \lambda$, let

$$\begin{aligned} &\phi_\alpha^{(1)}(\delta) := \inf\{\phi(x) : \ x \in X \text{ with } 0 < dist(x, S_\alpha) < \delta\}, \\ &\phi_\alpha^{(2)}(\delta) := \inf\{\phi(x) : \ x \in X \text{ with } 0 < f(x) - \alpha < \delta\}, \\ \text{and} \quad &\phi_\alpha^{(1,2)}(\delta) := \inf\{\phi(x) : \ x \in X \text{ with } 0 < \max\{dist(x, S_\alpha), f(x) - \alpha\} < \delta\}. \end{aligned}$$

Thus, one has that

$$\begin{aligned} \liminf_{x \to S_\alpha} \phi(x) &= \lim_{\delta \to 0^+} \phi_\alpha^{(1)}(\delta), \\ \liminf_{x_f \to S_\alpha} \phi(x) &= \lim_{\delta \to 0^+} \phi_\alpha^{(2)}(\delta), \\ \text{and} \quad \liminf_{x \xrightarrow{f} S_\alpha} \phi(x) &= \lim_{\delta \to 0^+} \phi_\alpha^{(1,2)}(\delta). \end{aligned}$$

Below we introduce various kinds of "lower derivatives" of f at the set S_α (rather than the usual ones at a point) with $\alpha \geq \lambda$. These derivatives are defined by

$$\begin{aligned} \underline{D}_1(f, S_\alpha) &:= \liminf_{x \to S_\alpha} \frac{f(x) - \alpha}{dist(x, S_\alpha)}, \\ \underline{D}_2(f, S_\alpha) &:= \liminf_{x_f \to S_\alpha} \frac{f(x) - \alpha}{dist(x, S_\alpha)}, \\ \text{and} \ \ \underline{D}_{1,2}(f, S_\alpha) &:= \liminf_{x \xrightarrow{f} S_\alpha} \frac{f(x) - \alpha}{dist(x, S_\alpha)}. \end{aligned}$$

If $\alpha \in (\lambda, \ \infty)$, then $S_\alpha \neq \emptyset$ and so $\underline{D}_1(f, S_\alpha)$, $\underline{D}_2(f, S_\alpha)$, and $\underline{D}_{1,2}(f, S_\alpha)$ are well defined. If $\alpha = \lambda$, it is possible that $S_\alpha = \emptyset$; in this case, we define $dist(x, S_\alpha) = \infty$ and these lower derivatives are to be understood as 0. Note also that

$$\underline{D}_{1,2}(f, S_\alpha) \geq \max\{\underline{D}_1(f, S_\alpha), \ \underline{D}_2(f, S_\alpha)\}.$$

Remark. By definition, $\underline{D}_{1,2}(f, S_\alpha) > 0$ means $S_\alpha \neq \emptyset$ and that there exist $\delta > 0$ and $\tau > 0$ such that

$$\tau dist(x, S_\alpha) \leq f(x) - \alpha \ \text{ for any } x \in X \text{ with } \max\{dist(x, S_\alpha), f(x) - \alpha\} < \delta.$$

Similar observations can be made for the case when $\underline{D}_1(f, S_\alpha) > 0$ or $\underline{D}_2(f, S_\alpha) > 0$. The following lemma will be a useful tool for us.

LEMMA 3.1. *Let $\alpha \geq \lambda$, $S_\alpha \neq \emptyset$, and $x \in \mathrm{dom}(f) \setminus S_\alpha$. Let $\varepsilon > 0$ and $\delta > 0$ be such that*

$$\delta < \max\{dist(x, S_\alpha),\ f(x) - \alpha\}. \tag{3.1}$$

Then there exist $s \in S_\alpha$ with $f(s) = \alpha$, and $y \in (s,\ x)$ such that

(a) $\delta = \max\{dist(y, S_\alpha), f(y) - \alpha\}$,
(b) $(s,\ x] \cap S_\alpha = \emptyset$,
(c) $\|y - s\| < (1 + \varepsilon) dist(y, S_\alpha)$,
(d) $\|x - s\| < (1 + \varepsilon) dist(x, S_\alpha)$,
(e)

$$\frac{f(y) - \alpha}{dist(y, S_\alpha)} \leq (1 + \varepsilon) \frac{f(x) - \alpha}{dist(x, S_\alpha)}.$$

Proof. Take a sequence $\{s_n\}$ in S_α such that

$$\|x - s_n\| \to dist(x, S_\alpha). \tag{3.2}$$

For each n, we may assume without loss of generality that $(s_n,\ x] \cap S_\alpha = \emptyset$ and $f(s_n) = \alpha$. It follows from (3.1) and the intermediate value theorem that there exists $y_n \in (s_n,\ x)$ such that

$$\delta = \max\{dist(y_n, S_\alpha),\ f(y_n) - \alpha\}. \tag{3.3}$$

Write $y_n = s_n + t_n(x - s_n)$ for some $t_n \in (0,\ 1)$; it follows from the convexity of f that

$$f(y_n) \leq (1 - t_n) f(s_n) + t_n f(x) = \alpha + t_n(f(x) - \alpha). \tag{3.4}$$

We claim that

$$\frac{t_n \|x - s_n\|}{dist(y_n, S_\alpha)} \to 1. \tag{3.5}$$

Indeed, if this is not the case, then, by $dist(y_n, S_\alpha) \leq \|y_n - s_n\| = t_n \|x - s_n\|$, one can assume that

$$\frac{t_n \|x - s_n\|}{dist(y_n, S_\alpha)} \to \beta > 1 \tag{3.6}$$

(passing to a subsequence if necessary). Since $\|s_n - x\| \to dist(x, S_\alpha)$, one can assume without loss of generality that for each n

$$t_n > \frac{dist(y_n, S_\alpha)}{dist(x, S_\alpha)}. \tag{3.7}$$

By (3.4), one also has that $t_n \geq \frac{f(y_n) - \alpha}{f(x) - \alpha}$ for each n. Setting $r = \min\{\frac{1}{dist(x, S_\alpha)}\ ,\ \frac{1}{f(x) - \alpha}\}$, it follows from (3.7) and (3.3) that for each n,

$$t_n \geq r \max\{dist(y_n, S_\alpha),\ f(y_n) - \alpha\} = r\delta.$$

Since $t_n \in (0,\ 1)$, one can assume without loss of generality that $t_n \to t$ for some t. Then $t \geq r\delta > 0$. Since

$$dist(x, S_\alpha) \leq \|x - y_n\| + dist(y_n, S_\alpha) = (1 - t_n)\|x - s_n\| + dist(y_n, S_\alpha),$$

$$\frac{dist(x, S_\alpha)}{\|x - s_n\|} \leq 1 - t_n + \frac{dist(y_n, S_\alpha)}{\|x - s_n\|}.$$

Letting $n \to \infty$, it follows from (3.2) and (3.6) that $1 \leq 1 - t + \frac{t}{\beta}$. This is not possible as $\beta > 1$ and $t > 0$. Therefore (3.5) is true and so

$$\frac{\|y_n - s_n\|}{dist(y_n, S_\alpha)} \to 1. \tag{3.8}$$

Moreover, rewrite (3.4) in the form

$$\frac{f(y_n) - \alpha}{dist(y_n, S_\alpha)} \leq \frac{t_n(f(x) - \alpha)}{dist(y_n, S_\alpha)} \frac{\|x - s_n\|}{\|x - s_n\|}$$

and note that the right-hand side converges to $\frac{f(x)-\alpha}{dist(x,S_\alpha)}$ by (3.5) and (3.2). It follows from (3.8) and (3.2) that if we take $y = y_n$ for large enough n, then (e), (d), and (c) are satisfied.

PROPOSITION 3.2. *$\underline{D}_1(f, S_\alpha)$, $\underline{D}_2(f, S_\alpha)$, and $\underline{D}_{1,2}(f, S_\alpha)$ are increasing on the interval $[\lambda, \infty)$ with respect to α.*

Proof. We need only show the conclusion to hold for $\underline{D}_1$ as proofs are similar for $\underline{D}_2$ and $\underline{D}_{1,2}$. Let $\alpha_1 > \alpha_2 \geq \lambda$. By way of contradiction we suppose that

$$\underline{D}_1(f, S_{\alpha_1}) < r < \underline{D}_1(f, S_{\alpha_2}) \tag{3.9}$$

for some $r \in R$. Then there exists $z \in \mathrm{dom}(f) \setminus S_{\alpha_1}$ such that

$$\frac{f(z) - \alpha_1}{dist(z, S_{\alpha_1})} < r. \tag{3.10}$$

By Lemma 3.1, for every natural number n there exist $x_n \in S_{\alpha_2}$ with

$$f(x_n) = \alpha_2 \quad \text{and} \quad (x_n,\ z] \cap S_{\alpha_2} = \emptyset$$

and $y_n \in (x_n,\ z]$ such that

$$\|y_n - x_n\| < \left(1 + \frac{1}{n}\right) dist(y_n, S_{\alpha_2}) \tag{3.11}$$

and

$$\max\{dist(y_n, S_{\alpha_2}), f(y_n) - \alpha_2\} = \frac{\alpha_1 - \alpha_2}{2n}. \tag{3.12}$$

In particular, $dist(y_n, S_{\alpha_2}) \to 0$ and by definition,

$$\underline{D}_1(f, S_{\alpha_2}) \leq \liminf_{n\to\infty} \frac{f(y_n) - \alpha_2}{dist(y_n, S_{\alpha_2})}.$$

By (3.12), one has that $f(y_n) < \alpha_1 < f(z)$. Take w_n in the open segment $(y_n,\ z)$ such that $f(w_n) = \alpha_1$. Since f is convex and since $w_n \in (y_n,\ z] \subset (x_n,\ z]$, one has

$$\frac{f(z) - f(w_n)}{\|z - w_n\|} \geq \frac{f(y_n) - f(x_n)}{\|y_n - x_n\|},$$

that is,

$$\frac{f(z)-\alpha_1}{\|z-w_n\|} \geq \frac{f(y_n)-\alpha_2}{\|y_n-x_n\|}.$$

Thus,

$$\frac{f(z)-\alpha_1}{dist(z,S_{\alpha_1})} \geq \frac{f(y_n)-\alpha_2}{\|y_n-x_n\|}.$$

It follows from (3.10) and (3.11) that

$$\left(1+\frac{1}{n}\right) r \geq \frac{f(y_n)-\alpha_2}{dist(y_n,S_{\alpha_2})} \qquad \text{for each } n$$

and so

$$r \geq \liminf_{n\to\infty} \frac{f(y_n)-\alpha_2}{dist(y_n,S_{\alpha_2})} \geq \underline{D}_1(f,S_{\alpha_2}),$$

contradicting (3.9).

THEOREM 3.3. *Let $\tau > 0$ be a constant. The following statements are equivalent.*

(i) $\inf\{\|x^*\| :\ x^* \in \partial f(X \setminus S_\lambda)\} \geq \tau$.
(ii) $\liminf_{x\ f\to S_\lambda} \inf\{\|x^*\| :\ x^* \in \partial f(x)\} \geq \tau$.
(iii) $S_\lambda \neq \emptyset$ *and* $\liminf_{x\to S_\lambda} \inf\{\|x^*\| :\ x^* \in \partial f(x)\} \geq \tau$.
(iv) $S_\lambda \neq \emptyset$ *and* $\liminf_{x\xrightarrow{f} S_\lambda} \inf\{\|x^*\| :\ x^* \in \partial f(x)\} \geq \tau$.
(v) $S_\lambda \neq \emptyset$ *and* $\underline{D}_1(f,S_\lambda) \geq \tau$.
(vi) $S_\lambda \neq \emptyset$ *and* $\underline{D}_2(f,S_\lambda) \geq \tau$.
(vii) $S_\lambda \neq \emptyset$ *and* $\underline{D}_{1,2}(f,S_\lambda) \geq \tau$.
(viii) *f has global weak sharp minima with constant τ.*

Consequently $\underline{D}_1(f,S_\lambda)$, $\underline{D}_2(f,S_\lambda)$, and $\underline{D}_{1,2}(f,S_\lambda)$ coincide: all equal zero if f does not have global weak sharp minima, and otherwise all equal the largest weak sharp minimum constant of f.

Proof. We need only prove the equivalence of (i)–(viii). It is clear that (i)⇒(ii) (and hence (i)⇒(iii) by virtue of the implication (ii)⇒(iv) to be proved below), (iii)⇒(iv), (v)⇒(vii), (vi)⇒(vii), (viii)⇒(v), and (viii)⇒(vi).

(ii)⇒(iv). Suppose that (ii) holds. We need only show that $S_\lambda \neq \emptyset$. Let $r \in (0,\ \tau)$; then by (ii) there exists $\delta > 0$ such that for each $x \in X \setminus S_\lambda$ with $f(x) < \lambda + \delta$

$$\inf\{\|x^*\| :\ x^* \in \partial f(x)\} > r. \tag{3.13}$$

Pick $x_1 \in X$ such that $f(x_1) < \lambda + \min\{\delta,\ r\}$. By the Ekeland variational principle there exists $y_1 \in X$ such that

(a) $f(y_1) \leq f(x_1)$,
(b) $f(y_1) < f(x) + \min\{\delta,\ r\}\|x - y_1\|$ for any $x \neq y_1$.

By (a), $f(y_1) < \lambda + \delta$; by (b), one has that

$$\inf\{\|x^*\| :\ x^* \in \partial f(y_1)\} \leq r.$$

It follows from (3.13) that $y_1 \in S_\lambda$, and so $S_\lambda \neq \emptyset$.

(iv)⇒(vii). Suppose that (iv) holds. Let $r \in (0,\ \tau)$. Then there exists $\delta > 0$ such that for any $x \in X \setminus S_\lambda$ with $\max\{dist(x,S_\lambda), f(x) - \lambda\} < \delta$,

$$\inf\{\|x^*\| :\ x^* \in \partial f(x)\} > r. \tag{3.14}$$

We claim that $\underline{D}_{1,2}(f, S_\lambda) \geq r$. To see this, suppose to the contrary that there exists $x_1 \in X \setminus S_\lambda$ with

$$\max\{dist(x_1, S_\lambda),\ f(x_1) - \lambda\} < \frac{\delta}{2}$$

such that $\frac{f(x_1)-\lambda}{dist(x_1,S_\lambda)} < r$, that is,

$$f(x_1) < \lambda + r dist(x_1, S_\lambda).$$

Using the Ekeland variational principle, there exists $y_1 \in X$ such that

(c) $f(y_1) \leq f(x_1)$,
(d) $\|y_1 - x_1\| < dist(x_1, S_\lambda) < \frac{\delta}{2}$,
(e) $f(y_1) < f(x) + r\|x - y_1\|$ for all $x \neq y_1$.

Note that (d) and (c) imply that $y_1 \in X \setminus S_\lambda$ and that $\max\{dist(y_1, S_\lambda), f(y_1) - \lambda\} < \delta$; (e) entails that

$$\inf\{\|y^*\| :\ y^* \in \partial f(y_1)\} \leq r.$$

This contradicts (3.14). Therefore $\underline{D}_{1,2}(f, S_\lambda) \geq r$ whenever $r \in (0, \tau)$, and (vii) is seen to hold.

(vii)⇒(viii). Suppose that (vii) holds. Let $r \in (0,\ \tau)$. Then there exists $\delta > 0$ such that for each $x \in X \setminus S_\lambda$ with $\max\{dist(x, S_\lambda), f(x) - \lambda\} \leq \delta$,

$$\frac{f(x) - \lambda}{dist(x, S_\lambda)} > r. \tag{3.15}$$

On the other hand, let $z \in X$ with $\max\{dist(z, S_\lambda), f(z) - \lambda\} > \delta$. Let $\varepsilon > 0$. By Lemma 3.1 there exists $y \in X$ such that

$$\max\{dist(y, S_\lambda), f(y) - \lambda\} = \delta \qquad \text{and} \qquad \frac{f(y) - \lambda}{dist(y, S_\lambda)} \leq (1 + \varepsilon)\frac{f(z) - \lambda}{dist(z, S_\lambda)}.$$

It follows from (3.15) that $r dist(z, S_\lambda) \leq (1+\varepsilon)(f(z)-\lambda)$. Letting $\varepsilon \to 0$, $r dist(z, S_\lambda) \leq f(z) - \lambda$. This and (3.15) imply that $r dist(\cdot, S_\lambda) \leq f(\cdot) - \lambda$ on X. Letting $r \to \tau$, (viii) is seen to hold.

(viii)⇒(i). Suppose that (viii) holds: for any $x \in X \setminus S_\lambda$

$$\tau dist(x, S_\lambda) \leq f(x) - \lambda.$$

Let $r \in (0,\ \tau)$. Then there exists $x_\lambda \in S_\lambda$ such that

$$r\|x - x_\lambda\| < \tau dist(x, S_\lambda) \leq f(x) - \lambda = f(x) - f(x_\lambda).$$

It follows that for any $x^* \in \partial f(x)$,

$$\langle x^*, x_\lambda - x\rangle \leq f(x_\lambda) - f(x) \leq -r\|x - x_\lambda\|,$$

and so $\|x^*\| \geq r$. This shows that

$$\inf\{\|x^*\| :\ x^* \in \partial f(x)\} \geq r.$$

Letting $r \to \tau$, (i) is seen to hold. This completes the proof.

Remarks. (i) Let $\mu := \sup\{f(x) : x \in \mathrm{dom}(f)\}$, and for each $\alpha \in [\lambda, \mu)$, let $f_\alpha(x) = \max\{f(x), \alpha\}$ $(x \in X)$. Then f_α is a proper lower semicontinuous convex function, $\inf\{f_\alpha(x) : x \in X\} = \alpha$. By definition,

$$\underline{D}_1(f, S_\alpha) = \underline{D}_1(f_\alpha, S_\alpha), \ \underline{D}_2(f, S_\alpha) = \underline{D}_2(f_\alpha, S_\alpha), \text{ and } \underline{D}_{1,2}(f, S_\alpha) = \underline{D}_{1,2}(f_\alpha, S_\alpha)$$

for any $\alpha \in [\lambda, \mu)$. By Theorem 3.3, these lower derivatives coincide: all equal zero if f_α does not have weak sharp minima, and otherwise all equal the largest weak sharp minima constant of f_α.

(ii) If $\mu \neq \infty$, then, for any $\alpha \in [\mu, \infty)$, $S_\alpha = S_\mu = \mathrm{dom}(f) \neq \emptyset$ and $f(x) - \alpha = \infty$ for any $x \notin S_\alpha$. Thus, by definition, for any $\alpha \in [\mu, \infty)$,

$$\underline{D}_1(f, S_\alpha) = \underline{D}_2(f, S_\alpha) = \underline{D}_{1,2}(f, S_\alpha) = \infty. \tag{3.16}$$

(iii) By virtue of the preceding remarks (i) and (ii), we shall write $\underline{D}$ in place of $\underline{D}_1$, $\underline{D}_2$, and $\underline{D}_{1,2}$.

The first assertion of following corollary is a consequence of remarks (i) and (ii), while the second assertion is that of Proposition 3.2.

COROLLARY 3.4. *For any $\alpha \in [\lambda, \infty)$, let*

$$\Gamma_\alpha := \{\gamma \geq 0 : \gamma dist(x, S_\alpha) \leq [f(x) - \alpha]_+ \text{ for any } x \in X\}$$

and $\tau_\alpha = \sup\{\gamma : \gamma \in \Gamma_\alpha\}$. Then

$$\underline{D}(f, S_\alpha) = \tau_\alpha.$$

Consequently, $\tau_\alpha \leq \tau_\beta$ whenever $\beta \geq \alpha \geq \lambda$.

COROLLARY 3.5. *Let $\lambda \leq \beta < \infty$. Then $\lim_{\alpha \to \beta^+} \underline{D}(f, S_\alpha) = \underline{D}(f, S_\beta)$.*

Proof. In view of (3.16), we need only consider the case when $\beta \in [\lambda, \mu)$. Then by Proposition 3.2,

$$\lim_{\alpha \to \beta^+} \underline{D}(f, S_\alpha) \geq \underline{D}(f, S_\beta). \tag{3.17}$$

If inequality in (3.17) is strict, then there exists $\tau \in R$ such that

$$\lim_{\alpha \to \beta^+} \underline{D}(f, , S_\alpha) > \tau > \underline{D}(f, S_\beta) \ (\geq 0). \tag{3.18}$$

Take a sequence $\{\alpha_n\}$ in (β, μ) with $\alpha_n \to \beta^+$. Then $\underline{D}(f, S_{\alpha_n}) > \tau$. By Corollary 3.4, this implies that $\tau \in \Gamma_{\alpha_n}$ and hence

$$\tau dist(x, S_{\alpha_n}) \leq [f(x) - \alpha_n]_+ \leq [f(x) - \beta]_+ \qquad \forall x \in X.$$

It follows that

$$\tau \liminf_{n \to \infty} dist(x, S_{\alpha_n}) \leq [f(x) - \beta]_+ \qquad \forall x \in X.$$

By Theorem 2.1 (applying to $[f(x) - \beta]_+$), one can show easily that

$$\tau dist(x, S_\beta) \leq [f(x) - \beta]_+ \qquad \forall x \in X.$$

Therefore, $\tau \in \Gamma_\beta$ and hence $\underline{D}(f, S_\beta) \geq \tau$, contradicting (3.18).

Corollary 3.5 and Theorem 3.3 imply that f has weak sharp minimum if and only if $\lim_{\alpha\to\lambda^+} \underline{D}(f, S_\alpha) > 0$.

If X is assumed to be reflexive, one has the following characterization.

THEOREM 3.6. *Let X be a reflexive Banach space and $S_\lambda \neq \emptyset$. Then f has global weak sharp minima with constant $\tau > 0$ if and only if for each $x \in S_\lambda$ there exists $\delta_x > 0$ such that*

$$\inf\{\|x^*\| : x^* \in \partial f(B(x, \delta_x) \setminus S_\lambda)\} \geq \tau. \tag{3.19}$$

Proof. The necessity is a consequence of Theorem 3.3.

For any $x \in \mathrm{dom}(f) \setminus S_\lambda$, there exists $x_0 \in S_\lambda$ such that $f(x_0) = \lambda$, $(x_0,\ x] \cap S_\lambda = \emptyset$, and $\|x - x_0\| = dist(x, S_\lambda)$ (because of the lower semicontinuity and convexity of f and the reflexivity of X). It is clear that $\|y - x_0\| = dist(y, S_\lambda)$ for each $y \in (x_0,\ x]$. We assert that

$$\tau dist(y, S_\lambda) \leq f(y) - \lambda \text{ for each } y \in (x_0,\ x] \cap B\left(x_0, \frac{\delta_{x_0}}{2}\right). \tag{3.20}$$

Granting this, one sees that the convexity of f implies that for each $y \in (x_0,\ x] \cap B(x_0, \frac{\delta_{x_0}}{2})$

$$\tau \leq \frac{f(y) - \lambda}{dist(y, S_\lambda)} = \frac{f(y) - f(x_0)}{\|y - x_0\|} \leq \frac{f(x) - f(x_0)}{\|x - x_0\|} = \frac{f(x) - \lambda}{dist(x, S_\lambda)}$$

and so $\tau dist(x, S_\lambda) \leq f(x) - \lambda$. Thus it remains to show that (3.20) holds. Suppose to the contrary that there exists $y_0 \in (x_0,\ x] \cap B(x_0, \frac{\delta_{x_0}}{2})$ and $0 < \gamma < \tau$ such that $f(y_0) < \lambda + \gamma\|y_0 - x_0\|$. By the Ekeland variational principle, there exists $v \in X$ such that

(a) $\|v - y_0\| < \|y_0 - x_0\|$,

(b) $f(v) < f(z) + \gamma\|z - v\|$ for all $z \neq v$.

Since $dist(y_0, S_\lambda) = \|y_0 - x_0\| < \frac{\delta_{x_0}}{2}$,

(a) implies that $v \in B(x_0, \delta_{x_0}) \setminus S_\lambda$. On the other hand,

(b) implies that $\inf\{\|v^*\| : v^* \in \partial f(v)\} \leq \gamma < \tau$.

This contradicts (3.19).

If f is further assumed to be continuous, one can obtain another characterization. For any $\alpha > \lambda$, let $L_\alpha := \{x \in X : f(x) = \alpha\}$.

THEOREM 3.7. *Let X be a reflexive Banach space and f a continuous convex function on X, and let $\tau > 0$ be a constant. Then f has weak sharp minimum with the constant τ if and only if there exists a sequence $\{\lambda_n\}$ with $\lambda_n \to \lambda^+$ such that*

$$\limsup_{n\to\infty} \inf\{\|x^*\| : x^* \in \partial f(L_{\lambda_n})\} \geq \tau. \tag{3.21}$$

Proof. Suppose that (3.21) holds for some sequence $\{\lambda_n\}$ with $\lambda_n \to \lambda^+$. Then

$$\tau \liminf_{n\to\infty} dist(x, S_{\lambda_n}) \leq f(x) - \lambda. \tag{3.22}$$

Indeed, if this is not the case, then there exist $\tau_0 \in (0,\ \tau)$ and $x_0 \in X$ with $f(x_0) > \lambda$ such that

$$\tau_0 \liminf_{n\to\infty} dist(x_0, S_{\lambda_n}) > f(x_0) - \lambda.$$

Since $\lambda_n \to \lambda_+$, we can assume without loss of generality that for all n, $f(x_0) > \lambda_n$ and

$$\tau_0 dist(x_0, S_{\lambda_n}) > f(x_0) - \lambda_n. \tag{3.23}$$

By the convexity of f and the reflexivity of X, there exists $x_n \in X$ with $f(x_n) = \lambda_n$ such that $\|x_0 - x_n\| = dist(x_0, S_{\lambda_n})$; thus

$$\mathrm{int}(B(x_0, \|x_0 - x_n\|)) \bigcap S_{\lambda_n} = \emptyset.$$

By the separation theorem there exists $z^* \in X^*$ with $\|z^*\| = 1$ such that

$$\langle z^*, x_0\rangle - \|x_0 - x_n\| = \langle z^*, x_n\rangle = \sup\{\langle z^*, x\rangle : \ x \in S_{\lambda_n}\}.$$

Thus $z^* \in N(S_{\lambda_n}, x_n)$. Since f is continuous and $\lambda_n > \lambda$, $N(S_{\lambda_n}, x_n) = \mathrm{cone}(\partial f(x_n))$. Hence there exist $r > 0$ and $y^* \in \partial f(x_n)$ such that $y^* = rz^*$. Therefore,

$$\|y^*\| dist(x_0, S_{\lambda_n}) = \|y^*\|\|x_0 - x_n\| = \langle y^*, x_0 - x_n\rangle \le f(x_0) - f(x_n) = f(x_0) - \lambda_n.$$

This and (3.23) imply that $\|y^*\| \le \tau_0$, and so

$$\inf\{\|x^*\| : \ x^* \in \partial f(x) \text{ and } x \in X \text{ with } f(x) = \lambda_n\} \le \tau_0,$$

contradicting (3.21). Therefore (3.22) holds, and it follows from Theorem 2.1 that f has global weak sharp minima with the constant τ. The sufficiency part is proved. The necessity part follows easily from Theorem 3.3.

From Theorem 3.3 and Theorem 3.7, one has the following result.

COROLLARY 3.8. *Let X be a reflexive Banach space and f a continuous convex function. Then $\lim_{\beta\to\alpha^+} \inf\{\|x^*\| : \ x^* \in \partial f(L_\beta)\}$ exists for each $\alpha \in [\lambda, \infty)$.*

Proof. Let τ_α be as in Corollary 3.4. Applying Theorem 3.7 to f_α, one can easily check that

$$\limsup_{\beta\to\alpha^+} \inf\{\|x^*\| : \ x^* \in \partial f(L_\beta)\} = \tau_\alpha. \tag{3.24}$$

On the other hand, by Theorem 3.3 (also applied to f_α), one has that $\tau_\alpha \le \inf\{\|x^*\| : x^* \in \partial f(X \setminus S_\alpha)\}$. Hence

$$\liminf_{\beta\to\alpha^+} \inf\{\|x^*\| : \ x^* \in \partial f(L_\beta)\} \ge \tau_\alpha.$$

It follows from (3.24) that

$$\lim_{\beta\to\alpha^+} \inf\{\|x^*\| : \ x^* \in \partial f(L_\beta)\} = \tau_\alpha.$$

This completes the proof.

Remark. Adopting the approach of Theorem 3.1 in [15], one sees that some other characterizations for f to have weak sharp minima can be given in terms of either local versions or the directional derivatives.

Let $x_i^* \in X^*$ and $c_i \in R$ for $i = 1, \ldots, n$. Define $\phi(x) = \max\{\langle x_i^*, x\rangle + c_i : \ i = 1, \ldots, n\}$ for each $x \in X$. Then, ϕ is convex but is in general not Fréchet differentiable on X. Moreover, ϕ has global weak sharp minima if (and only if) ϕ is bounded below. Indeed, let $I(x) = \{i \in I : \ \langle x_i^*, x\rangle + c_i = \phi(x)\}$ for each $x \in X$, where $I = \{1, \ldots, n\}$. Note that $0 \notin \partial\phi(x) = \mathrm{co}\{x_i^* : \ i \in I(x)\}$ for each $x \in X$ with $\phi(x) > \lambda$. It follows from the fact that $\{I(x) : \ x \in X \text{ with } \phi(x) > \lambda\}$ is a finite set that

$$\inf\{\|x^*\| : \ x^* \in \partial\phi(x) \text{ and } x \in X \text{ with } \phi(x) > \lambda\} > 0.$$

This and Theorem 3.6 imply that ϕ has global weak sharp minima.

4. Application to an abstract Hoffman error bound. In 1952, Hoffman [8] established an error bound for a finite system of linear inequalities. Since then, his pioneering work has been generalized in numerous ways. For details, readers may see Pang's survey paper [16]. Recently, Hu and Wang [10], Goberna, Lopez, and Todorov [7], and Hu [9] considered an error bound for the following infinite system of linear inequalities. Let U be an index set; $\mathcal{B}(U, R^n)$ denotes the set of bounded functions $a : U \to R^n$, that is, $a(U) := \{a(u) : u \in U\}$ is bounded. For $a \in \mathcal{B}(U, R^n)$ and $b \in \mathcal{B}(U, R^1)$, define a system of linear inequalities by

$$a(u)^T x \le b(u) \qquad \forall u \in U, \tag{4.1}$$

where $x \in R^n$ and $a(u)^T$ denotes the transpose of $a(u)$.

In this section, applying results in section 3, we will establish error bounds for a more general class of linear inequality systems. Let X and Y be Banach spaces. Let $C \subset Y$ be a closed convex cone, which specifies a preorder "$\le_C$" as follows: for $y_1, y_2 \in Y$, $y_1 \le_C y_2$ if and only if $y_2 - y_1 \in C$. Let $A : X \to Y$ be a bounded linear operator and $b \in Y$. Define an abstract linear inequality system (A, C, b) by

$$A(x) \le_C b, \tag{4.2}$$

where $x \in X$. Let $S_C := \{x \in X : A(x) \le_C b\}$ be the solution set of the system. The system (A, C, b) is said to have an error bound if there exists $\tau > 0$ such that

$$dist(x, S_C) \le \tau dist(Ax - b, -C) \qquad \forall x \in X.$$

In general such a constant τ does not necessarily exist even if X, Y are finite dimensional. In the following we present two sufficient conditions to ensure that (A, C, b) has an error bound.

If $X = R^n$, Y is the space $\mathcal{B}(U, R)$ of all bounded functions on the index set U equipped with the supremum norm, and the cone $C_U := \{y \in \mathcal{B}(U, R) : y(u) \ge 0$ for all $u \in U\}$, and if $A_U : R^n \to \mathcal{B}(U, R)$ is such that $A_U(x)(u) = a(u)^T x$ for each $x \in X$, then it is easily seen that the system (A_U, C_U, b) defined by (4.2) is exactly the system (4.1). Thus (4.2) may be viewed as a generalization of (4.1).

Let C^+ denote the dual cone of C, that is,

$$C^+ := \{y^* \in Y^* : \langle y^*, y\rangle \ge 0 \text{ all } x \in C\}.$$

DEFINITION 4.1. *The system (A, C, b) is said to have property (P) if there exist finitely many closed convex cones $C_1, \ldots, C_m$ in Y satisfying the following conditions:*

(i) *$C \subset C_i$ and $\text{int}C_i \ne \emptyset$ for each $i \in I := \{1, \ldots, m\}$.*

(ii) *$0 \notin A^*(C_i^+ \setminus \{0\})$ for each $i \in I$.*

(iii) *There exists $\alpha > 0$ such that $dist(x, S_C) \le \alpha \max\{dist(x, S_{C_i}) : i \in I\}$ for all $x \in X$, where $S_{C_i} = \{x \in X : A(x) \le_{C_i} b\}$.*

PROPOSITION 4.2. *Let $A : X \to R^m$ be a bounded linear operator and $b = (r_1, \ldots, r_m) \in R^m$. Then (A, R^m_+, b) has property (P).*

Proof. Pick $a_1^*, \ldots, a_m^* \in X^*$ such that $A(x) = (\langle a_1^*, x\rangle, \ldots, \langle a_m^*, x\rangle)$ for each $x \in X$. Write $S_{R^m_+}$ for the set $\{x \in X : Ax \le_C b\}$ with $C = R^m_+$. Let $\mathcal{J} = \{D \subset I : \{a_i^* : i \in D\}$ is linearly independent and $\langle a_i^*, x\rangle = r_i$ for some $x \in S_{R^m_+}$ and each $i \in D\}$. For each $D \in \mathcal{J}$, let $C_D = \{y \in R^m :$ the ith component of y is nonnegative for each $i \in D\}$; then $R^m_+ \subset C_D$, $\text{int}(C_D) \ne \emptyset$, and $C_D^+ = \{y \in R^m_+ :$ the ith component of y is

zero for each $i \in I \setminus D\}$. Therefore, for each $D \in \mathcal{J}$,

$$A^*(C_D^+ \setminus \{0\}) = \left\{ \sum_{i\in D} t_i a_i^* : \ t_i \geq 0 \text{ and } \sum_{i\in D} t_i \neq 0 \right\}.$$

Since $\{a_i^* : \ i \in D\}$ is linearly independent, it follows that $0 \notin A^*(C_D^+ \setminus \{0\})$. Let $X_0 = \{x \in X : \ \langle a_i^*, x\rangle = 0 \text{ for each } i \in I\}$. Then X/X_0 is finite dimensional. For each $x \in X$, let $[x]$ denote the equivalence class containing x in X/X_0, that is, $[x] = x + X_0$. Define $\hat{a}_i^* \in (X/X_0)^*$ such that $\langle \hat{a_i}^*, [x]\rangle = \langle a_i^*, x\rangle$ for each $x \in X$ and $i \in I$. Then, for any $D \subset I$, $\{\hat{a}_i^* : \ i \in D\}$ is linearly independent if and only if $\{a_i^* : \ i \in D\}$ is linearly independent. Let $\hat{S}_{R_+^m} = \{[x] : \ \langle \hat{a}_i^*, [x]\rangle \leq r_i \text{ for each } i \in I\}$ and $\hat{S}_{C_D} = \{[x] : \ \langle \hat{a}_i^*, [x]\rangle \leq r_i \text{ for each } i \in D\}$ for each $D \in \mathcal{J}$. It is easy to verify that $\hat{S}_{R_+^m} = \{[x] : \ x \in S_{R_+^m}\}$ and $\hat{S}_{C_D} = \{[x] : \ x \in S_{C_D}\}$ for each $D \in \mathcal{J}$. Equip X/X_0 with the norm $|||\cdot|||$: $|||[x]||| = \inf\{\|z\| : \ z \in [x] = x + X_0\}$ for each $x \in X$. Then for each $x \in X$ and $D \in \mathcal{J}$,

$$dist([x], \hat{S}_{R_+^m}) = dist(x, S_{R_+^m}) \text{ and } dist([x], \hat{S}_{C_D}) = dist(x, S_{C_D}).$$

Since X/X_0 is finite dimensional, and by [3, Corollary 1.1], one has that for each $x \in X$ there exists $D \in \mathcal{J}$ such that $dist([x], \hat{S}_{R_+^m}) = dist([x], \hat{S}_{C_D})$, that is, $dist(x, S_{R_+^m}) = dist(x, S_{C_D})$. It follows that $dist(x, S_{R_+^m}) \leq \max\{dist(x, S_{C_D}) : \ D \in \mathcal{J}\}$ for each $x \in X$. This shows that (A, R_+^m, b) has property (P).

LEMMA 4.3. *Let Y be a Banach space, $C \subset Y$ a closed convex cone, and $B(C^+) := \{y^* \in C^+ : \ \|y^*\| = 1\}$. Let $g(y) = dist(y, -C)$ for each $y \in Y$. Then the following assertions hold.*

(a) *$\partial g(y) \subset B(C^+)$ for each $y \in Y$ with $g(y) > 0$.*

(b) *$g(y) = \lim_{k\to\infty} \langle y^*, y - h_k\rangle$ for each $y^* \in \partial g(y)$ and each sequence $\{h_k\}$ in $-C$ with $g(y) = \lim_{k\to\infty} \|y - h_k\|$.*

Proof. Let $y \in Y$ and $y^* \in \partial g(y)$. It is easy to verify that $\|y^*\| \leq 1$. Let $\{h_k\}$ be a sequence in $-C$ with $\|y - h_k\| \to dist(y, -C) = g(y)$. Then $g(h_k) = 0$ and

$$\langle y^*, h_k - y\rangle \leq g(h_k) - g(y) = -dist(y, -C).$$

Passing to the limits in

$$g(y) = dist(y, -C) \leq \langle y^*, y - h_k\rangle \leq \|y - h_k\|,$$

it follows that

$$g(y) = \lim_{k\to\infty} \langle y^*, y - h_k\rangle = \lim_{k\to\infty} \|y - h_k\|, \tag{4.3}$$

proving (b). Moreover, if $g(y) > 0$, then (4.3) also entails that $\|y^*\| \geq 1$ and so $\|y^*\| = 1$. It remains to show that $y^* \in C^+$. To do this, let $h \in C$ and $t > 0$. Then $g(-th) = 0$ and $\langle y^*, -th - y\rangle \leq -g(y)$. Letting $t \to +\infty$ in $\langle y^*, h\rangle \geq \frac{g(y) - \langle y^*, y\rangle}{t}$, one has that $\langle y^*, h\rangle \geq 0$ and so $y^* \in C^+$.

THEOREM 4.4. *Let $A : X \to Y$ be a bounded linear operator, $C \subset Y$ a closed convex cone, and $b \in Y$. Assume that the system (A, C, b) has property (P) and $S_C = \{x \in X : \ A(x) \leq_C b\} \neq \emptyset$. Then there exists $\tau \in (0, \ +\infty)$ such that for each $x \in X$,*

$$dist(x, S_C) \leq \tau dist(A(x) - b, -C).$$

Proof. Let $C_1, \ldots, C_m$ be closed convex cones in Y which satisfy (i)–(iii) of Definition 4.1. For each $i \in I = \{1, \ldots, m\}$, pick a $c_i \in \text{int}(C_i)$ and let $\Theta_i := \{y^* \in C_i^+ : \langle y^*, c_i\rangle = 1\}$; it is easy to verify that Θ_i is a bounded weak*-closed (so weak*-compact) subset of Y^* and $C_i^+ = \{ty^* : t \geq 0 \text{ and } y^* \in \Theta_i\}$. For each $i \in I$, let $B_i = \{y^* \in C_i^+ : \|y^*\| = 1\}$; then there exist $0 < \alpha_i < \beta_i < +\infty$ such that

$$B_i \subset [\alpha_i,\ \beta_i]\Theta_i. \tag{4.4}$$

To see (4.4) we use U to denote the unit ball of Y and take $r > 0$ such that $c_i + rU \subset C_i$. Then, for each $y^* \in B_i$, one has that $\langle y^*, c_i - ru\rangle \geq 0$ for each $u \in U$. It follows from $\|y^*\| = 1$ that $\|c_i\| \geq \langle y^*, c_i\rangle \geq r$. Since $\frac{y^*}{\langle y^*, c_i\rangle} \in \Theta_i$, (4.4) is seen to hold with $[\alpha_i,\ \beta_i] = [r,\ \|c_i\|]$. Since the conjugate operator A^* is weak*-weak* continuous, $A^*(\Theta_i)$ is a weak*-compact (hence norm-closed). By (ii) of Definition 4.1, one has that $0 \notin A^*(\Theta_i)$. Therefore,

$$\inf\{\|x^*\| :\ x^* \in A^*(\Theta_i)\} > 0.$$

This and (4.4) imply that

$$\tau_i := \inf\{\|x^*\| :\ x^* \in A^*(B_i)\} > 0, \tag{4.5}$$

valid for each $i \in I$. For each $i \in I$, define $f_i(x) := dist(Ax - b, -C_i)$ and $g_i(y) := dist(y, -C_i)$ for each $x \in X$ and $y \in Y$. Note that f_i and g_i are continuous convex functions, $f_i(x) = g_i(Ax - b)$, $S_{C_i} = \{x \in X :\ f_i(x) \leq 0\}$, and $0 = \inf\{f_i(x) :\ x \in X\}$. Since $\partial f_i(x) = A^*(\partial g_i(Ax - b))$, it follows from Lemma 4.3 that $\partial f_i(x) \subset A^*(B_i)$ for each $x \in X$ with $f_i(x) > 0$. This and (4.5) imply that $\inf\{\|x^*\| :\ x^* \in \partial f_i(x)\} \geq \tau_i$ for each $x \in X$ with $f_i(x) > 0$. It follows from Theorem 3.3 that

$$dist(x, S_{C_i}) \leq \frac{1}{\tau_i} f_i(x) \qquad \forall x \in X.$$

By (iii) of Definition 4.1, one has that for each $x \in X$,

$$dist(x, S_C) \leq \tau \max\{f_i(x) :\ i \in I\},$$

where $\tau = \max\{\frac{\alpha}{\tau_i} :\ i \in I\}$. Since $C \subset C_i$,

$$f_i(x) = dist(Ax - b, -C_i) \leq dist(Ax - b, -C).$$

Hence,

$$dist(x, S_C) \leq \tau dist(Ax - b, -C) \qquad \forall x \in X.$$

This completes the proof.

If X, Y, A, and cone C are, respectively, taken as R^n, $\mathcal{B}(U, R)$, A_U, and C_U in the beginning of this section, then $dist(A_U x - b, -C_U) = \|(A_U x - b)_+\|_\infty$. In view of this and Proposition 4.2, we see that Theorem 4.4 is a generalization of Hoffman's error bound result.

THEOREM 4.5. *Let* $\text{Im}(A) := \{Ax :\ x \in X\}$ *be closed and suppose that* $C + b \subset \text{Im}(A)$. *Then there exists* $\tau \in [0,\ +\infty)$ *such that for each* $x \in X$,

$$dist(x, S_C) \leq \tau dist(Ax - b, -C).$$

Proof. Let $f(x) = dist(Ax - b, -C)$ and $g(x) = dist(y, -C)$ for each $x \in X$ and $y \in Y$. Then $\partial f(x) = A^*(\partial g(Ax - b))$ for each $x \in X$. Let

$$\tau := \sup\{dist(0, A^{-1}(y)) : \ y \in \mathrm{Im}(A) \text{ and } \|y\| = 1\}.$$

It follows from the closedness of $\mathrm{Im}(A)$ and the open mapping theorem that $\tau < +\infty$. For each $x \in X$ with $f(x) > 0$ and $x^* \in \partial f(x)$ there exists $y^* \in \partial g(Ax - b)$ such that $x^* = A^*(y^*)$. Pick a sequence $\{h_k\}$ in $-C$ such that

$$\|Ax - b - h_k\| \to dist(Ax - b, -C) = f(x). \tag{4.6}$$

Since $C + b \subset \mathrm{Im}(A)$, $Ax - b - h_k \in \mathrm{Im}(A)$ for each k. By the definition of τ, for each k there exists $x_k \in X$ such that $Ax_k = Ax - b - h_k$ and $\|x_k\| \le (1 + \frac{1}{k})\tau\|Ax - b - h_k\|$. It follows from $y^* \in \partial g(Ax - b)$ that

$$\begin{aligned} f(x) = g(Ax - b) &\le \langle y^*, Ax - b - h_k\rangle = \langle y^*, Ax_k\rangle = \langle A^*(y^*), x_k\rangle = \langle x^*, x_k\rangle \\ &\le \|x^*\|\|x_k\| \le \|x^*\| \left(1 + \frac{1}{k}\right) \tau\|Ax - b - h_k\|. \end{aligned}$$

This and (4.6) imply that $\|x^*\| \ge \frac{1}{\tau}$. This shows that $\inf\{\|x^*\| : \ x^* \in \partial f(x)\} \ge \frac{1}{\tau}$ for each $x \in X$ with $f(x) > 0$. Since $0 = \inf\{f(x) : \ x \in X\}$ and $S_C = \{x \in X : \ f(x) \le 0\}$, it follows from Theorem 3.3 that for each $x \in X$,

$$dist(x, S_C) \le \tau f(x) = \tau dist(Ax - b, -C).$$

REFERENCES

[1] F. A. Al-Khayyal and J. Kyparisis, *Finite convergence of algorithms for nonlinear programs and variational inequality inequalities*, J. Optim. Theory Appl., 70 (1991), pp. 319–332.

[2] A. Auslender, *Stability in mathematical programming with nondifferentiable data*, SIAM J. Control Optim., 22 (1984), pp. 239–254.

[3] C. Bergthaller and I. Singer, *The distance to a polyhedron*, Linear Algebra Appl., 169 (1992), pp. 111–129.

[4] J. V. Burke and M. C. Ferris, *Weak sharp minima in mathematical programming*, SIAM J. Control Optim., 31 (1993), pp. 1340–1359.

[5] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[6] L. Cromme, *Strong uniqueness*, Numer. Math., 29 (1978), pp. 179–193.

[7] M. A. Goberna, M. A. Lopez, and M. Todorov, *Stability theory for linear inequality systems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 730–743.

[8] A. J. Hoffman, *On approximate solutions of systems of linear inequalities*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 263–265.

[9] H. Hu, *Perturbation analysis of global error bounds for systems of linear inequalities*, Math. Program., 88 (2000), pp. 277–284.

[10] H. Hu and Q. Wang, *On approximate solutions of infinite systems of linear inequalities*, Linear Algebra Appl., 114/115 (1989), pp. 429–438.

[11] A. Jourani, *Hoffman's error bound, local controllability, and sensitivity analysis*, SIAM J. Control Optim., 38 (2000), pp. 947–970.

[12] A. Y. Kruger, *Properties of generalized differentials*, Siberian Math. J., 26 (1985), pp. 822–832.

[13] E. S. Levitin, A. A. Milyutin, and N. P. Osmolovski, *Conditions of high order for a local minima in problems with constraints*, Russian Math. Surveys, 33 (1978), pp. 97–168.

[14] B. S. Mordukhovich and Y. Shao, *Nonsmooth sequential analysis in Asplund space*, Trans. Amer. Math. Soc., 348 (1996), pp. 1235–1280.

[15] K. F. Ng and X. Y. Zheng, *Error bounds for lower semicontinuous functions in normed spaces*, SIAM J. Optim., 12 (2001), pp. 1–17.

[16] J. S. Pang, *Error bounds in mathematical programming*, Math. Program., 79 (1997), pp. 299–332.

[17] R. R. Phelps, *Convex Functions, Monotone Operators and Differentiability*, Lecture Notes in Math. 1364, Springer-Verlag, New York, 1989.
[18] M. Studniarski, *Necessary and sufficient conditions for isolated local minima of nonsmooth functions*, SIAM J. Control Optim., 24 (1986), pp. 1044–1049.
[19] M. Studniarski and D. E. Ward, *Weak sharp minima: Characterizations and sufficient conditions*, SIAM J. Control Optim., 38 (1999), pp. 219–236.
[20] D. E. Ward, *Characterizations of strict local minima and necessary conditions for weak sharp minima*, J. Optim. Theory Appl., 80 (1994), pp. 551–571.
[21] Z. Wu and J. J. Ye, *Sufficient conditions for error bounds*, SIAM J. Optim., 12 (2001), pp. 421–435.
[22] Q. J. Zhu, *The equivalence of several basic theorems for subdifferentials*, Set-Valued Anal., 6 (1998), pp. 171–185.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 6, pp. 1886–1900

CONTROLLABILITY OF THE SEMILINEAR PARABOLIC EQUATION GOVERNED BY A MULTIPLICATIVE CONTROL IN THE REACTION TERM: A QUALITATIVE APPROACH*

A. Y. KHAPALOV†

Abstract. In this paper we are concerned with the global "nonnegative" approximate controllability property of a rather general semilinear heat equation with superlinear term, governed in a bounded domain $\Omega \subset R^n$ by a multiplicative (bilinear) control in the reaction term like $vu(x,t)$, where v is the control. We show that any nonnegative target state in $L^2(\Omega)$ can approximately be reached from any nonnegative, nonzero initial state by applying at most three static bilinear $L^\infty(\Omega)$-controls subsequently in time. This result is further applied to discuss the controllability properties of the nonhomogeneous version of this problem with bilinear term like $v(u(x,t)-\theta(x))$, where θ is given. Our approach is based on an asymptotic technique allowing us to distinguish and make use of the pure diffusion and/or pure reaction parts of the dynamics of the system at hand, while suppressing the effect of a (general) nonlinear term.

Key words. semilinear parabolic equation, approximate controllability, bilinear control

AMS subject classifications. 35, 93

PII. S0363012901394607

1. Introduction.

We consider the following Dirichlet boundary problem, governed in a bounded domain $\Omega \subset R^n$ by a multiplicative (*bilinear*) control $v \in L^\infty(Q_T)$ in the reaction term:

$$\text{(S)} \qquad \frac{\partial u}{\partial t} = \Delta u + vu - f(x,t,u,\nabla u) \quad \text{in} \quad Q_T = \Omega \times (0,T),$$

$$u = 0 \quad \text{in} \ \Sigma_T = \partial\Omega \times (0,T), \qquad u\,|_{t=0} = u_0 \in L^2(\Omega),$$

assuming that f is the given function satisfying the following conditions:

- $f(x,t,q,p)$ is Lebesgue's measurable in x,t,q,p, is continuous in q,p for almost all $(x,t) \in Q_T$;
- there exists a nonnegative function ψ in $L^{1+n/(n+4)}(\Omega)$, $\beta > 0$, and

$$\text{(1.1a)} \qquad r_1 \in \left[0, 1+\frac{4}{n}\right), \qquad r_2 \in \left[0, 1+\frac{2}{n+2}\right)$$

such that

$$\text{(1.1b)} \qquad |f(x,t,q,p)| \le \psi(x,t) + \beta|q|^{r_1} + \beta\|p\|_{R^n}^{r_2} \quad \text{a.e. in } Q_T \text{ for } q \in R,\ p \in R^n;$$

- there exist $\rho > 0$ and $\nu > 0$ such that

$$\text{(1.1c)} \qquad \int_\Omega f(x,t,\phi,\nabla\phi)\phi\,dx \ge (\nu-1)\int_\Omega \|\nabla\phi\|_{R^n}^2 dx - \rho\int_\Omega (1+\phi^2)dx \qquad \forall \phi \in H_0^1(\Omega).$$

*Received by the editors August 31, 2001; accepted for publication (in revised form) September 23, 2002; published electronically February 27, 2003. This work was supported in part by NSF grant DMS-0204037.

http://www.siam.org/journals/sicon/41-6/39460.html

†Department of Pure and Applied Mathematics, Washington State University, Pullman, WA 99164-3113 (khapala@wsu.edu).

For $n = 1$ a simple example of a function f satisfying conditions (1.1a)–(1.1c) is $f(u) = u^3$.

Here and below we use the standard notations for Sobolev spaces such as $H_0^{1,0}(Q_T) = \{\phi | \phi, \phi_{x_i} \in L^2(Q_T),\ i = 1, \dots, n,\ \phi|_{\Sigma_T} = 0\}$ and $H_0^1(\Omega) = \{\phi | \phi, \phi_{x_i} \in L^2(\Omega),\ i = 1, \dots, n,\ \phi|_{\partial\Omega} = 0\}$.

We refer, e.g., to [19, p. 466], where it was shown that system (S), (1.1a)–(1.1c) admits at least one generalized solution in $C([0,T]; L^2(\Omega)) \bigcap H_0^{1,0}(Q_T) \bigcap L^{2+4/n}(Q_T)$, while its uniqueness is not guaranteed.

The multiplicative (bilinear) controls are essential in modeling reaction-diffusion-convection processes controlled by means of so-called catalysts that can accelerate or decelerate the reaction at hand, e.g., various chemical or biological chain reactions. In the context of heat-transfer, v is proportional to the heat-transfer coefficient, which depends on the substance at hand, its surface area, and the environment. If the heat-transfer (or mass-transfer in the case of the diffusion process) involves fluids (air), v also depends on the speed of the fluid. Alternatively, the surface area can be changed when the substance at hand is a polymer (e.g., a planar array of gel fibers can be controlled to maximize the surface area exposed to the surrounding fluid). We also refer to the so-called extended surface applications (fins, pins, studs, etc.) allowing one to increase or decrease the heat-exchange with an ambient fluid.

In this paper we intend to analyze the global controllability properties of the homogeneous bilinear system (S) and its nonhomogeneous version (NHS), given in the next section.

2. Main results. Let us remind the reader that it is said that the system at hand is globally *approximately* controllable in the given (*linear* phase) space H at time $T > 0$ if, by selecting a suitable available control, it can be steered in H from any given initial state into any desirable neighborhood of any desirable target state at time T.

It is immediate that, in general, system (S) is not approximately controllable in any (reasonable) linear space. This can be illustrated by a quick analysis of the linear truncated version of (S) with $f = 0$ (as in (3.2) below). Indeed, in this case the zero-state is the fixed point of the solution mapping, regardless of the choice of control v. In other words, the truncated linear version of (S) cannot be steered anywhere from the zero-state by applying any bilinear control. Furthermore, due to the maximum principle, if, e.g., the initial state $u_0(x)$ is nonnegative, then the maximum principle implies that the corresponding solution $u(x,t)$ to the truncated linear version of (S) must remain nonnegative for all $t > 0$, regardless of the choice of v. Hence, one is unable to reach any of the "negative" target states from a nonnegative initial state.

However, we showed in [15] that in the linear case, i.e., with $f = 0$, the one-dimensional (1-D) version of (S) can be steered in $L^2(0,1)$ from any nonzero, nonnegative initial state u_0 into any desirable neighborhood of any nonnegative target state u_d at a time $T > 0$, which depends on the choice of (u_0, u_d) and the desirable precision of steering, by means of static controls $v = v(x)$, $v \in L^\infty(0,1)$ only. By making use of at most three static bilinear controls, applied subsequently in time, this "nonnegative controllability" result was further extended in [15] to prove the nonnegative controllability of the 1-D semilinear parabolic equation like (S), (1.1a)–(1.1c), admitting multiple solutions, in the case when $n = 1$, $\psi = 0$, $\nu = 1$, and $\rho = 0$ in the sense of the following definition.

DEFINITION 2.1 (see [15]). *We will say that system* (S), (1.1a)–(1.1c), *generally admitting multiple solutions, is nonnegatively globally approximately controllable in*

$L^2(\Omega)$ if for every $\varepsilon > 0$ and nonnegative $u_0, u_d \in L^2(\Omega), u_0 \neq 0$ there exist a $T = T(\varepsilon, u_0, u_d)$ and a bilinear control $v \in L^\infty(Q_T)$ such that for all (i.e., possibly multiple) solutions of (S), (1.1a)–(1.1c), *corresponding to the latter,*

$$\|u(\cdot, T) - u_d\|_{L^2(\Omega)} \leq \varepsilon.$$

The central idea of the method in [15] is to select the bilinear control $v = v(x)$ in such a way that the target state (or its "close" approximation) becomes colinear to the first (nonnegative) eigenfunction for the truncated linear version of (S), which is then approached by the system at hand as t increases. Thus, on the one hand, this method allows one to deal with "practical" relatively small and simple static controls, but, on the other hand, the control time can be relatively large. Moreover, such an approach requires the first eigenvalues to be always simple—hence one needs to assume that $n = 1$.

In the semilinear case another principal limitation of the method of [15] is the assumption that the nonlinear term must be superlinear near the origin as well (i.e., in particular, it must vanish at the origin, which is an equilibrium in this case). In this way it can be assured that the system at hand behaves "almost" like a linear one near the origin, which enabled us to make use of the bilinear controllability properties of the latter. However, this limitation did not allow us to extend the methods of [15] to a nonhomogeneous bilinear system like

$$\text{(NHS)} \qquad \frac{\partial z}{\partial t} = \Delta z + v(z - \theta(x)) - f(x, t, z, \nabla z) \qquad \text{in } Q_T,$$

$$z = 0 \quad \text{in } \Sigma_T, \qquad z\,|_{t=0} = z_0 \in L^2(\Omega),$$

where f does not necessarily vanish at the origin and $\theta \neq 0$ is given.

In the context of heat-transfer the term $v(x,t)(z(x,t) - \theta(x))$ in (NHS) can model the heat-exchange at point x at time t of the given substance with the surrounding medium of temperature $\theta(x)$ according to Newton's law (the classical examples here are the heat equations for a rod or a plate [23]).

In this paper we intend to prove the same, as in [15], nonnegative controllability result, but now (a) in *any* space dimension and (b) within *an arbitrarily small time-interval* $(0, T)$ given in advance. We employ a different qualitative approach which allows us to *eliminate the assumptions of* [15] *on the one dimensionality* of the system at hand and (c) to *get rid of the superlinearity assumption on the nonlinear term near the origin*. Our central idea below is to view the evolution of system (S) as an interaction of the following three dynamics associated with the three terms in the right-hand side of (S):

- *Pure diffusion dynamics*—when $v = 0$ and $f = 0$;
- *Pure reaction dynamics*—caused by the reaction term only, namely, we further associate it with the system

$$\frac{\partial y}{\partial t} = vy \qquad \text{in} \quad Q_T, \tag{2.1}$$

$$y\,|_{t=0} = y_0.$$

- *Nonlinear "disturbance"*—the dynamics caused by the nonlinear term of class (1.1a)–(1.1c).

Accordingly, our strategy to achieve the desirable controllability result will be to try to select the bilinear control in such a way that the corresponding trajectories of (S) can be approximated by those associated with the pure diffusion and/or the pure reaction like in (2.1), while the effect of nonlinearity is to be suppressed. The latter appears to be unavoidable when dealing with a general class of highly nonlinear terms like in (1.1a)–(1.1c).

Our main results are as follows.

THEOREM 2.1. *Let $T > 0$ be given and the pair of the initial and target states $u_0 \in H_0^1(\Omega)$ and $u_d \in L^2(\Omega)$ be such that*

$$\frac{u_d}{u_0} \in H^2(\Omega), \qquad \nabla\left(\frac{u_d}{u_0}\right) \in [L^\infty(\Omega)]^n, \qquad \Delta\left(\frac{u_d}{u_0}\right) \in L^\infty(\Omega), \tag{2.2a}$$

and

$$0 < c_1 \le \frac{u_d(x)}{u_0(x)} \le c_2 < 1 \quad \text{a.e. in } \Omega, \tag{2.2b}$$

where c_1 and c_2 are some positive constants. Then for every $\varepsilon > 0$ there is a $T_* \in (0, T)$ such that for all, i.e., possibly multiple, solutions to (S), (1.1a)–(1.1c) *generated by control*

$$v(x) = \frac{1}{T_*} \ln\left(\frac{u_d(x)}{u_0(x)}\right), \tag{2.3}$$

we have the same uniform estimate

$$\|u(\cdot, T_*) - u_d\|_{L^2(\Omega)} \le \varepsilon. \tag{2.4}$$

Theorem 2.1 can be reformulated as follows.

THEOREM 2.2. *Given $T > 0$, let $v_*(x)$ be any function such that*

$$v_* \in L^\infty(\Omega) \bigcap H^2(\Omega),\ \nabla v_* \in [L^\infty(\Omega)]^n,\ \Delta v_* \in L^\infty(\Omega),\ v_*(x) \le L < 0 \quad \text{a.e. in } \Omega,$$

where L is some negative constant. Then for any $u_0 \in H_0^1(\Omega)$ and every $\varepsilon > 0$ there is a $T_* \in (0, T)$ such that for all, i.e., possibly multiple, solutions to (S), (1.1a)–(1.1c) *with control*

$$v(x) = \frac{1}{T_*} v_*(x),$$

we have the same estimate

$$\|u(\cdot, T_*) - e^{v_*(\cdot)} u_0\|_{L^2(\Omega)} \le \varepsilon.$$

Theorems 2.1 and 2.2 provide the basis for the following nonnegative controllability result within any time-interval $(0, T)$ given in advance.

THEOREM 2.3 (nonnegative controllability of (S)). *Given $T > 0$, assume that the boundary $\partial\Omega$ of domain Ω is of class $C^{3+[n/2]}$ (where $[n/2]$ denotes the largest nonnegative integer which does not exceed $n/2$). Then for every $\varepsilon > 0$ and nonnegative $u_0, u_d \in L^2(\Omega)$, $u_0 \not\equiv 0$, there exists a $T_* = T_*(\varepsilon, u_0, u_d) \in (0, T)$ and a bilinear control $v \in L^\infty(Q_{T_*})$ such that for all (i.e., possibly multiple) solutions of* (S), (1.1a)–(1.1c), *corresponding to the latter,* (2.4) *holds. Suitable v can be selected as a combination of at most three static controls applied subsequently in time.*

An immediate consequence of Theorem 2.3 for the nonhomogeneous system (NHS) is as follows.

THEOREM 2.4 (nonhomogeneous case (NHS)). *Given $T > 0$, let $\theta \in H^2(\Omega) \bigcap H_0^1(\Omega)$ and the boundary $\partial\Omega$ of domain Ω be of class $C^{3+[n/2]}$. For every $\varepsilon > 0$ and $z_0, z_d \in L^2(\Omega)$, $z_0 \neq \theta$, $z_0(x) \geq \theta(x)$, $z_d(x) \geq \theta(x)$, a.e. in Ω, there exist a $T_* = T_*(\varepsilon, z_0, z_d) \in (0, T)$ and a bilinear control $v \in L^\infty(Q_{T_*})$ such that for all (i.e., possibly multiple) solutions of* (NHS), (1.1a)–(1.1c), *corresponding to the latter,* (2.4) *holds with z_0, z_d in place of u_0, u_d. Again, suitable v can be selected as a combination of at most three static controls applied subsequently in time.*

Indeed, to prove Theorem 2.4 it is sufficient to notice that the substitution $w = z - \theta$ transforms (NHS) into

$$\frac{\partial w}{\partial t} = \Delta w + vw - f_*(x, t, w, \nabla w) \text{ in } Q_T, \tag{2.5}$$

$$w = 0 \qquad \text{in } \Sigma_T, \qquad w\,|_{t=0} = z_0 - \theta \in L^2(\Omega),$$

where $f_*(x, t, w, \nabla w) = -\Delta\theta + f(x, t, w+\theta, \nabla w+\theta)$. Making use of Young's inequality and the inequality $(a + b)^\gamma \leq C(a^\gamma + b^\gamma)$ $(a, b, \gamma \geq 0,\ C = C(\gamma) > 0)$, one can check that conditions (1.1a)–(1.1c) hold for this function as well (in general, with a different set of parameters β, ν, ρ). Then Theorem 2.4 follows immediately from Theorem 2.3 applied to (2.5).

Remark 2.1 (some references on bilinear controllability).

- In the pioneering work [4] by Ball, Mardsen, and Slemrod the global approximate controllability of the rod equation $u_{tt} + u_{xxxx} + k(t)u_{xx} = 0$ with hinged ends and of the wave equation $u_{tt} - u_{xx} + k(t)u = 0$ with Dirichlet boundary conditions, where k is control (the axial load), was shown by making use of the nonharmonic Fourier series approach under the additional (nontraditional) assumption that all the modes in the initial data are active. We also refer to [18] exploring the ideas of [4] in the context of simultaneous control of the rod equation and Schrödinger equation.
- In [14] the global approximate controllability of a semilinear heat equation like (S), (1.1a)–(1.1c) was established at any positive time $T > 0$ (fixed in advance) in the case when a *pair* of controls govern the system at hand: (a) the traditional internal either locally distributed or lumped control *and* (b) a piecewise constant bilinear control v. (Recall along these lines that without the latter one does not have the global approximate controllability for this class of PDEs [6], [12], [8].) In one space dimension the method of [14] was further extended in [15] to the case dealing with bilinear controls only, as we discussed in detail in the beginning of this section.
- The works [9] and [16] deal with a different approach to the bilinear controllability, which is as follows. It is known that a rather general class of semilinear parabolic equations with globally Lipschitz terms is approximately controllable by the traditional additive locally distributive controls (see, e.g., [6], [12], [10], [7], [8] and the references therein), while a system like (S), (1.1a)–(1.1c) is globally approximately controllable by the additive static controls with support everywhere in Ω [13]. Denote, e.g., in the latter case, the additive static control by $\alpha(x)$. Then a suitable bilinear control for (S) can be sought as some "well-posed modification" of the expression $v(x, t) = \alpha(x)/u(x, t)$ in the homogeneous case and of the expression $v(x, t) = \alpha(x)/(z(x, t) - \theta(x))$ in the

nonhomogeneous case (NHS). In this way the "original" additive static control $\alpha(x)$ is "transformed," respectively, into the homogeneous bilinear term $v(x,t)u(x,t)$ or into its nonhomogeneous version $v(x,t)(z(x,t)-\theta)$. Note, however, that this approach deals with essentially more "complex" controls (e.g., in terms of numerical implementation), namely, as functions of both x and t. In the paper [9] it was applied in the context of the nonnegative approximate controllability and in [16] it was used to investigate the exact null-controllability of a nonhomogeneous bilinear problem.

- Aside from bilinear controllability, a very close issue is *stabilization* by means of bilinear controls. We can point out only a very limited number of publications in this area in terms of PDEs; see [2], [3], [22]. Regarding the issues of optimal control for bilinear systems, we refer to [20], [5], and the references therein.
- An extensive and thorough bibliography on controllability on bilinear *ODEs* is available; see, e.g., the survey in [1]. On the issue of the qualitative approach in the context of controllability for bilinear ODEs, we refer to [17].

The remainder of this paper is organized as follows. In section 3 we introduce several auxiliary technical estimates, proven in Appendices A and B. In section 4 we prove Theorems 2.1 and 2.2, while Theorem 2.3 is proven in section 5.

3. Auxiliary estimates. In this section we formulate three lemmas containing several estimates, which are heavily used in the proofs of our main results.

Denote $\mathcal{B}(0,T) = C([0,T];\, L^2(\Omega)) \bigcap H_0^{1,0}(Q_T)$ and

$$\|\phi\|_{\mathcal{B}(0,T)} = \left(\max_{t\in[0,T]} \|\phi(\cdot,t)\|^2_{L^2(\Omega)} + 2\nu \int_0^T \int_\Omega \|\nabla\phi\|^2_{R^n} dxds \right)^{1/2},$$

where $\nu > 0$ is from (1.1c).

LEMMA 3.1. *Given $T > 0$ and $v \in L^\infty(\Omega)$, any solution to system* (S), (1.1a)–(1.1c) *(if there are multiple solutions), satisfies the following* two *estimates:*

$$\|u\|_{\mathcal{B}(0,T)}, \qquad \|u\|_{L^{2+4/n}(Q_T)} \le Ce^{(\alpha+\rho)T} \left(\|u_0\|^2_{L^2(\Omega)} + 2\rho T \right)^{1/2}, \tag{3.1}$$

where $\alpha = \|v\|_{L^\infty(\Omega)}$ and C is a generic positive constant (it does not depend on v).

Consider now the truncated version of system (S) as follows:

$$h_t = \Delta h + vh \qquad \text{in} \quad Q_T, \tag{3.2}$$

$$h = 0 \text{ in } \Sigma_T, \qquad h\,|_{t=0} = h_0 \in L^2(\Omega).$$

LEMMA 3.2. *Given $T > 0$, $v \in L^\infty(\Omega)$, $\delta \in (0, 1/4)$, we have the following two estimates for the difference $\xi = u - h$ between any corresponding solution u to* (S), (1.1a)–(1.1c) *(if there are multiple ones), and the unique corresponding solution to* (3.2)*:*

$$\begin{aligned} \|\xi\|_{\mathcal{B}(0,T)}, \|\xi\|_{L^{2+4/n}(Q_T)} \le Ce^{2\sqrt{2}\alpha T} \Bigg\{ & \|u_0 - h_0\|^2_{L^2(\Omega)} \\ & + \frac{1}{\sqrt{\delta}} T^{\frac{n+4}{2(n+2)}(1-\frac{r_1 n}{n+4})} \|u\|^{r_1}_{L^{2+4/n}(Q_T)} \\ & + T^{\frac{n+4}{2(n+2)}(1-\frac{(n+2)r_2}{(n+4)})} \|\nabla u\|^{r_2}_{[L^2(Q_T)]^n} + \|\psi\|_{L^{1+n/(n+4)}(Q_T)} \Bigg\}, \end{aligned} \tag{3.3}$$

where the (generic) constant C does not depend on $\alpha = \|v\|_{L^\infty(\Omega)}$.

Lemmas 3.1 and 3.2 are proven in Appendix A.

Next we have the following result, proven in Appendix B, based on Lemma 3.1.

LEMMA 3.3. *Given $T > 0$, the following estimate holds for solutions to* (3.2) *with $h_0 \in H_0^1(\Omega)$, $v \in L^\infty(\Omega) \bigcap H^2(\Omega)$, $\nabla v \in [L^\infty(\Omega)]^n$, $\Delta v \in L^\infty(\Omega)$, $v(x) \le 0$:*

$$\begin{aligned} &\int\int_{Q_T} (\Delta h)^2 dxdt \ + \ \frac{1}{2}\int_\Omega \|\nabla h(x,T)\|^2_{R^n} dx \\ &\le \frac{1}{2}\int_\Omega \|\nabla h_0\|^2_{R^n} dx + \frac{1}{2}\|\Delta v\|_{L^\infty(\Omega)} TC^2 e^{2\alpha T} \|h_0\|^2_{L^2(\Omega)}, \end{aligned} \tag{3.4}$$

where C is from Lemma 3.1 *and $\alpha = \|v\|_{L^\infty(\Omega)}$.*

4. Proofs of Theorems 2.1 and 2.2.

4.1. Proof of Theorem 2.1: The case $f = 0$. In this subsection we consider the truncated version (3.2) of system (S).

Step 1. Consider h_0 and h_d satisfying the assumptions of Theorem 2.1 in place of u_0 and u_d.

Denote

$$v_*(x) = \ln\left(\frac{h_d(x)}{h_0(x)}\right). \tag{4.1}$$

Then

$$h_d(x) = e^{v_*(x)} h_0(x). \tag{4.2}$$

Remark 4.1. Note that, since (2.1) is, in fact, a linear ODE in $L^2(\Omega)$, in view of (4.1) and (4.2), its solution y satisfies the following property:

$$y(x, 1/s) = h_d(x) \qquad \text{when } y_0 = h_0, v(x) = sv_*(x) \tag{4.3}$$

for any number $s > 0$.

Step 2. To prove Theorem 2.1, we need to show that h_d can be approximated by a suitable solution to (3.2). To this end we intend to study the difference between the solutions to (3.2) and to (2.1) on $(0, T)$.

Consider any $v \in L^\infty(\Omega)$ and denote $g = h - y$. Then, assuming that $y_0 = h_0$ in (2.1) and (3.2), we obtain

$$g_t = vg + \Delta h \qquad \text{in} \quad Q_T,$$

$$g|_{t=0} = 0.$$

Thus

$$g(x,t) = \int_0^t e^{v(x)(t-\tau)} \Delta h(x,\tau) d\tau, \qquad t \in [0,T]. \tag{4.4}$$

Let now v be of the form as in (4.3), namely,

$$v(x) = sv_*(x),$$

where we now treat the positive number s as a parameter (its value will be selected later in Step 4).

Then, since (see (2.2b))

$$\ln c_1 \le v_* \le \ln c_2 < 0,$$

formula (4.4) yields

$$\|g(\cdot,t)\|^2_{L^2(\Omega)} \le \left(\frac{e^{2st\ln c_2}-1}{s\ln c_2}\right)\|\Delta h\|^2_{L^2(Q_t)}, \qquad Q_t=\Omega\times(0,t),\ t\in[0,T]. \tag{4.5}$$

Step 3. Making use of the estimate (3.4) from Lemma 3.3 applied with $v = sv_*$, we derive from (4.5) that

$$\begin{aligned}\|g(\cdot,t)\|^2_{L^2(\Omega)} \ &\le\ \left(\frac{e^{2st\ln c_2}-1}{s\ln c_2}\right)\\ &\times\left(\frac{1}{2}\int_\Omega\|\nabla h_0\|^2_{R^n}dx\ +\ \frac{1}{2}s\|\Delta v_*\|_{L^\infty(\Omega)}tC^2e^{2\alpha st}\|h_0\|^2_{L^2(\Omega)}\right),\end{aligned} \tag{4.6}$$

where $\alpha = \|v_*\|_{L^\infty(\Omega)}$.

Step 4. Select now $s > 0$ and $T_* \in (0,T)$ such that

$$T_* = \frac{1}{s} \text{ or } T_*s = 1. \tag{4.7}$$

Then we obtain from (4.6), (4.7), and the property (4.3) that

$$\|g(\cdot,T_*)\|_{L^2(\Omega)} = \|h(\cdot,T_*)-y(\cdot,T_*)\|_{L^2(\Omega)} = \|h(\cdot,T_*)-h_d\|_{L^2(\Omega)} \to 0 \tag{4.8}$$

as $s\to\infty$ (or $T_*\to 0+$), which ensures (2.4) for any given in advance ε for some pair (s,T_*) as in (4.7). This ends the proof of Theorem 2.1 when $f = 0$. □

4.2. Proof of Theorem 2.1: The general case. It follows from the above argument by making use of Lemma 3.2, in which we evaluated the difference between the (possible multiple) solutions u to (S) and h to (3.2) in a uniform way as given in (3.3). It follows from (3.3) that under condition (4.7) and with $h_0 = u_0$,

$$\|u(\cdot,T_*)-h(\cdot,T_*)\|_{L^2(\Omega)}\ \to\ 0$$

as $s = 1/T_* \to \infty$, which ensures (2.4) whenever it holds for h as in (4.8). This ends the proof of Theorem 2.1 in the general case. □

4.3. Proof of Theorem 2.2. It is immediate from the proof of Theorem 2.1, since the assumptions on h_0 and h_d in the latter are used specifically to ensure the properties of v_* required in Theorem 2.2. □

5. Proof of Theorem 2.3.

5.1. Proof of Theorem 2.3: The case $f = 0$. Again we study first the truncated problem (3.2).

Consider any pair of initial and target states h_0, $h_d \in L^2(\Omega)$, which are nonnegative (almost everywhere) in Ω and $h_0 \ne 0$. Since we study the issue of approximate controllability and because the set of infinitely differentiable functions with compact

support (denoted by $C_0^\infty(\Omega)$) is dense in $L^2(\Omega)$, without loss of generality we can further assume that

$$h_d \in C_0^\infty(\Omega), \qquad h_d \neq 0, \quad h_d(x) \geq 0 \quad \forall x \in \Omega. \tag{5.1}$$

We plan to approximate h_d by using three static bilinear controls, applied subsequently in time:

(a) First, we will use $v = 0$ on some time-interval $(0, t_1)$ to steer our system to a state $h(\cdot, t_1)$, which is strictly positive in the interior of Ω.

(b) Second, we will use a relatively "large" positive constant control v on some time-interval (t_1, t_2) to steer our system to a state which is "larger" than the given h_d in (5.1).

(c) Finally, we will use a static control as described in Theorem 2.1 on some time-interval (t_2, t_3) to steer our system further to a desirable neighborhood of h_d.

Step 1. Pick any $t_1 > 0$ and apply in (3.2) the zero bilinear control $v = 0$ on $(0, t_1)$. Then, at time t_1 system (3.2) reaches the state

$$h(\cdot, t_1) \in H_0^1(\Omega) \bigcap H^{3+[n/2]}(\Omega) \subset C^2(\bar{\Omega}). \tag{5.2}$$

(We refer, e.g., to [21] for the corresponding regularity and embedding results.)

Note also that, due to the smoothing effect, the solution h to (3.2) is classical in $\bar{\Omega} \times [\beta, t_1]$, for any $\beta \in (0, t_1]$ [21]. Furthermore, due to the strong maximum principle (see, e.g., [11]),

$$h(x, t_1) > 0 \quad \text{in the interior of } \Omega, \qquad h(x, t_1)|_{\partial\Omega} = 0. \tag{5.3}$$

Step 2. Consider any $t_2 > t_1$. On the interval (t_1, t_2) we apply a positive constant control $v(x) = v$ (its value will be chosen later). Then for the corresponding solution h to (3.2) on (t_1, t_2) we have

$$h(x, t_2) = e^{v(t_2 - t_1)} \sum_{k=1}^{\infty} e^{\lambda_k (t_2 - t_1)} \left(\int_\Omega h(r, t_1) \omega_k(r) dr \right) \omega_k(x), \tag{5.4}$$

where λ_k ($\lambda_k \to -\infty$ as $k \to \infty$) and $\omega_k(x)$ ($\|\omega_k\|_{L^2(\Omega)} = 1$), $k = 1, \ldots$, are, respectively, the eigenvalues and eigenfunctions associated with the spectral problem $\Delta\omega = \lambda\omega$, $\omega|_{\partial\Omega} = 0$ in $H_0^1(\Omega)$.

Consider any number $\gamma > 1$ (its value will be chosen more precisely a little bit later) and select a constant (in t and x) control $v > 0$ such that

$$e^{v(t_2 - t_1)} = \gamma, \text{ namely, } v = \frac{\ln \gamma}{t_2 - t_1}. \tag{5.5a}$$

(Thus, v depends on γ and $t_2 - t_1$.)

Then, it follows from (5.4) that with this control

$$h(\cdot, t_2) \to \gamma h(\cdot, t_1) \text{ as } t_2 \to t_1 + \text{ in } C(\bar{\Omega}), \tag{5.5b}$$

as implied by the estimate

$$\|h(\cdot, t_2) - \gamma h(\cdot, t_1)\|_{C(\bar{\Omega})} \leq C \|h(\cdot, t_2) - \gamma h(\cdot, t_1)\|_{H^{1+[n/2]}(\Omega)}$$

$$\leq C \left(\sum_{k=1}^{\infty} \lambda_k^{1+[n/2]} \left(e^{\lambda_k (t_2 - t_1)} - 1 \right)^2 \left(\int_\Omega h(r, t_1) \omega_k(r) dr \right)^2 \right)^{1/2},$$

where $C > 0$ is a (generic) constant associated with the continuous embedding $H^{1+[n/2]}(\Omega) \subset C(\bar{\Omega})$ (see, e.g., [21]).

Select now the value of $\gamma > 1$ in such a way that

$$\gamma h(x, t_1) \geq h_d(x) + 1 \qquad \forall x \in \operatorname{supp} h_d,$$

which is possible due to (5.3) and (5.1), where supp h_d stands for the set of all x where $h_d(x) \neq 0$ (i.e., where $h_d(x) > 0$).

For this γ and any given $\sigma \in (0, 1)$ (to be selected more precisely later) select any positive number $t_2 > t_1, t_2 = t_2(\gamma, \sigma)$ and v as in (5.5a) such that

$$\gamma h(x, t_1) + \sigma/2 \geq h(x, t_2) \geq \gamma h(x, t_1) - \sigma/2 \geq -\sigma/2 \qquad \forall x \in \Omega, \tag{5.6a}$$

$$h(x, t_2) \geq h_d(x) \qquad \forall x \in \text{ supp } h_d. \tag{5.6b}$$

This is possible due to (5.1), (5.3), and (5.5b).

Step 3. We will now apply Theorem 2.2 to the system (3.2) on some interval (t_2, t_3) with the initial state $h(x, t_2)$ and the static control

$$v(x) = \frac{1}{t_3 - t_2} v_\sigma(x),$$

where

$$v_\sigma = \ln\left(\frac{h_d + \sigma^2/2}{h(\cdot, t_2) + \sigma}\right) \in C^2(\bar{\Omega}).$$

(Note that the additional "regularizing" terms $\sigma^2/2$ and σ ensure that the argument of the logarithmic function in the above is positive everywhere in $\bar{\Omega}$.)

Since, in view of (5.6a)–(5.6b),

$$\frac{\sigma^2/2}{\max_{x\in\bar{\Omega}} h(x, t_2) + \sigma} \leq \frac{h_d(x) + \sigma^2/2}{h(x, t_2) + \sigma}$$

$$\leq \begin{cases} \frac{h(x,t_2)+\sigma-\sigma+\sigma^2/2}{h(x,t_2)+\sigma} \leq 1 - \frac{\sigma-\sigma^2/2}{\max_{x\in\bar{\Omega}} h(x,t_2)+\sigma} & \text{for } \in \text{ supp } h_d, \\ \frac{\sigma^2/2}{-\sigma/2+\sigma} = \sigma & \text{for } x \in \Omega \backslash \text{ supp } h_d, \end{cases}$$

we have

$$\ln\left(\frac{\sigma^2/2}{\max_{x\in\bar{\Omega}} h(x, t_2) + \sigma}\right) < v_\sigma(x)$$

$$\leq \ln\left(\max\left\{s, 1 - \frac{\sigma - \sigma^2/2}{\max_{x\in\bar{\Omega}} h(x, t_2) + \sigma}\right\}\right) < 0 \qquad \text{in } \Omega.$$

According to Theorem 2.2, this v will steer (3.2) in $L^2(\Omega)$ at some time t_3 from $h(\cdot, t_2) \in H_0^1(\Omega)$ as close as we wish to a state

$$e^{v_\sigma(x)} h(x, t_2) = h(x, t_2)\left(\frac{h_d(x) + \sigma^2/2}{h(x, t_2) + \sigma}\right), \tag{5.7}$$

provided that t_3 is sufficiently close to t_2 from the right. For example, there is a $t_3 > t_2$ ($t_3 = t_3(\sigma)$) such that

$$\left\| h(\cdot, t_3) - h(\cdot, t_2)\left(\frac{h_d + \sigma^2/2}{h(\cdot, t_2) + \sigma}\right)\right\|_{L^2(\Omega)} \leq \sigma. \tag{5.8}$$

To finish the proof of Theorem 2.3 for the case $f = 0$, it remains to notice that, in view of (5.6a)–(5.6b), the expression in (5.7) converges in $L^2(\Omega)$ to the desirable target state h_d as $\sigma \to 0+$.

Indeed, we have

$$\begin{aligned}\left|h(x,t_2)\left(\frac{h_d(x)+\sigma^2/2}{h(x,t_2)+\sigma}\right)-h_d(x)\right| &= \left|\frac{h(x,t_2)\sigma^2/2-h_d(x)\sigma}{h(x,t_2)+\sigma}\right| \\ &\le \sigma^2/2\frac{h(x,t_2)}{h(x,t_2)+\sigma}+\sigma\frac{h_d(x)}{h(x,t_2)+\sigma} \le \sigma^2/2+\sigma\end{aligned} \tag{5.9}$$

for all $x \in \operatorname{supp} h_d$, where, in view of (5.6b), $h(x,t_2) \ge h_d(x) > 0$ and hence

$$0 \le \frac{h(x,t_2)}{h(x,t_2)+\sigma} \le 1, \tag{5.10}$$

and

$$0 \le \frac{h_d(x)}{h(x,t_2)+\sigma} \le \frac{h(x,t_2)+\sigma}{h(x,t_2)+\sigma} = 1. \tag{5.11}$$

In turn, since h_d vanishes elsewhere, for $x \in \Omega\setminus \operatorname{supp} h_d$ we have

$$h(x,t_2)\left(\frac{h_d(x)+\sigma^2/2}{h(x,t_2)+\sigma}\right)-h_d(x) = h(x,t_2)\left(\frac{\sigma^2/2}{h(x,t_2)+\sigma}\right), \tag{5.12}$$

where, due to (5.6a),

$$\left|h(x,t_2)\left(\frac{\sigma^2/2}{h(x,t_2)+\sigma}\right)\right| \le \left|h(x,t_2)\left(\frac{\sigma^2/2}{-\sigma/2+\sigma}\right)\right| \le (\gamma\|h(\cdot,t_1)\|_{C(\bar\Omega)}+\sigma/2)\sigma. \tag{5.13}$$

Thus, combining (5.9)–(5.13), we obtain that

$$\left\|h(\cdot,t_2)\left(\frac{h_d+\sigma^2/2}{h(\cdot,t_2)+\sigma}\right)-h_d\right\|_{C(\bar\Omega)} \le \sigma^2+\sigma+\sigma\gamma\|h(\cdot,t_1)\|_{C(\bar\Omega)} \to 0$$

as $\sigma \to 0+$, which, in view of (4.8), completes the proof of Theorem 2.3 in the case when $f = 0$. □

5.2. Proof of Theorem 2.3: The general case. As in the case of Theorem 2.1, it follows from estimate (3.3) evaluating the difference between the solutions u to (S) and h to (3.2) (uniformly with respect to possible multiple solutions to (S)).

It follows from (3.3) that whenever the product of the $L^\infty(\Omega)$-norm of the static bilinear control $v(x)$, applied, say, on the time-interval (a,b), and its duration $(b-a)$, namely, $\|v\|_{L^\infty(\Omega)}(b-a)$ remains bounded, we have

$$\|u(\cdot,b)-h(\cdot,b)\|_{L^2(\Omega)} \to 0,$$

provided that $b \to a+$ and $\|u(\cdot,a)-h(\cdot,a)\|_{L^2(\Omega)} \to 0$.

Indeed, in the above,

- on the interval $(0,t_1)$ we used $v = 0$ and can select t_1 as small as we wish;
- on the interval (t_1,t_2) we applied a constant v as in (5.5a) such that for any (fixed) $\gamma > 1$ we have $v(t_2-t_1) = \ln\gamma$;

- on the interval (t_2, t_3) we used the condition like in (4.7), exactly as it is described in the proof of Theorem 2.1 in the general case. Again, $(t_3 - t_2)$ can be arbitrarily small.

Thus applying (3.3) subsequently three times on the aforementioned intervals, while selecting sufficiently small t_i, $i = 1, 2, 3$, and the bilinear controls as described in the above, we can ensure (2.4). This completes the proof of Theorem 2.3 in the general case. □

Appendix A. Proof of Lemmas 3.1 and 3.2.

A.1. Proof of Lemma 3.1. Recall (see, e.g., [19]) that $f(\cdot, \cdot, u, \nabla u) \in L^{1+n/(n+4)}(Q_T)$ and that the following energy equality holds for (S) treated as a linear equation with the source term $f(x, t, u, \nabla u)$, e.g., [19, p. 142]:

$$\frac{1}{2}\|u\|^2_{L^2(\Omega)}|_0^t + \int_0^t \int_\Omega \left(\|\nabla u\|^2_{R^n} - vu^2 + f(x, s, u, \nabla u)u\right) dxds = 0 \qquad \forall t \in [0, T]. \tag{A.1}$$

Here and everywhere below, if there exist several solutions to (S), we always deal separately with a selected one, while noticing that all the estimates hold uniformly.

Combining (A.1) and (1.1c) yields for $t \in [0, T]$

$$\begin{aligned} &\|u(\cdot, t)\|^2_{L^2(\Omega)} + 2\nu \int_0^t \int_\Omega \|\nabla u\|^2_{R^n}\, dxds \\ &\le \|u_0\|^2_{L^2(\Omega)} + 2(\alpha + \rho) \int_0^t \int_\Omega u^2 dxds + 2\rho T \\ &\le \left(\|u_0\|^2_{L^2(\Omega)} + 2\rho T\right) + 2(\alpha + \rho) \int_0^t \left(\|u(\cdot, \tau)\|^2_{L^2(\Omega)} + 2\nu \int_0^\tau \int_0^1 \|\nabla u\|^2_{R^n}\, dxds\right) d\tau. \end{aligned} \tag{A.2}$$

Applying the Gronwall–Bellman inequality to (A.2) yields the first estimate in (3.1) with respect to the $\mathcal{B}(0, T)$-norm with $C = \sqrt{2}$. The second estimate follows by the continuity of the embedding of $\mathcal{B}(0, T)$ into $L^{2+4/n}(Q_T)$ (e.g., [19, pp. 467, 475]), due to which

$$\|\xi\|_{L^{2+4/n}(Q_T)} \le c\|\xi\|_{\mathcal{B}(0,T)} \tag{A.3}$$

for some constant $c > 0$ independent of T. In this case $C = c\sqrt{2}$ in (3.1). This ends the proof of Lemma 3.1. □

A.2. Proof of Lemma 3.2. We now intend to evaluate the difference between any possible (multiple) solution u to (S) and its truncated version (3.2).

Denote $\xi = u - h$; then

$$\xi_t = \Delta\xi + v\xi - f(x, t, u, \nabla u) \qquad \text{in} \quad Q_T, \tag{A.4}$$

$$\xi|_{\Sigma_T} = 0, \qquad \xi\,|_{t=0} = u_0 - h_0.$$

Multiplying (A.4) by ξ and integrating it by parts in $Q_t = \Omega \times (0, t)$, we obtain

the following chain of estimates for all $t \in [0,T]$:

$$\begin{aligned}
&\|\xi(\cdot,t)\|^2_{L^2(\Omega)} + 2\nu \int_0^t \int_\Omega \|\nabla\xi\|^2_{R^n}(x,s)\, dxds \\
&= \|\xi(\cdot,0)\|^2_{L^2(\Omega)} + 2\int_0^t \int_\Omega v\xi^2 dxds - 2\int_0^t \int_\Omega \xi f(x,s,u,\nabla u) dxds \\
&\le \|\xi(\cdot,0)\|^2_{L^2(\Omega)} + 2\alpha \int_0^t \int_0^1 \xi^2 dxds + 2\|\xi\|_{L^{2+4/n}(Q_t)} \|f(x,t,u,\nabla u)\|_{L^{1+n/(n+4)}(Q_t)} \\
&\le \|\xi(\cdot,0)\|^2_{L^2(\Omega)} + 2\alpha \int_0^t \int_\Omega \xi^2 dxds + 2c\|\xi\|_{\mathcal{B}(0,t)} \|f(x,t,u,\nabla u)\|_{L^{1+n/(n+4)}(Q_T)} \\
&\le \|\xi(\cdot,0)\|^2_{L^2(\Omega)} + 2\alpha \int_0^t \|\xi\|^2_{\mathcal{B}(0,s)} ds + \delta\|\xi\|^2_{\mathcal{B}(0,t)} + \frac{c^2}{\delta} \|f(\cdot,\cdot,u,\nabla u)\|^2_{L^{1+n/(n+4)}(Q_T)},
\end{aligned} \tag{A.5}$$

where we made use of (A.3) and Hölder's and Young's inequalities.

It follows from (A.5) that we have

$$\|\xi\|^2_{\mathcal{B}(0,t)} \le 2\|\xi(\cdot,0)\|^2_{L^2(\Omega)} + 4\alpha \int_0^t \|\xi\|^2_{\mathcal{B}(0,s)} ds + 2\delta\|\xi\|^2_{\mathcal{B}(0,t)}$$

$$+ \frac{2c^2}{\delta} \|f(\cdot,\cdot,u,\nabla u)\|^2_{L^{1+n/(n+4)}(Q_T)} \qquad \forall t \in [0,T].$$

Thus,

$$\begin{aligned}
\|\xi\|^2_{\mathcal{B}(0,t)} \le 4\|\xi(\cdot,0)\|^2_{L^2(\Omega)} &+ 8\alpha \int_0^t \|\xi\|^2_{\mathcal{B}(0,s)} ds \\
&+ \frac{4c^2}{\delta} \|f(\cdot,\cdot,u,\nabla u)\|^2_{L^{1+n/(n+4)}(Q_T)},
\end{aligned} \tag{A.6}$$

provided that $0 < \delta < \frac{1}{4}$.

Making use of the Gronwall–Bellman inequality, we derive from (A.6) that

$$\|\xi\|_{\mathcal{B}(0,T)} \le e^{2\sqrt{2}\alpha T} \left(2\|\xi(\cdot,0)\|_{L^2(\Omega)} + \frac{2c}{\sqrt{\delta}} \|f(\cdot,\cdot,u,\nabla u)\|_{L^{1+n/(n+4)}(Q_T)} \right). \tag{A.7}$$

Now, using (1.1b) and Hölder's inequality (as in [19, p. 469], [13, p. 863]), we obtain

$$\begin{aligned}
\|f(\cdot,\cdot,u,\nabla u)\|_{L^{1+n/(n+4)}(Q_T)} &\le \beta T^{\frac{n+4}{2(n+2)}(1-\frac{r_1 n}{n+4})} \|u\|^{r_1}_{L^{2+4/n}(Q_T)} \\
&+ \beta T^{\frac{n+4}{2(n+2)}(1-\frac{(n+2)r_2}{(n+4)})} \|\nabla u\|^{r_2}_{[L^2(Q_T)]^n} + \|\psi\|_{L^{1+n/(n+4)}(Q_T)}.
\end{aligned} \tag{A.8}$$

Combining (A.8), (A.7) yields the result of Lemma 3.2. □

Appendix B. Proof of Lemma 3.3. We need to evaluate Δh in $L^2(Q_T)$, where h satisfies (3.2) with $h_0 \in H^1_0(\Omega)$.

Consider any

$$v \in L^\infty(\Omega) \bigcap H^2(\Omega), \quad \nabla v \in [L^\infty(\Omega)]^n, \quad \Delta v \in L^\infty(\Omega), \quad v(x) \le 0 \qquad \text{in } \Omega. \tag{B.1}$$

Multiplying (3.2) by Δh and further integrating it over Q_T yields

$$\int\int_{Q_T} \left(-h_t \Delta h + (\Delta h)^2\right) dxdt = -\int\int_{Q_T} vh\Delta h dxdt. \tag{B.2}$$

In turn,

$$-\int\int_{Q_T} h_t \Delta h dxdt = \frac{1}{2}\int_\Omega \sum_{k=1}^n h_{x_k}^2(x,T)dx - \frac{1}{2}\int_\Omega \sum_{k=1}^n h_{x_k}^2(x,0)dx,$$

while, having (B.1) in mind,

$$-\int\int_{Q_T} vh\Delta h dxdt = -\int\int_{Q_T} \sum_{k=1}^n vhh_{x_k x_k} dxdt$$

$$= \int\int_{Q_T} \sum_{k=1}^n vh_{x_k}^2 dxdt + \frac{1}{2}\int\int_{Q_T} \sum_{k=1}^n v_{x_k}(h^2)_{x_k} dxdt \le \frac{1}{2}\|\Delta v\|_{L^\infty(\Omega)} \int\int_{Q_T} h^2 dxdt.$$

Combining all the above yields

$$\begin{aligned} &\int\int_{Q_T} (\Delta h)^2 dxdt + \frac{1}{2}\int_\Omega \|\nabla h(x,T)\|_{R^n}^2 dx \\ &\le \frac{1}{2}\int_\Omega \|\nabla h_0\|_{R^n}^2 dx + \frac{1}{2}\|\Delta v\|_{L^\infty(\Omega)} \int\int_{Q_T} h^2 dxdt \\ &\le \frac{1}{2}\int_\Omega \|\nabla h_0\|_{R^n}^2 dx + \frac{1}{2}\|\Delta v\|_{L^\infty(\Omega)} T \max_{t\in[0,T]} \int_\Omega h^2(x,t)dx. \end{aligned} \tag{B.3}$$

From (B.3) and estimate (3.1) (applied to system (3.2)) we obtain the estimate (3.4). This ends the proof of Lemma 3.3. □

REFERENCES

[1] A. Baciotti, *Local Stabilizability of Nonlinear Control Systems,* Ser. Adv. Math. Appl. Sci. 8, World Scientific, River Edge, NJ, 1992.

[2] J.M. Ball and M. Slemrod, *Feedback stabilization of semilinear control systems*, Appl. Math. Optim., 5 (1979), pp. 169–179.

[3] J.M. Ball and M. Slemrod, *Nonharmonic Fourier series and the stabilization of distributed semi-linear control systems*, Comm. Pure Appl. Math., 32 (1979), pp. 555–587.

[4] J.M. Ball, J.E. Mardsen, and M. Slemrod, *Controllability for distributed bilinear systems*, SIAM J. Control Optim., 20 (1982), pp. 575–597.

[5] M.E. Bradley, S. Lenhart, and J. Yong, *Bilinear optimal control of the velocity term in a Kirchhoff plate equation*, J. Math. Anal. Appl., 238 (1999), pp. 451–467.

[6] C. Fabre, J.-P. Puel, and E. Zuazua, *Approximate controllability of the semilinear heat equations*, Proc. Roy. Soc. Edinburgh Sect. A, 125 (1995), pp. 31–61.

[7] E. Fernández-Cara and E. Zuazua, *The cost of approximate controllability for heat equations: The linear case*, Adv. Differential Equations, 5 (2000), pp. 465–514.

[8] E. Fernández-Cara and E. Zuazua, *Null and approximate controllability for weakly blowing-up semilinear heat equations,* Ann. Inst. H. Poincaré Anal. Non Linéaire, 17 (2000), pp. 583–616.

[9] L.A. Fernández, *Controllability of some semilinear parabolic problems with multiplicative control*, in Proceedings of the Fifth SIAM Conference on Control and Its Applications, San Diego, CA, 2001.

[10] L.A. Fernández and E. Zuazua, *Approximate controllability for the semilinear heat equation involving gradient terms*, J. Optim. Theory Appl., 101 (1999), pp. 307–328.
[11] A. Friedman, *Partial Differential Equations of Parabolic Type,* Prentice–Hall, Englewood Cliffs, NJ, 1964.
[12] A. Fursikov and O. Imanuvilov, *Controllability of Evolution Equations,* Lect. Note Ser. Seoul 34, Seoul National University, Seoul, 1996.
[13] A.Y. Khapalov, *Some aspects of the asymptotic behavior of the solutions of the semilinear heat equation and approximate controllability*, J. Math. Anal. Appl., 194 (1995), pp. 858–882.
[14] A.Y. Khapalov, *Bilinear control for global controllability of the semilinear parabolic equations with superlinear terms*, in Control of Nonlinear Distributed Parameter Systems, G. Chen, I. Lasiecka, and J. Zhou, eds., Marcel Dekker, New York, 2001, pp. 139–155.
[15] A.Y. Khapalov, *Global non-negative controllability of the semilinear parabolic equation governed by bilinear control*, ESAIM Control Optim. Calc. Var., 7 (2002), pp. 269–283.
[16] A.Y. Khapalov, *On bilinear controllability of the parabolic equation with the reaction-diffusion term satisfying Newton's Law*, J. Comput. Appl. Math., 21 (2002), pp. 1–23.
[17] A.Y. Khapalov and R.R. Mohler, *Reachable sets and controllability of bilinear time-invariant systems: A qualitative approach*, IEEE Trans. Automat. Control, 41 (1996), pp. 1342–1346.
[18] K. Kime, *Simultaneous control of a rod equation and a simple Schrödinger equation*, Systems Control Lett., 24 (1995), pp. 301–306.
[19] O.H. Ladyzhenskaya, V.A. Solonikov, and N.N. Ural'ceva, *Linear and Quasi-Linear Equations of Parabolic Type*, AMS, Providence, RI, 1968.
[20] S. Lenhart, *Optimal control of convective-diffusive fluid problem*, Math. Models Methods Appl. Sci., 5 (1995), pp. 225–237.
[21] V.P. Mikhailov, *Partial Differential Equations,* Mir, Moscow, 1978.
[22] S. Müller, *Strong convergence and arbitrarily slow decay of energy for a class of bilinear control problems*, J. Differential Equations, 81 (1989), pp. 50–67.
[23] A.N. Tichonov and A.A. Samarski, *Partial Differential Equations of Mathematical Physics*, Vol. 1, Holden–Day, San Francisco, CA, 1964.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 6, pp. 1901–1921

CONFIGURATION CONTROLLABILITY OF MECHANICAL SYSTEMS UNDERACTUATED BY ONE CONTROL*

JORGE CORTÉS† AND SONIA MARTÍNEZ‡

Abstract. We investigate local configuration controllability for mechanical control systems within the affine connection formalism. We rely on previous results on controllability and series expansions for the evolution of mechanical systems starting from rest. Extending the work by Lewis for the single-input case, we are able to characterize local configuration controllability for systems with n degrees of freedom and $n-1$ input forces.

Key words. nonlinear control, configuration controllability, symmetric product

AMS subject classifications. 53B05, 70Q05, 93B03, 93B05, 93B29

PII. S0363012900374099

1. Introduction.

Mechanical control systems belong to a class of nonlinear systems whose controllability properties have not been fully characterized yet. Much work has been devoted to the study of their rich geometrical structure, both in the Hamiltonian framework (see [30] and references therein) and in the Lagrangian one, which has been receiving increasing attention in the last few years [5, 8, 17, 21, 23, 24, 25, 31]. This research is providing new insights and a bigger understanding of the accessibility and controllability aspects associated with them. In particular, the affine connection formalism was revealed to be very useful for modeling different types of mechanical systems, such as natural ones (Lagrangian equal to kinetic energy minus potential energy) [24, 25], with symmetries [5, 9], with nonholonomic constraints [6, 23], etc. and, on the other hand, it has led to the development of some new techniques and control algorithms for approximate trajectory generation in controller design [4, 37]. Certainly, we shall see further progress in these directions in future years.

Underactuated mechanical control systems are interesting to study both from a theoretical and a practical point of view. From a theoretical perspective, they offer a control challenge as they have nonzero drift, their linearization at zero velocity is not controllable, they are not static feedback linearizable, and it is not known if they are dynamic feedback linearizable. That is, they are not amenable to standard techniques in control theory [13, 30]. From the practical point of view, they appear in numerous applications as a result of design choices motivated by the search for less costly devices, or as a result of a failure regime in fully actuated mechanical systems.

The work by Lewis and Murray [24, 25] on simple mechanical control systems has rendered strong conditions for configuration accessibility and sufficient conditions

*Received by the editors June 21, 2000; accepted for publication (in revised form) August 7, 2002; published electronically February 27, 2003. This research was partially supported by FPU and FPI grants from the Spanish Ministerio de Educación y Cultura and Ministerio de Ciencia y Tecnología, respectively, and grant DGICYT PGC2000-2191-E.

http://www.siam.org/journals/sicon/41-6/37409.html

†Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, 1308 W. Main St., Urbana, IL 61801 (jcortes@uiuc.edu). Former address: Laboratory of Dynamical Systems, Mechanics and Control, Instituto de Matemáticas y Física Fundamental, Consejo Superior de Investigaciones Científicas, Serrano 123, 28006 Madrid, Spain.

‡Escola Universitària Politècnica de Vilanova i la Geltrú, Universidad Politécnica de Cataluña, Av. V. Balaguer s/n, Vilanova i la Geltrú 08800, Spain (soniam@mat.upc.es). Former address: Laboratory of Dynamical Systems, Mechanics and Control, Instituto de Matemáticas y Física Fundamental, Consejo Superior de Investigaciones Científicas, Serrano 123, 28006 Madrid, Spain.

for configuration controllability. The conditions for the latter are based on the sufficient conditions that Sussmann obtained for general affine control systems [35]. It is worth noting that these conditions are not invariant under input transformations. As controllability is the more interesting property in practice, more research is needed in order to sharpen the configuration controllability conditions. Whatever these conditions might be, they will be harder to check than the ones for accessibility, since controllability is inherently a more difficult property to establish [14, 33]. Lewis [21] investigated the single-input case, building on previous results by Sussmann for general scalar-input systems [34]. The recent work by Bullo [3] on series expansions for the evolution of a mechanical control system starting from rest gave the necessary tools to tackle this problem in the much more involved multi-input case. In this paper, we characterize local configuration controllability for systems whose number of inputs and degrees of freedom differ by one. Examples include autonomous vehicles (like aircraft takeoff and landing models [11, 28], underwater vehicles [32]), robotic manipulators with a passive joint [26], and locomotion devices (such as the robotic leg [23] or the quadrotor [29]). In addition, fully actuated mechanical systems may temporarily suffer from an actuator failure turning them into underactuated systems by one control, in which case the knowledge of their controllability properties becomes relevant within a robust design perspective. Interestingly, the differential flatness properties of this type of underactuated mechanical control systems have also been characterized in intrinsic geometric terms [32].

Both results, Lewis's and ours, can be seen as particular cases of the following conjecture, which remains open: *The system is locally configuration controllable at a point if and only if there exists a basis of inputs satisfying the sufficient conditions for local configuration controllability at that point.* The conjecture relies on the fact we have mentioned before: the lack of invariance of the sufficient conditions under input transformations. It is remarkable to note that local controllability has not been characterized yet for general control systems, even for the single-input case (in this regard see [12, 34, 35]).

The paper is organized as follows. In section 2, we describe the affine connection framework for mechanical control systems and recall the controllability notions we shall consider on them. In section 3 we review the existing results concerning configuration controllability [24, 25] and the series expansion for the evolution of a mechanical control system starting from rest developed by Bullo in [3]. In section 4 we briefly recall the single-input case solved by Lewis and properly state his conjecture. Section 5 contains the main contributions of this paper. In section 6 we treat two examples to illustrate the results. Finally, we present our conclusions in section 7.

2. Simple mechanical control systems. Let Q be a n-dimensional manifold. We will denote by TQ the tangent bundle of Q, by $\mathfrak{X}(Q)$ the set of vector fields on Q, and by $C^\infty(Q)$ the set of smooth functions on Q. Throughout the paper, the manifold Q and the mathematical objects defined on it will be assumed analytic.

A *simple mechanical control system* is defined by a triple $(Q, g, \mathcal{F})$, where Q is the manifold of configurations of the system, g is a Riemannian metric on Q, and $\mathcal{F} = \{F^1, \ldots, F^m\}$ is a set of m linearly independent 1-forms on Q, which physically correspond to forces or torques.

Associated with the metric g is a natural affine connection, called the *Levi–Civita* connection. An *affine connection* [1, 18] is defined as an assignment

$$\begin{array}{rccc} \nabla : & \mathfrak{X}(Q) \times \mathfrak{X}(Q) & \longrightarrow & \mathfrak{X}(Q), \\ & (X, Y) & \longmapsto & \nabla_X Y, \end{array}$$

which is $\mathbb{R}$-bilinear and satisfies $\nabla_{fX}Y = f\nabla_X Y$ and $\nabla_X(fY) = f\nabla_X Y + X(f)Y$, for any $X,\ Y \in \mathfrak{X}(Q)$, $f \in C^\infty(Q)$. A curve $c : [a,b] \longrightarrow Q$ is a *geodesic* for ∇ if $\nabla_{\dot c(t)}\dot c(t) = 0$. Locally, the condition for a curve $t \longmapsto (q^1(t), \dots, q^n(t))$ to be a geodesic can be expressed as

$$\ddot q^a + \Gamma^a_{bc}\dot q^b \dot q^c = 0\,, \qquad 1 \le a \le n\,, \tag{2.1}$$

where the $\Gamma^a_{bc}(q)$ are the Christoffel symbols of the affine connection, that is, they are given by $\nabla_{\frac{\partial}{\partial q^b}}\frac{\partial}{\partial q^c} = \Gamma^a_{bc}\frac{\partial}{\partial q^a}$. The geodesic equation (2.1) is a first-order differential equation on TQ. The vector field corresponding to this first-order equation is given in coordinates by

$$S = v^a\frac{\partial}{\partial q^a} - \Gamma^a_{bc}v^b v^c\frac{\partial}{\partial v^a}$$

and is called the *geodesic spray* of the affine connection ∇. Hence, the integral curves of the geodesic spray S, $(q^a, \dot q^a)$ are the solutions of the geodesic equation.

The Levi–Civita connection ∇^g is determined by the formula

$$\begin{aligned} 2\,g(\nabla^g_X Y, Z) = {} & (X(g(Y,Z)) + Y(g(Z,X)) - Z(g(X,Y)) \\ & + g(Y,[Z,X]) - g(X,[Y,Z]) + g(Z,[X,Y]))\,, \qquad X,Y,Z \in \mathfrak{X}(Q)\,. \end{aligned}$$

One can compute the Christoffel symbols of ∇^g to be

$$\Gamma^a_{bc} = \frac{1}{2}g^{ad}\left(\frac{\partial g_{db}}{\partial q^c} + \frac{\partial g_{dc}}{\partial q^b} - \frac{\partial g_{bc}}{\partial q^d}\right),$$

where (g^{ad}) denotes the inverse of the inertia matrix $(g_{da}) = (g(\frac{\partial}{\partial q^d}, \frac{\partial}{\partial q^a}))$.

The metric tensor g induces a bundle isomorphism $\flat_g : TQ \longrightarrow T^*Q$ given by $\flat_g(X)(Y) = g(X,Y)$. Instead of the input forces $F^1, \dots, F^m$, we shall make use of the vector fields $Y_1, \dots, Y_m$, defined as $Y_i = \flat_g^{-1}(F^i)$. Roughly speaking, this corresponds to considering "accelerations" rather than forces. If $Y_i = Y_i^a(q)\frac{\partial}{\partial q^a}$, the control equations for the simple mechanical control system read in coordinates as

$$\begin{aligned} \dot q^a &= v^a\,, \\ \dot v^a &= -\Gamma^a_{bc}\dot q^b \dot q^c + \sum_{i=1}^m u_i(t)Y_i^a(q)\,, \qquad 1 \le a \le n\,. \end{aligned}$$

These equations can be written in a coordinate-free way as

$$\nabla^g_{\dot c(t)}\dot c(t) = \sum_{i=1}^m u^i(t)Y_i(c(t))\,. \tag{2.2}$$

The inputs we will consider come from the set $\mathcal{U} = \{u : [0,T] \to \mathbb{R}^m \mid T > 0,\ u \text{ is measurable and } \|u\| \le 1\}$, where

$$\|u\| = \sup_{t\in[0,T]}\|u(t)\|_\infty = \sup_{t\in[0,T]}\max_{l=1,\dots,m}|u_l(t)|\,.$$

We can use a general affine connection in (2.2) instead of the Levi–Civita connection without changing the structure of the equation. This is particularly interesting, since nonholonomic mechanical control systems also give rise to equations of

the form (2.2) by means of the so-called nonholonomic affine connection (see [23]). Therefore, the discussion throughout the paper is carried out for a general affine connection ∇.

We can turn (2.2) into a general affine control system with drift

$$\dot{x}(t) = f(x(t)) + \sum u^i(t) g_i(x(t)) \,. \tag{2.3}$$

To do this we need another bit of notation. The vertical lift of a vector field X on Q is the vector field X^v on TQ defined as

$$X^v(v_q) = \frac{d}{dt}\Big|_{t=0} (v_q + tX(q)) \,.$$

In coordinates, if $X = X^a \frac{\partial}{\partial q^a}$, one can check that $X^v = X^a \frac{\partial}{\partial v^a}$. Then, the second-order equation (2.2) on Q can be written as the first-order system on TQ:

$$\dot{v} = S(v) + \sum_{i=1}^{m} u^i(t) Y_i^v(v) \,, \tag{2.4}$$

where S is the geodesic spray associated with the affine connection ∇.

2.1. Controllability notions. The control equations for the mechanical system (2.4) are nonlinear. The standard techniques in control theory [30], like, for example, the linearization around an equilibrium point or linearization by feedback, do not yield satisfactory results in the analysis of its controllability properties, in the sense that they do not provide necessary and sufficient conditions characterizing them.

The point in the approach of Lewis and Murray to simple mechanical control systems is precisely to focus on what is happening to configurations, rather than to states, since in many of these systems, configurations may be controlled, but not configurations and velocities at the same time. The basic question they pose is, What is the set of configurations which are attainable from a given configuration starting from rest? Moreover, since we deal with objects defined on the configuration manifold Q, we expect to find answers on Q, although the control system (2.4) lives in TQ.

DEFINITION 2.1. *A* solution *of* (2.2) *is a pair* (c, u)*, where* $c : [0, T] \longrightarrow Q$ *is a piecewise smooth curve and* $u \in \mathcal{U}$ *such that* $(\dot{c}, u)$ *satisfies the first-order control system* (2.4)*.*

Consider $q_0 \in Q$, $(q_0, 0_{q_0}) \in T_{q_0}Q$ and let $U \subset Q$, $\overline{U} \subset TQ$ be neighborhoods of q_0 and $(q_0, 0_{q_0})$, respectively. Define

$$\mathcal{R}_Q^U(q_0, T) = \left\{ q \in Q \,\middle|\, \begin{array}{c} \text{there exists a solution } (c,u) \text{ of (2.2) such that} \\ \dot{c}(0) = 0_{q_0},\ c(t) \in U \text{ for } t \in [0,T], \text{ and } \dot{c}(T) \in T_qQ \end{array} \right\},$$

$$\mathcal{R}_{TQ}^{\overline{U}}(q_0, T) = \left\{ (q, v) \in TQ \,\middle|\, \begin{array}{l} \text{there exists a solution } (c,u) \text{ of (2.2) such that } \dot{c}(0) = \\ 0_{q_0},\ (c(t), \dot{c}(t)) \in \overline{U} \text{ for } t \in [0,T], \text{ and } \dot{c}(T) = v \in T_qQ \end{array} \right\}$$

and denote

$$\mathcal{R}_Q^U(q_0, \leq T) = \cup_{0 \leq t \leq T} \mathcal{R}_Q^U(q_0, t) \,, \qquad \mathcal{R}_{TQ}^{\overline{U}}(q_0, \leq T) = \cup_{0 \leq t \leq T} \mathcal{R}_{TQ}^{\overline{U}}(q_0, t) \,.$$

Now, we recall the notions of accessibility considered in [24].

DEFINITION 2.2. *The system* (2.2) *is* locally configuration accessible (LCA) at $q_0 \in Q$ *if there exists* $T > 0$ *such that* $\mathcal{R}^U_Q(q_0, \leq t)$ *contains a nonempty open set of* Q, *for all neighborhoods* U *of* q_0 *and all* $0 \leq t \leq T$. *If this holds for any* $q_0 \in Q$, *then the system is called LCA.*

DEFINITION 2.3. *The system* (2.2) *is* locally accessible (LA) at $q_0 \in Q$ and zero velocity *if there exists* $T > 0$ *such that* $\mathcal{R}^{\overline{U}}_{TQ}(q_0, \leq t)$ *contains a nonempty open set of* TQ, *for all neighborhoods* $\overline{U}$ *of* $(q_0, 0_{q_0})$ *and all* $0 \leq t \leq T$. *If this holds for any* $q_0 \in Q$, *then the system is called LA at zero velocity.*

We shall focus our attention on the following concepts of controllability [24].

DEFINITION 2.4. *The system* (2.2) *is* small-time locally configuration controllable (STLCC) at $q_0 \in Q$ *if there exists* $T > 0$ *such that* $\mathcal{R}^U_Q(q_0, \leq t)$ *contains a nonempty open set of* Q *to which* q_0 *belongs, for all neighborhoods* U *of* q_0 *and all* $0 \leq t \leq T$. *If this holds for any* $q_0 \in Q$, *then the system is called STLCC.*

DEFINITION 2.5. *The system* (2.2) *is* small-time locally controllable (STLC) at $q_0 \in Q$ and zero velocity *if there exists* $T > 0$ *such that* $\mathcal{R}^{\overline{U}}_{TQ}(q_0, \leq t)$ *contains a nonempty open set of* TQ *to which* $(q_0, 0_{q_0})$ *belongs, for all neighborhoods* $\overline{U}$ *of* $(q_0, 0_{q_0})$ *and all* $0 \leq t \leq T$. *If this holds for any* $q_0 \in Q$, *then the system is called STLC at zero velocity.*

3. Existing results. Here we review some accessibility and controllability results obtained in [24, 25] and summarize the work by Bullo [3] in describing the evolution of mechanical control systems via a series expansion.

3.1. On controllability. Given an affine connection ∇ on Q, the *symmetric product* of two vector fields $X, Y \in \mathfrak{X}(Q)$ is defined by

$$\langle X : Y \rangle = \nabla_X Y + \nabla_Y X \,.$$

The geometric meaning of the symmetric product is the following [22]: a *geodesically invariant* distribution $\mathcal{D}$ is a distribution such that for every geodesic $c(t)$ of ∇ starting from a point in $\mathcal{D}$, $\dot{c}(0) \in \mathcal{D}_{c(0)}$, we have that $\dot{c}(t) \in \mathcal{D}_{c(t)}$. Then, one can prove that $\mathcal{D}$ is geodesically invariant if and only if $\langle X : Y \rangle \in \mathcal{D}$, for all X, $Y \in \mathcal{D}$.

Given the input vector fields $\mathcal{Y} = \{Y_1, \ldots, Y_m\}$, let us denote by $\overline{Sym}(\mathcal{Y})$ the distribution obtained by closing the set $\mathcal{Y}$ under the symmetric product and by $\overline{Lie}(\mathcal{Y})$ the involutive closure of $\mathcal{Y}$. With these ingredients, one can prove the following theorem.

THEOREM 3.1 (see [24]). *The control system* (2.2) *is LCA at* q *(respectively, LA at* q *and zero velocity) if* $\overline{Lie}(\overline{Sym}(\mathcal{Y}))_q = T_qQ$ *(respectively,* $\overline{Sym}(\mathcal{Y})_q = T_qQ$*).*

If P is a symmetric product of vector fields in $\mathcal{Y}$, we let $\gamma_i(P)$ denote the number of occurrences of Y_i in P. The *degree* of P will be $\gamma_1(P) + \cdots + \gamma_m(P)$. We shall say that P is *bad* if $\gamma_i(P)$ is even for each $1 \leq i \leq m$. We say that P is *good* if it is not bad. The following theorem gives sufficient conditions for STLCC.

THEOREM 3.2. *Suppose that the system* (2.2) *is LCA at* q *(respectively, LA at* q *and zero velocity) and that* $\mathcal{Y}$ *is such that every bad symmetric product* P *at* q *in* $\mathcal{Y}$ *can be written as a linear combination of good symmetric products at* q *of lower degree than* P. *Then* (2.2) *is STLCC at* q *(respectively, STLC at* q *and zero velocity).*

This theorem was proved in [24], adapting previous work by Sussmann [35] on general control systems of the form (2.3). Throughout the paper, we will refer to the conditions of every bad symmetric product at q being a linear combination of good symmetric products at q of lower degree as the *sufficient conditions for STLCC.*

3.2. Series expansion. Within the realm of geometric control theory, series expansions play a key role in the study of nonlinear controllability [2, 15, 34, 35], trajectory generation and motion planning problems [4, 19, 20, 29], etc. In [27], Magnus describes the evolution of systems on a Lie group. In [7, 10, 16, 36] a general framework is developed to describe the evolution of a nonlinear system via the so-called Chen–Fliess series and its factorization.

In the context of mechanical control systems, the work by Bullo in [3] describes the evolution of the trajectories with zero initial velocity via a series expansion on the configuration manifold Q. In this section we describe the series expansion, which will be key in the subsequent discussion. Before doing so, however, we need to introduce some notation on analyticity over complex neighborhoods.

Let $q_0 \in Q$. By selecting a coordinate chart around q_0, we locally identify $Q \equiv \mathbb{R}^n$. In this way, we write $q_0 \in \mathbb{R}^n$. Let σ be a positive scalar, and define the complex σ-neighborhood of q_0 in $\mathbb{C}^n$ as $B_\sigma(q_0) = \{z \in \mathbb{C}^n \mid \|z - q_0\| < \sigma\}$. Let f be a real analytic function on $\mathbb{R}^n$ that admits a bounded analytic continuation over $B_\sigma(q_0)$. The norm of f is defined as

$$\|f\|_\sigma \triangleq \max_{z \in B_\sigma(q_0)} |f(z)| \,,$$

where f denotes both the function over $\mathbb{R}^n$ and its analytic continuation. Given a time-varying vector field $(q,t) \mapsto Z(q,t) = Z_t(q)$, let Z_t^i be its ith component with respect to the usual basis on $\mathbb{R}^n$. Assuming $t \in [0,T]$, and assuming that every component function Z_t^i is analytic over $B_\sigma(q_0)$, we define the norm of Z as

$$\|Z\|_{\sigma,T} \triangleq \max_{t\in[0,T]} \max_{i\in\{1,\dots,n\}} \|Z_t^i\|_\sigma \,.$$

In what follows, we will often simplify notation by neglecting the subscript T in the norm of a time-varying vector field. Finally, given an affine connection ∇ with Christoffel symbols $\{\Gamma^i_{jk} \mid i,j,k \in \{1,\dots,n\}\}$, we introduce the following notation:

$$\|\Gamma\|_\sigma \triangleq \max_{i,j,k} \|\Gamma^i_{jk}\|_\sigma \,.$$

In what follows, we let

$$Z(q,t) = \sum_{i=1}^m u_i(t) Y_i(q) \,.$$

THEOREM 3.3 (see [3]). *Let $c(t)$ be the solution of (2.2) with input given by $Z(q,t)$ and with initial conditions $c(0) = q_0$, $\dot c(0) = 0$. Let the Christoffel symbols $\Gamma^i_{jk}(q)$ and the vector field $Z(q,t)$ be uniformly integrable and bounded analytic in Q. Define recursively the time-varying vector fields*

$$\begin{aligned} V_1(q,t) &= \int_0^t Z(q,s) ds \,, \\ V_k(q,t) &= -\frac{1}{2} \sum_{j=1}^{k-1} \int_0^t \langle V_j(q,s) : V_{k-j}(q,s) \rangle ds \,, \qquad k \geq 2 \,, \end{aligned}$$

where q is maintained fixed at each integral. Select a coordinate chart around the point $q_0 \in Q$, let $\sigma > \sigma'$ be two positive constants, and assume that

$$\|Z\|_\sigma T^2 < L \triangleq \min \left\{ \frac{\sigma - \sigma'}{2^4 n^2 (n+1)}, \frac{1}{2^4 n (n+1) \|\Gamma\|_\sigma}, \frac{\eta^2 (\sigma' n^2 \|\Gamma\|_{\sigma'})}{n^2 \|\Gamma\|_{\sigma'}} \right\} . \tag{3.1}$$

Then the series $(q,t) \longmapsto \sum_{k=1}^{\infty} V_k(q,t)$ *converges absolutely and uniformly in* t *and* q, *for all* $t \in [0,T]$ *and for all* $q \in B_{\sigma'}(q_0)$, *with the* V_k *satisfying the bound*

$$\|V_k\|_{\sigma'} \leq L^{1-k}\, \|Z\|_{\sigma}^{k}\, t^{2k-1}\,, \tag{3.2}$$

Over the same interval, the solution $c(t)$ *satisfies*

$$\dot{c}(t) = \sum_{k=1}^{\infty} V_k(c(t),t)\,. \tag{3.3}$$

This theorem generalizes previous results obtained in [4] under the assumption of small amplitude forcing. The first few terms of the series (3.3) can be computed to obtain

(3.4)

$$\begin{aligned}\dot{c}(t) = {} & \overline{Z}(c(t),t) - \frac{1}{2}\overline{\langle \overline{Z} : \overline{Z} \rangle}(c(t),t) + \frac{1}{2}\overline{\left\langle \langle \overline{Z} : \overline{Z} \rangle : \overline{Z} \right\rangle}(c(t),t) \\ & - \frac{1}{2}\overline{\left\langle \left\langle \langle \overline{Z} : \overline{Z} \rangle : \overline{Z} \right\rangle : \overline{Z} \right\rangle}(c(t),t) - \frac{1}{8}\overline{\left\langle \langle \overline{Z} : \overline{Z} \rangle : \langle \overline{Z} : \overline{Z} \rangle \right\rangle}(c(t),t) + O(\|Z\|_{\sigma}^{5} t^{9})\,,\end{aligned}$$

where $\overline{Z}(q,t) \equiv \int_0^t Z(q,s)ds$ and so on.

4. The single-input case. Theorem 3.2 gives us sufficient conditions for STLCC. A natural concern both from the theoretical and the practical points of view is to try to sharpen this controllability test. Lewis [21] investigated the single-input case and proved the next result.

THEOREM 4.1. *Let* (Q,g) *be an analytic manifold with an affine connection* ∇. *Let* Y *be an analytic vector field on* Q *and* $q_0 \in Q$. *Then the system*

$$\nabla_{\dot{c}(t)}\dot{c}(t) = u(t)Y(c(t))$$

is locally configuration controllable at $q_0 \in Q$ *if and only if* $\dim Q = 1$.

The fact of being able to completely characterize STLCC in the single-input case (something which has not been accomplished yet for general control systems of the form (2.3)) suggests that understanding local configuration controllability for mechanical systems may be possible. More precisely, examining the single-input case, one can deduce that if (2.2) is STLCC at q_0, then $\dim Q = 1$, which implies $\langle Y : Y \rangle(q_0) \in \operatorname{span}\{Y(q_0)\}$, i.e., sufficient conditions for STLCC are also necessary. Can this be extrapolated to the multi-input case? The following conjecture was posed by Lewis:

> *Let a mechanical control system* (2.2) *be LCA at* $q_0 \in Q$. *Then it is STLCC at* q_0 *if and only if there exists a basis of input vector fields which satisfies the sufficient conditions for STLCC at* q_0.

Theorem 4.1 implies that the conjecture is true for $m = 1$. In the following section we prove that this conjecture is also valid for $m = n - 1$.

5. Mechanical systems underactuated by one control. Here we focus our attention on mechanical control systems of the form (2.2) which have n degrees of freedom and $m = n - 1$ control input vector fields. The following lemma, taken from [34], will be helpful in the proof of the theorem of this section.

LEMMA 5.1. *Let* Q *be a* n*-dimensional analytic manifold. Given* $q_0 \in Q$ *and* $X_1, \ldots, X_p \in \mathfrak{X}(Q)$, $p \leq n$, *linearly independent vector fields, there exists a function* $\phi : Q \longrightarrow \mathbb{R}$ *satisfying the properties*

1. ϕ *is analytic,*
2. $\phi(q_0) = 0$*,*
3. $X_1(\phi) = \cdots = X_{p-1}(\phi) = 0$ *on a neighborhood* V *of* q_0*,*
4. $X_p(\phi)(q_0) = -1$*,*
5. *within any neighborhood of* q_0 *there exist points* q*, where* $\phi(q) < 0$ *and* $\phi(q) > 0$*.*

Proof. Let $Z_1, \ldots, Z_n$ be vector fields defined in a neighborhood of q_0 such that $\{Z_1(q_0), \ldots, Z_n(q_0)\}$ forms a basis for $T_{q_0}Q$ and $Z_i = X_i$, $1 \leq i \leq p-1$, $Z_p = -X_p$. Let $t_i \longmapsto \Psi_i(t)$ be the flow of Z_i, $1 \leq i \leq n$. In a sufficiently small neighborhood V of q_0, any point q may be expressed as $q = \Psi_1(t_1) \circ \cdots \circ \Psi_n(t_n)(q_0)$ for some unique n-tuple $(t_1, \ldots, t_n) \in \mathbb{R}^n$. Define $\phi(q) = t_p$. It is a simple exercise to verify that ϕ satisfies the required properties. □

Next, we state and prove the main result of the paper.

THEOREM 5.2. *Let* Q *be a* n*-dimensional analytic manifold and let* $Y_1, \ldots, Y_{n-1}$ *be analytic vector fields on* Q*. Consider the control system*

$$\nabla_{\dot{c}(t)}\dot{c}(t) = \sum_{i=1}^{n-1} u_i(t) Y_i(c(t)), \tag{5.1}$$

and assume that it is LCA at $q_0 \in Q$*. Then the system is locally configuration controllable at* q_0 *if and only if there exists a basis of input vector fields satisfying the sufficient conditions for STLCC at* q_0*.*

A rough sketch of the proof is the following: because of the hypotheses of the theorem, we need only to check that the symmetric products of degree two of a given basis of the input distribution, when evaluated at q_0, are linear combinations of good products of degree one. To verify this, we associate with the given basis a symmetric matrix A, in such a way that this basis satisfies the sufficient conditions for STLCC if and only if the diagonal elements of A are all zero. If this is not the case, we search for a change of basis B such that the new basis has an associated matrix A with zeros in its diagonal. This is equivalent to solving a quadratic equation in B. In order to ensure that a solution to this equation exists, we have to explore the different possibilities that may occur regarding the various radicands involved. Finally, we discard the situations in which the equation is not solvable by a contradiction argument with the controllability assumption (see Figure 5.1).

Proof. We need only to prove one implication (the other one is Theorem 3.2). Let us suppose that the system is locally configuration controllable at q_0. Let $\mathcal{D}$ denote the input distribution. One of the following is true:

1. For all $Y_1, Y_2 \in \mathcal{D}$, $\langle Y_1 : Y_2 \rangle(q_0) \in \mathcal{D}_{q_0}$.
2. There exist Y_1, $Y_2 \in \mathcal{D}$ such that $\langle Y_1 : Y_2 \rangle(q_0) \notin \mathcal{D}_{q_0}$.

In case 1, there is nothing to prove since any basis of input vector fields satisfies the sufficient conditions for STLCC at q_0. In case 2, it is clear that one can choose Y_1, $Y_2 \in \mathcal{D}$, linearly independent at q_0 and such that $\langle Y_1 : Y_2 \rangle(q_0) \notin \mathcal{D}_{q_0}$. (If Y_1, Y_2 in case 2 are linearly dependent, then $\langle Y_1 : Y_1 \rangle(q_0) \notin \mathcal{D}_{q_0}$. Take any Y_2 linearly independent with Y_1. If $\langle Y_1 : Y_2 \rangle(q_0) \in \mathcal{D}_{q_0}$, define a new Y_2' by $Y_1 + Y_2$.) Therefore, we can complete the set $\{Y_1(q_0), Y_2(q_0)\}$ to a basis of $\mathcal{D}_{q_0}$,

$$\{Y_1(q_0), Y_2(q_0), \ldots, Y_m(q_0)\}$$

such that $\operatorname{span}\{Y_1(q_0), Y_2(q_0), \ldots, Y_m(q_0), \langle Y_1 : Y_2 \rangle(q_0)\} = T_{q_0}Q$. In this basis, the

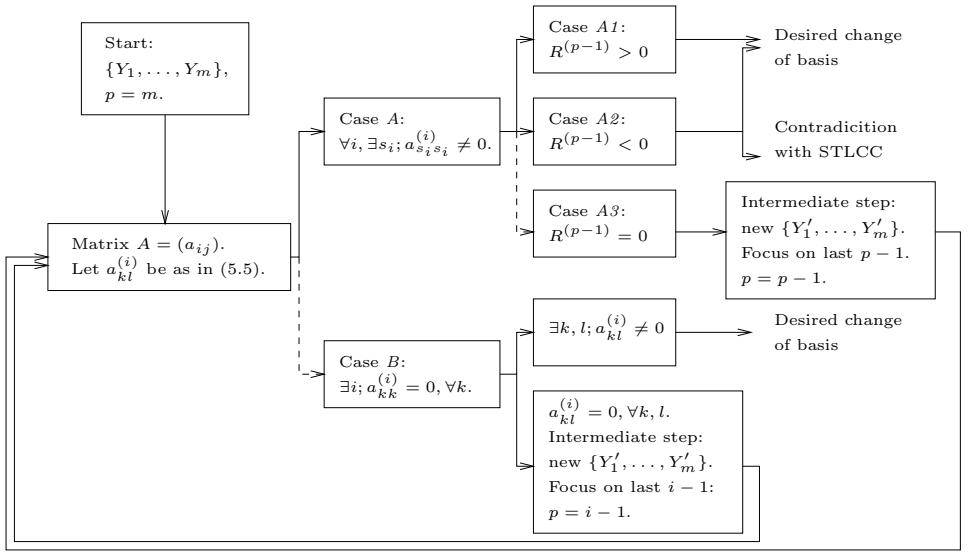


FIG. 5.1. *Illustration of the proof of Theorem* 5.2. $R^{(p-1)}$ *denotes* $(a^{(p-1)}_{s_{p-1}s_p})^2 - a^{(p-1)}_{s_{p-1}s_{p-1}} a^{(p-1)}_{s_p s_p}$. *The dashed lines mean that one cannot fall repeatedly in Cases* A3 *or* B *without contradicting STLCC.*

symmetric products of degree two of the vector fields $\{Y_1, \ldots, Y_m\}$ at q_0 are expressed,

$$\begin{aligned}
\langle Y_1 : Y_1 \rangle(q_0) &= lc(Y_1(q_0), \ldots, Y_m(q_0)) + a_{11} \langle Y_1 : Y_2 \rangle(q_0)\,, \\
&\vdots \\
\langle Y_m : Y_m \rangle(q_0) &= lc(Y_1(q_0), \ldots, Y_m(q_0)) + a_{mm} \langle Y_1 : Y_2 \rangle(q_0)\,, \\
\langle Y_1 : Y_2 \rangle(q_0) &= a_{12} \langle Y_1 : Y_2 \rangle(q_0)\,, \\
\langle Y_1 : Y_3 \rangle(q_0) &= lc(Y_1(q_0), \ldots, Y_m(q_0)) + a_{13} \langle Y_1 : Y_2 \rangle(q_0)\,, \\
&\vdots \\
\langle Y_{m-1} : Y_m \rangle(q_0) &= lc(Y_1(q_0), \ldots, Y_m(q_0)) + a_{m-1\,m} \langle Y_1 : Y_2 \rangle(q_0)\,,
\end{aligned}$$

where $lc(Y_1(q_0), \ldots, Y_m(q_0))$ means a linear combination of $Y_1(q_0), \ldots, Y_m(q_0)$. The coefficients a_{ij} define a symmetric matrix $A = (a_{ij}) \in \mathbb{R}^{m \times m}$. Observe that if $a_{11} = \cdots = a_{mm} = 0$, then the bad symmetric products $\langle Y_i : Y_i \rangle(q_0)$ are in $\mathcal{D}_{q_0}$ and we have finished. Suppose then that the opposite situation is true, that is, there exists $s = s_1$ such that $a_{s_1 s_1} \neq 0$.

What we are going to prove now is that, under the hypothesis of STLCC at q_0, there exists a change of basis $B = (b_{jk})$, $\det B \neq 0$, providing new vector fields in $\mathcal{D}$,

$$Y'_j = \sum_{k=1}^{m} b_{jk} Y_k\,, \qquad 1 \leq j \leq m\,,$$

which satisfy the sufficient conditions for STLCC at q_0. Since

$$\begin{aligned}\langle Y_j' : Y_j' \rangle(q_0) &= \sum_{k,l=1}^{m} b_{jk} b_{jl} \langle Y_k : Y_l \rangle(q_0) \\ &= \sum_{k=1}^{m} b_{jk}^2 \langle Y_k : Y_k \rangle(q_0) + 2 \sum_{1 \le k < l \le m} b_{jk} b_{jl} \langle Y_k : Y_l \rangle(q_0) \\ &= lc(Y_1'(q_0), \dots, Y_m'(q_0)) + \left(\sum_{k=1}^{m} b_{jk}^2 a_{kk} + 2 \sum_{1 \le k < l \le m} b_{jk} b_{jl} a_{kl} \right) \langle Y_1 : Y_2 \rangle(q_0) \,,\end{aligned} \tag{5.2}$$

the matrix B we are looking for must fulfill

$$\sum_{k=1}^{m} b_{jk}^2 a_{kk} + 2 \sum_{1 \le k < l \le m} b_{jk} b_{jl} a_{kl} = 0 \,, \qquad 1 \le j \le m \,, \tag{5.3}$$

or, equivalently,

$$(BAB^T)_{jj} = 0 \,, \qquad 1 \le j \le m \,.$$

Note that, since $a_{s_1 s_1} \neq 0$, this is equivalent to

$$\begin{aligned} b_{j s_1} = & \frac{-\sum_{k \neq s_1} b_{jk} a_{k s_1}}{a_{s_1 s_1}} \\ & \pm \frac{\sqrt{(\sum_{k \neq s_1} b_{jk} a_{k s_1})^2 - a_{s_1 s_1} (\sum_{k \neq s_1} b_{jk}^2 a_{kk} + 2 \sum_{k<l,\, k,l \neq s_1} b_{jk} b_{jl} a_{kl})}}{a_{s_1 s_1}} \,, \end{aligned}$$

for each $1 \le j \le m$. After some computations, the radicand of this expression becomes

$$\sum_{k \neq s_1} b_{jk}^2 (a_{k s_1}^2 - a_{s_1 s_1} a_{kk}) + 2 \sum_{k<l,\, k,l \neq s_1} b_{jk} b_{jl} (a_{k s_1} a_{l s_1} - a_{s_1 s_1} a_{kl}) \,.$$

If this radicand is zero, it would imply that the matrix B should be singular in order to satisfy (5.3). We must ensure then that it is possible to select B such that the radicand is different from zero. We do this in the following, studying several cases that can occur. Letting

$$a_{kl}^{(2)} = a_{k s_1} a_{l s_1} - a_{s_1 s_1} a_{kl} \,, \qquad k,l \in \{1, \dots, m\} \setminus \{s_1\} \,,$$

we have that the radicand would vanish if

$$\sum_{k \neq s_1} b_{jk}^2 a_{kk}^{(2)} + 2 \sum_{k<l,\, k,l \neq s_1} b_{jk} b_{jl} a_{kl}^{(2)} = 0 \,. \tag{5.4}$$

Note the similarity between (5.3) and (5.4). Define recursively

$$\begin{aligned} & a_{kl}^{(1)} = a_{kl} \,, \\ & a_{kl}^{(i)} = a_{k s_{i-1}}^{(i-1)} a_{l s_{i-1}}^{(i-1)} - a_{s_{i-1} s_{i-1}}^{(i-1)} a_{kl}^{(i-1)} \,, \quad i \ge 2 \,, \quad k,l \in \{1, \dots, m\} \setminus \{s_1, \dots, s_{i-1}\} \,. \end{aligned} \tag{5.5}$$

Case A. Here we treat the case when for each i there exists s_i such that $a^{(i)}_{s_i s_i} \neq 0$. Several subcases are discussed.

Reasoning as before, (5.4) would imply that for $1 \leq j \leq m$

$$b_{js_2} = lc(b_{j1}, \ldots, \hat{b}_{js_1}, \ldots, \hat{b}_{js_2}, \ldots, b_{jm}) \pm \frac{1}{a^{(2)}_{s_2 s_2}} \sqrt{\sum_{k \neq s_1, s_2} b^2_{jk} a^{(3)}_{kk} + 2 \sum_{k<l,\, k,l \neq s_1, s_2} b_{jk} b_{jl} a^{(3)}_{kl}},$$

where the symbol $\hat{b}$ means that the term b has been removed. Iterating this procedure, we finally obtain the following equations for the $b_{js_{m-1}}$:

$$b_{js_{m-1}} = b_{js_m} \frac{-a^{(m-1)}_{s_{m-1}s_m} \pm \sqrt{(a^{(m-1)}_{s_{m-1}s_m})^2 - a^{(m-1)}_{s_{m-1}s_{m-1}} a^{(m-1)}_{s_m s_m}}}{a^{(m-1)}_{s_{m-1}s_{m-1}}}, \quad 1 \leq j \leq m\,.$$

Let $(b_{js_m})_{1 \leq j \leq m}$ be a nonzero vector in $\mathbb{R}^m$. Now, we distinguish three possibilities.

Case A1. We show that if the radicand $(a^{(m-1)}_{s_{m-1}s_m})^2 - a^{(m-1)}_{s_{m-1}s_{m-1}} a^{(m-1)}_{s_m s_m}$ is positive, then it is possible to obtain the desired change of basis.

If $(a^{(m-1)}_{s_{m-1}s_m})^2 - a^{(m-1)}_{s_{m-1}s_{m-1}} a^{(m-1)}_{s_m s_m} > 0$, then the quadratic polynomial in $b_{js_{m-1}}$,

$$a^{(m-1)}_{s_{m-1}s_{m-1}} b^2_{js_{m-1}} + 2a^{(m-1)}_{s_{m-1}s_m} b_{js_{m-1}} b_{js_m} + a^{(m-1)}_{s_m s_m} b^2_{js_m}\,, \tag{5.6}$$

has two real roots and we can choose $(b_{js_{m-1}})_{1 \leq j \leq m} \in \mathbb{R}^m$, linearly independent with $(b_{js_m})_{1 \leq j \leq m}$ such that (5.6) is positive for all $1 \leq j \leq m$. As this polynomial is the radicand of the preceding one,

$$\sum_{k \neq s_1, \ldots, s_{m-3}} b^2_{jk} a^{(m-2)}_{kk} + 2 \sum_{k<l,\, k,l \neq s_1, \ldots, s_{m-3}} b_{jk} b_{jl} a^{(m-2)}_{kl}\,, \tag{5.7}$$

our choice of $(b_{js_{m-1}})_{1 \leq j \leq m}$ ensures that we can again take $(b_{js_{m-2}})_{1 \leq j \leq m} \in \mathbb{R}^m$, linearly independent with $(b_{js_{m-1}})_{1 \leq j \leq m}$ and $(b_{js_m})_{1 \leq j \leq m}$ such that (5.7) is positive for all $1 \leq j \leq m$. This is propagated step by step through the iteration process and we are able to choose a nonsingular matrix (b_{jk}) satisfying (5.3).

Case A2. We show that when the radicand $(a^{(m-1)}_{s_{m-1}s_m})^2 - a^{(m-1)}_{s_{m-1}s_{m-1}} a^{(m-1)}_{s_m s_m}$ is negative, then either it is possible to find the change of basis or the system is not STLCC at q_0.

If $(a^{(m-1)}_{s_{m-1}s_m})^2 - a^{(m-1)}_{s_{m-1}s_{m-1}} a^{(m-1)}_{s_m s_m} < 0$, then for all $b_{js_{m-1}}$, b_{js_m} (5.6) does not change its sign. If this sign is positive, the same argument as in Case A1 ensures us the choice of the desired matrix. If negative, it implies that for all $b_{js_{m-2}}$, $b_{js_{m-1}}$, b_{js_m} (5.7) does not change its sign. Then, the unique problem we must face is when, through the iteration process, all the radicands are negative. In the following, we discard this latter case by contradiction with the hypothesis of controllability. Apply Lemma 5.1 to the vector fields $\{Y_1, \ldots, Y_m, \langle Y_1 : Y_2 \rangle\}$ to find a function ϕ satisfying properties 1–5. By (3.4), we have that

$$\begin{aligned}\dot{c}(t) &= \sum_{i=1}^m \bar{u}_i Y_i - \frac{1}{2} \overline{\left\langle \sum_{j=1}^m \bar{u}_j Y_j : \sum_{k=1}^m \bar{u}_k Y_k \right\rangle} + O(\|Z\|^3_\sigma t^5) \\ &= \sum_{i=1}^m \bar{u}_i Y_i - \frac{1}{2} \left(\overline{\sum_{j=1}^m \bar{u}_j^2 \langle Y_j : Y_j \rangle - \sum_{j<k} \bar{u}_j \bar{u}_k \langle Y_j : Y_k \rangle} \right) + O(\|Z\|^3_\sigma t^5)\,,\end{aligned}$$

where $Z = \sum_{i=1}^m u_i Y_i$. Now, observe that $\frac{d}{dt}(\phi(c(t))) = \dot{c}(t)(\phi)$. Then, using properties 3 and 4 of ϕ, we get

$$\frac{d}{dt}(\phi(c(t))) = \frac{1}{2}\left(\overline{\sum_{j=1}^m a_{jj}\bar{u}_j^2 + 2\sum_{j<k} a_{jk}\bar{u}_j\bar{u}_k}\right) + O(\|Z\|_\sigma^3 t^5)\,.$$

The expression $\sum_{j=1}^m a_{jj}\bar{u}_j^2 + 2\sum_{j<k} a_{jk}\bar{u}_j\bar{u}_k$ does not change its sign, whatever the functions $u_1(t), \ldots, u_m(t)$ might be, because as a quadratic polynomial in $\bar{u}_{s_1}$ its radicand is always negative. Therefore, $\frac{d}{dt}(\phi(c(t)))$ has constant sign for sufficiently small t, since $\overline{\sum_{j=1}^m a_{jj}\bar{u}_j^2 + 2\sum_{j<k} a_{jk}\bar{u}_j\bar{u}_k}$ is $O(\|u\|^2 t^3)$ and dominates $O(\|Z\|_\sigma^3 t^5) = O(\|u\|^3 t^5)$ when $t \to 0$. Finally,

$$\phi(c(t)) = \phi(q_0) + \int_0^t \frac{d}{ds}(\phi(c(s))) = \int_0^t \frac{d}{ds}(\phi(c(s)))$$

will have constant sign for t small enough. As a consequence, all the points in a neighborhood of q_0 where ϕ has the opposite sign (property 5) are unreachable in small time, which contradicts the hypothesis of controllability.

Case A3. We show that if the radicand $(a^{(m-1)}_{s_{m-1}s_m})^2 - a^{(m-1)}_{s_{m-1}s_{m-1}} a^{(m-1)}_{s_m s_m}$ vanishes, then an intermediate change of basis reduces the problem to considering $m-1$ input vector fields. The preceding discussion can be then reproduced.

The situation now is similar to that of Case A2. However, the argument employed above to discard the possibility of all the radicands being negative does not apply, since in this case there *do* exist controls such that $\sum_{j=1}^m a_{jj}\bar{u}_j^2 + 2\sum_{j<k} a_{jk}\bar{u}_j\bar{u}_k$ is zero, and hence we should really investigate the sign of $O(\|Z\|_\sigma^3 t^5)$ to reach a contradiction. Instead, what we are going to do is to get a new basis $\{Y'_j\}$ such that $\langle Y'_1 : Y'_j\rangle(q_0) \in \mathcal{D}_{q_0}$, $1 \le j \le m$, and thus remove one vector field (Y'_1) from the discussion. By repeating this procedure, we finally come to consider a limit case, which we will discard by contradiction with the controllability hypothesis.

For $j = 1$, we choose $b_{1s_m} \neq 0$ and

$$\begin{aligned}
b_{1s_{m-1}} &= -b_{1s_m}\frac{a^{(m-1)}_{s_{m-1}s_m}}{a^{(m-1)}_{s_{m-1}s_{m-1}}} = C_{s_{m-1}} b_{1s_m}\,,\\
b_{1s_{m-2}} &= -\frac{a^{(m-2)}_{s_{m-2}s_{m-1}} b_{1s_{m-1}} + a^{(m-2)}_{s_{m-2}s_m} b_{1s_m}}{a^{(m-2)}_{s_{m-2}s_{m-2}}} = C_{s_{m-2}} b_{1s_m}\,,\\
&\vdots\\
b_{1s_1} &= -\frac{\sum_{k\neq s_1} b_{1k} a_{ks_1}}{a_{s_1 s_1}} = C_{s_1} b_{1s_m}\,.
\end{aligned} \tag{5.8}$$

We denote $C_{s_m} = 1$. For $j > 1$, we select the $(b_{jk})_{1\le k\le m}$ such that the matrix B is nonsingular. Consequently, we change our original basis $\{Y_1, \ldots, Y_m\}$ to a new one $\{Y'_1, \ldots, Y'_m\}$. In this basis, following (5.2), one has

$$\begin{aligned}
&\langle Y'_1 : Y'_1\rangle(q_0) = lc(Y'_1(q_0), \ldots, Y'_m(q_0))\,,\\
&\langle Y'_j : Y'_j\rangle(q_0) = lc(Y'_1(q_0), \ldots, Y'_m(q_0)) + a'_{jj}\langle Y_1 : Y_2\rangle(q_0)\,, \qquad 2 \le j \le m\,.
\end{aligned}$$

In addition, one can check that for each $2 \le j \le m$,

$$\begin{aligned}\langle Y_1' : Y_j' \rangle(q_0) &= lc(Y_1'(q_0), \dots, Y_m'(q_0)) + \left(\sum_{k,l} a_{kl} b_{1k} b_{jl} \right) \langle Y_1 : Y_2 \rangle(q_0) \\ &= lc(Y_1'(q_0), \dots, Y_m'(q_0)) + b_{1 s_m} \left(\sum_l b_{jl} \left(\sum_k a_{kl} C_k \right) \right) \langle Y_1 : Y_2 \rangle(q_0) \,.\end{aligned}$$

Now, when the C_k are given by (5.8), we have

$$\sum_k a_{kl} C_k = 0 \,, \qquad 1 \le l \le m$$

(see Lemma A.1 in the appendix), and this guarantees that

$$\langle Y_1' : Y_j' \rangle(q_0) = lc(Y_1'(q_0), \dots, Y_m'(q_0)) \,, \qquad 2 \le j \le m \,.$$

If the $a'_{jj} = 0$, $2 \le j \le m$, we are done. Assume then that $a'_{33} \neq 0$, reordering the input vector fields if necessary. Assume further that $\langle Y_2' : Y_3' \rangle(q_0)$ is not a linear combination of $\{Y_1', \dots, Y_m'\}$ (otherwise, redefine a new Y_2'' as $Y_2' + Y_3'$). Then,

$$\begin{aligned}\langle Y_2' : Y_2' \rangle(q_0) &= lc(Y_1'(q_0), \dots, Y_m'(q_0)) + a'_{22} \langle Y_2' : Y_3' \rangle(q_0) \,, \\ &\vdots \\ \langle Y_m' : Y_m' \rangle(q_0) &= lc(Y_1'(q_0), \dots, Y_m'(q_0)) + a'_{mm} \langle Y_2' : Y_3' \rangle(q_0) \,, \\ \langle Y_2' : Y_3' \rangle(q_0) &= a'_{23} \langle Y_2' : Y_3' \rangle(q_0) \,, \\ \langle Y_2' : Y_4' \rangle(q_0) &= lc(Y_1'(q_0), \dots, Y_m'(q_0)) + a'_{24} \langle Y_2' : Y_3' \rangle(q_0) \,, \\ &\vdots \\ \langle Y_{m-1}' : Y_m' \rangle(q_0) &= lc(Y_1'(q_0), \dots, Y_m'(q_0)) + a'_{m-1m} \langle Y_2' : Y_3' \rangle(q_0) \,,\end{aligned}$$

where we have denoted with a slight abuse of notation by a'_{jk} the new coefficients corresponding to $\langle Y_2' : Y_3' \rangle$. Consequently, we can now reproduce the preceding discussion, but with the $m-1$ vector fields $\{Y_2', \dots, Y_m'\}$. That is, we look for one change of basis B' in the vector fields $\{Y_2', \dots, Y_m'\}$ such that the new ones $\{Y_2'', \dots, Y_m''\}$ together with Y_1' verify the sufficient conditions for STLCC at q_0. Accordingly, we must consider the vanishing of the new polynomials

$$\sum_{k=2}^{m} {b_{jk}^2}' a'_{kk} + 2 \sum_{2 \le k < l \le m} b'_{jk} b'_{jl} a'_{kl} = 0 \,, \quad 2 \le j \le m \,.$$

The cases in which the last radicand $(a_{s_{m-1} s_m}^{(m-1)\prime})^2 - a_{s_{m-1} s_{m-1}}^{(m-1)\prime} a_{s_m s_m}^{(m-1)\prime}$ does not vanish are treated as before (Cases A1 and A2). When it vanishes, we obtain a new basis $\{Y_1'' = Y_1', Y_2'', \dots, Y_m''\}$ such that

$$\begin{aligned}&\langle Y_1'' : Y_1'' \rangle(q_0) \,, \langle Y_2'' : Y_2'' \rangle(q_0) \in \mathcal{D}_{q_0} \,, \\ &\langle Y_j'' : Y_j'' \rangle(q_0) = lc(Y_1''(q_0), \dots, Y_m''(q_0)) + c'_{jj} \langle Y_2' : Y_3' \rangle(q_0) \,, \qquad 3 \le j \le m \,, \\ &\langle Y_1'' : Y_j'' \rangle \,, \langle Y_2'' : Y_{j+1}'' \rangle \in \mathcal{D}_{q_0} \,, \qquad 2 \le j \le m \,,\end{aligned}$$

where there could exist some $3 \le j \le m$ such that $c'_{jj} \neq 0$. By an induction procedure, we finally come to consider discarding the case of a certain basis $\{Z_1 = Y_1', Z_2 = Y_2'',$

$\dots, Z_m\}$ of $\mathcal{D}$ satisfying $\langle Z_i : Z_j\rangle(q_0) \in \operatorname{span}\{Z_1(q_0), \dots, Z_m(q_0)\}$, $1 \le i < j \le m$, and the sufficient conditions for STLCC at q_0 for $Z_1, \dots, Z_{m-1}$, but such that $\langle Z_m : Z_m\rangle(q_0) \notin \operatorname{span}\{Z_1(q_0), \dots, Z_m(q_0)\}$. Similarly as we have done above, the application of Lemma 5.1 with the vector fields $\{Z_1, \dots, Z_m, \langle Z_m : Z_m\rangle\}$ implies that the system is not controllable at q_0, yielding a contradiction.

Case B. Finally, we prove that if there exists an $i \ge 2$ such that $a^{(i)}_{kk} = 0$ for all $k \in \{1, \dots, m\} \setminus \{s_1, \dots, s_{i-1}\}$, then either the desired change of basis is straightforward or an intermediate step can be done that reduces the problem to considering $i-1$ input vector fields.

In this case, the polynomial

$$\sum_{k \neq s_1, \dots, s_{i-1}} b^2_{jk} a^{(i)}_{kk} + \quad 2 \sum_{k<l,\, k,l \neq s_1, \dots, s_{i-1}} b_{jk} b_{jl} a^{(i)}_{kl}$$

takes the form

$$2 \sum_{k<l,\, k,l \neq s_1, \dots, s_{i-1}} b_{jk} b_{jl} a^{(i)}_{kl}\,. \tag{5.9}$$

If any of the $a^{(i)}_{kl}$ are different from zero, then it is clear that we can choose the b_{jk}, $k \notin \{s_1, \dots, s_{i-1}\}$, such that (5.9) is positive. Then, reasoning as before, we find a regular matrix B yielding the desired change of basis. If this is not the case, i.e., $a^{(i)}_{kl} = 0$, for all $k < l$, $k, l \notin \{s_1, \dots, s_{i-1}\}$, we can do the following. Choose $\{(b_{jk})_{1 \le j \le m}\}$, with $k \notin \{s_1, \dots, s_{i-1}\}$, $m-i+1$ linearly independent vectors in $\mathbb{R}^m$, such that the minor $\{b_{jk}\}^{k \neq s_1, \dots, s_{i-1}}_{1 \le j \le m-i+1}$ is regular. Now, let j in (5.8) vary between 1 and $m-i+1$; that is, take

$$\begin{aligned} b_{js_{i-1}} &= -\frac{\sum^m_{k \neq s_1, \dots, s_{i-1}} b_{jk} a^{(i-1)}_{s_{i-1}k}}{a^{(i-1)}_{s_{i-1}s_{i-1}}}\,, \\ b_{js_{i-2}} &= -\frac{\sum^m_{k \neq s_1, \dots, s_{i-2}} b_{jk} a^{(i-2)}_{s_{i-2}k}}{a^{(i-2)}_{s_{i-2}s_{i-2}}}\,, \quad \dots, \quad b_{js_1} = -\frac{\sum_{k \neq s_1} b_{jk} a_{ks_1}}{a_{s_1 s_1}}\,, \end{aligned} \tag{5.10}$$

for $1 \le j \le m-i+1$. Finally, for $j > m-i+1$, we select the b_{jk} such that the matrix B is nonsingular. In this manner, in a unique step, we would change to a new basis $\{Y'_1, \dots, Y'_m\}$ verifying

$$\begin{aligned} &\langle Y'_1 : Y'_1\rangle(q_0), \dots, \langle Y'_{m-i+1} : Y'_{m-i+1}\rangle(q_0) \in \mathcal{D}_{q_0}\,, \\ &\langle Y'_j : Y'_j\rangle(q_0) = lc(Y'_1(q_0), \dots, Y'_m(q_0)) + a'_{jj}\langle Y_1 : Y_2\rangle(q_0)\,, \qquad m-i+2 \le j \le m\,, \\ &\langle Y'_k : Y'_l\rangle(q_0) \in \mathcal{D}_{q_0}\,, \qquad k < l,\ 1 \le k \le m-i+1\,, \end{aligned}$$

with possibly some of the $(a'_{jj})_{m-i+1 \le j \le m}$ being different from zero. Now, the above discussion can be redone in this context to assert the validity of the theorem. That is, we have to look for a change of basis B' in the vector fields $\{Y'_{m-i+2}, \dots, Y'_m\}$ such that the new ones, $\{Y''_{m-i+2}, \dots, Y''_m\}$, together with $\{Y'_1, \dots, Y'_{m-i+1}\}$, verify the sufficient conditions for STLCC at q_0. To find the change of basis for $\{Y'_{m-i+2}, \dots, Y'_m\}$, we have to consider the corresponding versions of Cases A and B. If we repeatedly fall into Case B, then we come to discard the same possibility that we encountered in the treatment of Case A3, which can be done again by means of Lemma 5.1. □

To recap, the steps of the proof can be summarized as follows (see Figure 5.1). First, we have considered the case when there exists for all i an s_i such that $a^{(i)}_{s_i s_i} \neq 0$. We have seen that this case can be subdivided into three: one (Case A1) ensuring the desired change of basis, another one (Case A2) in which either one obtains the basis or one contradicts the hypothesis of STLCC, and a third one (Case A3), where an intermediate change of basis is performed that allows us to focus on the search for a change of basis for $m-1$ of the new vector fields. Then, under the same assumption on the new coefficients, a'_{jk} (i.e., for all i, there exists an s_i such that $a^{(i)\prime}_{s_i s_i} \neq 0$), we can reproduce the former discussion. We cannot repeatedly fall into Case A3, since we would contradict the controllability assumption. Finally, we have treated the case when this type of "circular" process is broken (Case B); that is, when there exists an i such that $a^{(i)}_{kk} = 0$ for all $k \neq s_1, \ldots, s_{i-1}$. What we have shown then is that this leads to either a new basis of input vector fields satisfying the sufficient conditions for STLCC or a reduced situation where we can at the same time "get rid" of the problems associated with $m-i+1$ vector fields.

Remark 5.3. Notice that the proof of this result can be reproduced for the corresponding notions of accessibility and controllability at zero velocity. Indeed, a mechanical control system of the form (2.2) with $m = n-1$, which is STLC at q_0 and zero velocity is, in particular, STLCC at q_0. Then, Theorem 5.2 implies that there exists a basis of input vector fields $\mathcal{Y}$ satisfying the sufficient conditions of Theorem 3.2, so the same result is also valid for local controllability at zero velocity.

COROLLARY 5.4. *Let Q be a three-dimensional analytic manifold and let Y_1, Y_2 be analytic vector fields on Q. Consider the control system* (5.1) *and assume that it is LCA at $q_0 \in Q$. Let A be the 2×2 symmetric matrix whose elements are given by*

$$\begin{aligned}
\langle Y_1 : Y_1 \rangle(q_0) &= lc(Y_1(q_0), Y_2(q_0)) + a_{11}\langle Y_1 : Y_2 \rangle(q_0)\,,\\
\langle Y_2 : Y_2 \rangle(q_0) &= lc(Y_1(q_0), Y_2(q_0)) + a_{22}\langle Y_1 : Y_2 \rangle(q_0)\,,\\
\langle Y_1 : Y_2 \rangle(q_0) &= a_{12}\langle Y_1 : Y_2 \rangle(q_0)\,.
\end{aligned}$$

Then the system is locally configuration controllable at q_0 if and only if $\det A < 0$.

Proof. The result follows from the proof of Theorem 5.2 by noting that $\det A < 0$ corresponds to Case A1, $\det A > 0$ to Case A2, and $\det A = 0$ to Case A3. ∎

Remark 5.5. Note that Corollary 5.4 together with Theorem 5.2 completely characterize the configuration controllability properties of mechanical control systems with three degrees of freedom, since fully actuated systems are obviously STLCC.

6. Examples.

6.1. The planar rigid body. Consider a planar rigid body [24]. Fix a point $P \in \mathbb{R}^2$ and let $\{e_1, e_2\}$ be the standard orthonormal frame at that point. Let $\{d_1, d_2\}$ be an orthonormal frame attached to the body at its center of mass. The configuration manifold is then $SE(2)$, with coordinates (x, y, θ), where (x, y) describe the position of the center of mass and θ the orientation of the frame $\{d_1, d_2\}$ with respect to $\{e_1, e_2\}$.

The inputs of the system consist of a force F^1 applied at a distance h from the center of mass CM and a torque, F^2, about CM (see Figure 6.1). In coordinates, the input forces are given by

$$F^1 = -\sin\theta dx + \cos\theta dy - h d\theta\,, \qquad F^2 = d\theta\,.$$

The Riemannian metric is

$$g = m dx \otimes dx + m dy \otimes dy + J d\theta \otimes d\theta\,,$$

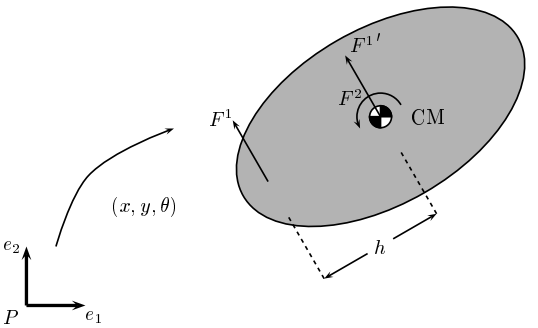


FIG. 6.1. *The planar rigid body.*

where m is the mass of the body and J its moment of inertia.

The input vector fields can be computed via $\flat_g^{-1}$ as

$$Y_1 = -\frac{\sin\theta}{m}\frac{\partial}{\partial x} + \frac{\cos\theta}{m}\frac{\partial}{\partial y} - \frac{h}{J}\frac{\partial}{\partial\theta}d\theta\,, \qquad Y_2 = \frac{1}{J}\frac{\partial}{\partial\theta}\,.$$

One can easily show that the planar body is LCA [24]. However, the inputs Y_1, Y_2 fail to satisfy the sufficient conditions for STLCC. In fact,

$$\begin{aligned}
\langle Y_1 : Y_1\rangle &= \frac{2h\cos\theta}{mJ}\frac{\partial}{\partial x} + \frac{2h\sin\theta}{mJ}\frac{\partial}{\partial y}\,,\\
\langle Y_1 : Y_2\rangle &= -\frac{\cos\theta}{mJ}\frac{\partial}{\partial x} - \frac{\sin\theta}{mJ}\frac{\partial}{\partial y}\,,\\
\langle Y_2 : Y_2\rangle &= 0\,.
\end{aligned}$$

Therefore, $\{Y_1, Y_2, \langle Y_1 : Y_2\rangle\}$ are linearly independent and $\langle Y_1 : Y_1\rangle = -2h\langle Y_1 : Y_2\rangle$. Theorem 5.2 ensures STLCC if and only if there exists a basis of input vector fields satisfying the sufficient conditions. We have that

$$\det A = \det\begin{pmatrix} -2h & 1\\ 1 & 0\end{pmatrix} = -1 < 0\,,$$

and consequently, by Corollary 5.4, the system is locally configuration controllable. Indeed, this example falls into Case A1 of the proof of Theorem 5.2. Accordingly, we obtain the change of basis $Y_1' = Y_1 + hY_2$, $Y_2' = Y_2$. This yields

$$\langle Y_1' : Y_1'\rangle = \langle Y_2' : Y_2'\rangle = 0\,, \qquad \langle Y_1' : Y_2'\rangle = \langle Y_1 : Y_2\rangle\,,$$

which satisfies the sufficient conditions for STLCC. The new input vector field precisely corresponds to the force $F^{1'}$ in Figure 6.1.

6.2. A simple example. The following example does not necessarily correspond to a physical example, but it illustrates the proof of Theorem 5.2. Consider a mechanical control system on $\mathbb{R}^3$, with coordinates (x, y, z). The Riemannian metric is given by

$$g = dx \otimes dx + dy \otimes dy + dz \otimes dz$$

and the input vector fields are

$$Y_1 = z\frac{\partial}{\partial x} + \frac{\partial}{\partial y} + \frac{1}{4}\frac{\partial}{\partial z}\,, \qquad Y_2 = y\frac{\partial}{\partial x} + \frac{1}{4}\frac{\partial}{\partial y} - \frac{1}{2}\frac{\partial}{\partial z}\,.$$

In coordinates, we have the following control equations:

$$\ddot{x} = u_1 z + u_2 y\,, \quad \ddot{y} = u_1 + \frac{u_2}{4}\,, \quad \ddot{z} = \frac{u_1}{4} - \frac{u_2}{2}\,. \tag{6.1}$$

Since

$$\langle Y_1 : Y_1\rangle = \langle Y_1 : Y_2\rangle = \langle Y_2 : Y_2\rangle = \frac{1}{2}\frac{\partial}{\partial x}\,,$$

we deduce that $\operatorname{span}\{Y_1(q), Y_2(q), \langle Y_1 : Y_2\rangle(q)\} = T_qQ$ for all $q \in Q$ and the system (6.1) is LCA. However, Corollary 5.4 implies that it is not STLCC, since $\det A = 0$. Going through the proof of Theorem 5.2, we see that this example falls into Case A3. Choosing the change of basis

$$B = \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix},$$

we get the new input vector fields $Y_1' = -Y_1 + Y_2$ and $Y_2' = Y_1 + Y_2$. Now, we have

$$\langle Y_1' : Y_1'\rangle = 0\,, \quad \langle Y_1' : Y_2'\rangle = 0\,, \quad \langle Y_2' : Y_2'\rangle = 2\frac{\partial}{\partial x}\,.$$

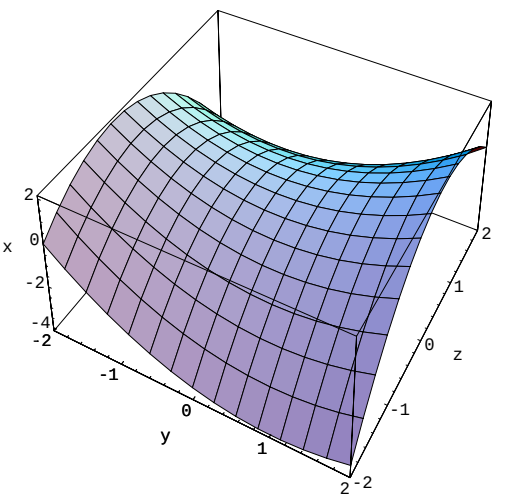


FIG. 6.2. *The level surface* $\phi(x, y, z) = 0$.

We can compute explicitly the function ϕ of Lemma 5.1 for this example. The flows of $Z_1 = Y_1'$, $Z_2 = Y_2'$, $Z_3 = -\langle Y_2' : Y_2'\rangle$ are given by

$$\begin{aligned}
\Psi_1(t)(x, y, z) &= (x + (y - z)t, y - 3t/4, z - 3t/4)\,,\\
\Psi_2(t)(x, y, z) &= (x + (y + z)t + t^2/2, y + 5t/4, z - t/4)\,,\\
\Psi_3(t)(x, y, z) &= (x - 2t, y, z)\,.
\end{aligned}$$

Letting (x_0, y_0, z_0) be an arbitrary point, one verifies

$$\begin{aligned}\Psi_1(t_1) \circ \Psi_2(t_2) \circ \Psi_3(t_3)(x_0, y_0, z_0) \\ = \left(x_0 - 2t_3 + \left(y_0 + z_0 + \frac{1}{2} t_2 \right) t_2 + t_1 \left(y_0 - z_0 - \frac{3}{2} t_2 \right), \right. \\ \left. y_0 - \frac{3}{4} t_1 + \frac{5}{4} t_2, z_0 - \frac{3}{4} t_1 - \frac{1}{4} t_2 \right).\end{aligned}$$

We may solve for $\phi(x, y, z) = t_3$ as

$$\begin{aligned}&\phi(x, y, z) \\ &= \frac{1}{18} \left(-9(x - x_0) + 4(y^2 - y y_0 + yz - 5 y_0 z - 2z^2 + y z_0 + 3 y_0 z_0 + 5 z z_0 - 3 z_0^2) \right).\end{aligned}$$

In Figure 6.2, we show the level set $\phi(x, y, z) = 0$ for $(x_0, y_0, z_0) = (0, 0, 0)$. The locally accessible configurations from $(0, 0, 0)$ are contained below the surface, where $\phi(x, y, z) \geq 0$.

7. Conclusions. In this paper, we have built on previous results on controllability and series expansions for the evolution of mechanical control systems within the affine connection formalism to demonstrate that the sufficient conditions encountered in [24] for STLCC are also necessary when the configuration manifold is n-dimensional and the system is actuated by $n - 1$ inputs, in the sense that there exists some basis of input vector fields that verifies them.

However $n - 1$ controls is a special case and is the simplest case next to fully actuated systems, which are always STLCC. For an arbitrary number of inputs, higher-order controllability will necessarily play a key role. Future research will be devoted to investigating the validity of the controllability conjecture in the full general case.

Appendix A. A simple lemma.

LEMMA A.1. *With the notation of Theorem* 5.2, *assume that* $(a^{(m-1)}_{s_{m-1}s_m})^2 - a^{(m-1)}_{s_{m-1}s_{m-1}} a^{(m-1)}_{s_m s_m} = 0$. *Then the coefficients* C_k *given by* (5.8) *verify*

$$\sum_{k=1}^{m} a_{kl} C_k = 0\,, \qquad 1 \leq l \leq m\,.$$

Proof. From (5.8), one can obtain the following recurrence formula for the coefficients C_k:

$$C_{s_m} = 1\,, \quad C_{s_j} = -\frac{1}{a^{(j)}_{s_j s_j}} \left(\sum_{i=j+1}^{m} a^{(j)}_{s_i s_j} C_{s_i} \right), \quad 1 \leq j \leq m-1\,. \tag{A.1}$$

Let us denote

$$\Sigma(l) = \sum_{k=1}^{m} a_{kl} C_k\,.$$

It is easy to see that $\Sigma(s_1) = 0$. Indeed, using (A.1), we have that

$$\Sigma(s_1) = a_{s_1 s_1} C_{s_1} + \sum_{i=2}^{m} a_{s_i s_1} C_{s_i} = -\sum_{i=2}^{m} a_{s_i s_1} C_{s_i} + \sum_{i=2}^{m} a_{s_i s_1} C_{s_i} = 0\,.$$

To prove the result for the remaining indices we can do the following. First, note that

$$a_{s_1 s_j} C_{s_1} = -\frac{a_{s_1 s_j}}{a_{s_1 s_1}} \left(\sum_{i=2}^{m} a_{s_i s_j} C_{s_i} \right) = -\sum_{i=2}^{m} \left(\frac{a_{s_1 s_j} a_{s_i s_j}}{a_{s_1 s_1}} \right) C_{s_i} \,.$$

Then, substituting in $\Sigma(s_j)$, we get

$$\begin{aligned} \Sigma(s_j) &= -\sum_{i=2}^{m} \left(\frac{a_{s_1 s_j} a_{s_i s_j}}{a_{s_1 s_1}} \right) C_{s_i} + \sum_{i=2}^{m} a_{s_i s_j} C_{s_i} \\ &= \sum_{i=2}^{m} \left(\frac{a_{s_i s_j} a_{s_1 s_1} - a_{s_1 s_j} a_{s_i s_j}}{a_{s_1 s_1}} \right) C_{s_i} = -\frac{1}{a_{s_1 s_1}} \left(\sum_{i=2}^{m} a^{(2)}_{s_i s_j} C_{s_i} \right), \end{aligned}$$

where we have used the definition (5.5) for the coefficients $a^{(j)}_{kl}$. This procedure can be iterated to obtain the general expression

$$\Sigma(s_j) = \frac{(-1)^k}{a_{s_1 s_1} a^{(2)}_{s_2 s_2} \dots a^{(k)}_{s_k s_k}} \left(\sum_{i=k+1}^{m} a^{(k+1)}_{s_i s_j} C_{s_i} \right), \tag{A.2}$$

which is valid for any $1 \le k \le m-2$.

Now, consider the cases $2 \le j \le m-1$. Take $k = j-1$. Then, using (A.2),

$$\begin{aligned} \Sigma(s_j) &= \frac{(-1)^{j-1}}{a_{s_1 s_1} a^{(2)}_{s_2 s_2} \dots a^{(j-1)}_{s_{j-1} s_{j-1}}} \left(\sum_{i=j}^{m} a^{(j)}_{s_i s_j} C_{s_i} \right) \\ &= \frac{(-1)^{j-1}}{a_{s_1 s_1} a^{(2)}_{s_2 s_2} \dots a^{(j-1)}_{s_{j-1} s_{j-1}}} \left(a^{(j)}_{s_j s_j} C_{s_j} + \sum_{i=j+1}^{m} a^{(j)}_{s_i s_j} C_{s_i} \right) = 0 \,, \end{aligned}$$

where in the last equality we have used (A.1). Finally, if $j = m$, we have that

$$\begin{aligned} \Sigma(s_m) &= \frac{(-1)^{m-2}}{a_{s_1 s_1} a^{(2)}_{s_2 s_2} \dots a^{(m-2)}_{s_{m-2} s_{m-2}}} \left(a^{(m-1)}_{s_{m-1} s_m} C_{s_{m-1}} + a^{(m-1)}_{s_m s_m} C_{s_m} \right) \\ &= \frac{(-1)^{m-2}}{a_{s_1 s_1} a^{(2)}_{s_2 s_2} \dots a^{(m-2)}_{s_{m-2} s_{m-2}}} \left(-\frac{(a^{(m-1)}_{s_{m-1} s_m})^2}{a^{(m-1)}_{s_{m-1} s_{m-1}}} + a^{(m-1)}_{s_m s_m} \right). \end{aligned}$$

From the hypothesis $(a^{(m-1)}_{s_{m-1} s_m})^2 - a^{(m-1)}_{s_{m-1} s_{m-1}} a^{(m-1)}_{s_m s_m} = 0$, we conclude that $\Sigma(s_m) = 0$, and this completes the proof. □

Acknowledgments. We wish to thank F. Cantrijn for several helpful suggestions and the Department of Mathematical Physics and Astronomy of the University of Ghent for its kind hospitality. We would also like to thank the anonymous reviewers for their useful comments which helped us improve the presentation of the manuscript.

REFERENCES

[1] R. Abraham and J.E. Marsden, *Foundations of Mechanics*, 2nd ed., Benjamin-Cummings, Reading, MA, 1978.

[2] A.A. Agrachev and R.V. Gamkrelidze, *Local controllability and semigroups of diffeomorphisms*, Acta Appl. Math., 32 (1993), pp. 1–57.

[3] F. Bullo, *Series expansions for the evolution of mechanical control systems*, SIAM J. Control Optim., 40 (2001), pp. 166–190.
[4] F. Bullo, N.E. Leonard, and A.D. Lewis, *Controllability and motion algorithms for underactuated Lagrangian systems on Lie groups*, IEEE Trans. Automat. Control, 45 (2000), pp. 1437–1454.
[5] F. Bullo and A.D. Lewis, *Configuration controllability of mechanical systems on Lie groups*, in Proceedings of the International Symposium on the Mathematical Theory of Networks and Systems, St. Louis, MO, 1996.
[6] F. Bullo and M. Zefran, *On mechanical control systems with nonholonomic constraints and symmetries*, Systems Control Lett., 45 (2002), pp. 133–143.
[7] K.T. Chen, *Integration of paths, geometric invariants and a generalized Baker-Hausdorff formula*, Ann. of Math., 67 (1957), pp. 164–178.
[8] J. Cortés, S. Martínez, and F. Bullo, *On nonlinear controllability and series expansions for Lagrangian systems with dissipative forces*, IEEE Trans. Automat. Control, 47 (2002), pp. 1396–1401.
[9] J. Cortés, S. Martínez, J.P. Ostrowski, and H. Zhang, *Simple mechanical control systems with constraints and symmetry*, SIAM J. Control Optim., 41 (2002), pp. 851–874.
[10] M. Fliess, *Fonctionelles causales non linéares et indeterminées non commutatives*, Bull. Soc. Math., 109 (1981), pp. 3–40.
[11] J.E. Hauser, S.S. Sastry, and G. Meyer, *Nonlinear control design for slightly nonminimum phase systems: Application to V/STOL aircraft*, Automatica J. IFAC, 28 (1992), pp. 665–679.
[12] H. Hermes, *Control systems which generate decomposable Lie algebras*, J. Differential Equations, 44 (1982), pp. 166–187.
[13] A. Isidori, *Nonlinear Control Systems*, 3rd ed., Communications and Control Engineering Series, Springer–Verlag, Berlin, 1995.
[14] M. Kawski, *The complexity of deciding controllability*, Systems Control Lett., 15 (1990), pp. 9–14.
[15] M. Kawski, *High-order small-time local controllability*, in Nonlinear Controllability and Optimal Control, H.J. Sussmann, ed., Marcel Dekker, New York, 1990, pp. 441–477.
[16] M. Kawski and H.J. Sussmann, *Noncommutative power series and formal Lie-algebraic techniques in nonlinear control theory*, in Operators, Systems and Linear Algebra, U. Helmke, D. Pratzel-Wolters, and E. Zerz, eds., Teubner, Stuttgart, 1997, pp. 111–128.
[17] S.D. Kelly and R.M. Murray, *Geometric phases and robotic locomotion*, J. Robotic Systems, 12 (1995), pp. 417–431.
[18] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry*, Vol. I, Wiley-Interscience, New York, London, 1963.
[19] G. Lafferriere and H.J. Sussmann, *A differential geometric approach to motion planning*, in Nonholonomic Motion Planning, Z.X. Li and J.F. Canny, eds., Kluwer Academic Publishers, Norwell, MA, 1997, pp. 235–270.
[20] N.E. Leonard and P.S. Krishnaprasad, *Motion control of drift-free, left-invariant systems on Lie groups*, IEEE Trans. Automat. Control, 40 (1995), pp. 1539–1554.
[21] A.D. Lewis, *Local configuration controllability for a class of mechanical systems with a single input*, in Proceedings of the 4th European Control Conference, European Union Control Association, Brussels, Belgium, 1997.
[22] A.D. Lewis, *Affine connections and distributions with applications to nonholonomic mechanics*, Rep. Math. Phys., 42 (1998), pp. 135–164.
[23] A.D. Lewis, *Simple mechanical control systems with constraints*, IEEE Trans. Automat. Control, 45 (2000), pp. 1420–1436.
[24] A.D. Lewis and R.M. Murray, *Configuration controllability of simple mechanical control systems*, SIAM J. Control Optim., 35 (1997), pp. 766–790.
[25] A.D. Lewis and R.M. Murray, *Configuration controllability of simple mechanical control systems*, SIAM Rev., 41 (1999), pp. 555–574.
[26] K.M. Lynch, N. Shiroma, H. Arai, and K. Tanie, *Collision-free trajectory planning for a 3-DOF robot with a passive joint*, Int. J. Robotics Research, 19 (2000), pp. 1171–1184.
[27] W. Magnus, *On the exponential solution of differential equations for a linear operator*, Comm. Pure Appl. Math., 7 (1954), pp. 649–673.
[28] P. Martin, S. Devasia, and B. Paden, *A different look at output tracking: Control of a VTOL aircraft*, Automatica J. IFAC, 32 (1996), pp. 101–107.
[29] S. Martínez and J. Cortés, *Motion control algorithms for mechanical systems with symmetries*, Acta Appl. Math., to appear.
[30] H. Nijmeijer and A.J. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer–Verlag, New York, 1990.

[31] J.P. Ostrowski and J.W. Burdick, *Controllability tests for mechanical systems with symmetries and constraints*, J. Appl. Math. Comp. Sci., 7 (1997), pp. 101–127.

[32] M. Rathinam and R.M. Murray, *Configuration flatness of Lagrangian systems underactuated by one control*, SIAM J. Control Optim., 36 (1998), pp. 164–179.

[33] E.D. Sontag, *Controllability is harder to decide than accessibility*, SIAM J. Control Optim., 26 (1988), pp. 1106–1118.

[34] H.J. Sussmann, *Lie brackets and local controllability: A sufficient condition for scalar-input systems*, SIAM J. Control Optim., 21 (1983), pp. 686–713.

[35] H.J. Sussmann, *A general theorem on local controllability*, SIAM J. Control Optim., 25 (1987), pp. 158–194.

[36] H.J. Sussmann, *A product expansion of the Chen series*, in Theory and Applications of Nonlinear Control Systems, C.I. Byrnes and A. Lindquist, eds., North–Holland, Amsterdam, 1986, pp. 323–335.

[37] H. Zhang and J.P. Ostrowski, *Control algorithms using affine connections on principal fiber bundles*, in Proceedings of the IFAC Workshop on Lagrangian and Hamiltonian Methods for Nonlinear Control, Princeton, NJ, 2000.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 6, pp. 1922–1945

TRAJECTORY-BASED LOCAL APPROXIMATIONS OF ORDINARY DIFFERENTIAL EQUATIONS*

LUC MOREAU† AND DIRK AEYELS†

Abstract. The present paper introduces a new definition of local approximation for ordinary differential equations locally around an equilibrium point. This definition generalizes the well-known linear and homogeneous approximations. The approach is based on approximating trajectories near the origin. This concept of local approximation is applied to the study of local uniform asymptotic stability, leading to alternative proofs for and extensions of several existing stability results.

Key words. asymptotic stability, ordinary differential equations, averaging, perturbations

AMS subject classifications. 34C29, 34D20, 34E10, 93D20

PII. S0363012900370776

1. Introduction.

It is of course well-known that in general the solutions of a nonlinear differential equation cannot be obtained in closed form. Accordingly, various methods have been developed to approximate these solutions. Apart from numerical integration techniques, several asymptotic methods are available. These asymptotic methods typically reflect the structural properties possessed by the differential equation under consideration. Averaging techniques [24], for example, are applicable when the time variation in the constitutive relation is much faster then the rate of change of the state with time. Singular perturbation techniques [10] may be applied when some components of the state evolve on a much faster time scale than other state components.

In a standard setting, asymptotic methods are concerned with differential equations depending on a small parameter. Asymptotic methods enable us to approximate the solutions of this equation by the solutions of a simpler equation (or several simpler equations). Standard results in this context are concerned with closeness properties of solutions. Typically, it is proven that solutions of the original equation converge uniformly on compact time-intervals to solutions of the simpler, limiting equation as the parameter tends to zero.

A different class of results that has been obtained in the literature on asymptotic methods is concerned with stability properties. The subject of these studies is the extent to which stability properties of the original differential equation (for small enough values of the parameter) may be inferred from stability properties of the simpler, limiting system. These results are typically obtained by means of Lyapunov techniques, as in [9, 13, 20], or by means of generalized Lyapunov techniques, as in [4] and [1, 23, 22].

In previous papers [17, 18] we have initiated a new approach to obtain such sta-

*Received by the editors March 28, 2000; accepted for publication (in revised form) July 18, 2002; published electronically February 27, 2003. This paper presents research results of the Belgian Programme on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture. The scientific responsibility rests with its authors. A preliminary version of this work has appeared in Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, AZ, 1999.

http://www.siam.org/journals/sicon/41-6/37077.html

†SYSTeMS group, Ghent University, Technologiepark 914, 9052 Zwijnaarde, Belgium (luc.moreau@rug.ac.be, dirk.aeyels@rug.ac.be). The first author is a Postdoctoral Fellow of the Fund for Scientific Research - Flanders (Belgium) (F.W.O.-Vlaanderen). Part of the first author's work was done while supported by a BOF grant of the Ghent University.

bility results. This approach is centered around the observation that many of these stability results may be obtained directly as a consequence of the closeness property for solutions, even if the convergence is only proven to be uniform on compact time-intervals. In other words, stated as a general principle, *it is possible to deduce stability properties for a dynamical system based upon an approximate analysis of its trajectories, even if this approximation is only proven to be valid on finite time-intervals.* Important features of the present approach are that it only makes use of elementary mathematical techniques and that, in contrast with other approaches based on (generalized) Lyapunov techniques, the present approach does not rely on converse Lyapunov theorems. Consequently, the present approach lends itself naturally to generalizations to differential equations with delay and to the study of input-to-state stability. (Such extensions are being studied in [19] and [29].)

The present work is a continuation of this line of research initiated in [17, 18]. In those papers, the original system, whose stability properties are to be analyzed, depends on a small parameter—recall that this is the natural setting for asymptotic methods. In the present paper we show that the presence of a small parameter is not always needed; the general principle mentioned above may also be applicable when the original system does not depend on a small parameter. In this case, instead of assuming the presence of a small parameter, we typically impose appropriate *homogeneity* assumptions on the right-hand side (RHS) of the differential equation. The stability property that will be studied in this case is local uniform asymptotic stability.

The paper is organized as follows. After the preliminaries in section 2, we present in section 3 the main theorem of the paper. This theorem relates closeness properties for trajectories with stability properties. It states that, if the solutions of an ordinary differential equation starting near the origin (which is assumed to be an equilibrium point) are sufficiently close (in some well-defined sense) to the solutions of another differential equation, which has a locally uniformly asymptotically stable equilibrium point at the origin, then the origin of the original system is also locally uniformly asymptotically stable. This result leads to the introduction of a new notion of local approximation for ordinary differential equations, whose definition is based on approximating trajectories near the origin. This new notion generalizes the well-known linear and homogeneous approximations. As an immediate consequence of our main result, we obtain that the null-solution of a given system is locally uniformly asymptotically stable if the origin of its local approximation is locally uniformly asymptotically stable. In section 4 we show how this concept of local approximation relates to more standard closeness results for differential equations depending on a small parameter. This relation is made explicit by means of a rescaling mechanism. Applications are given in section 5, where it is shown how the general theory of the paper enables us to recover and extend several known stability results. We conclude the paper in section 6.

We end this introduction with some references to related work. A rescaling mechanism similar to the present one has been used, for example, in [3] and [30, p. 227] in the context of, respectively, bifurcation analysis and renormalization techniques. The rescaling mechanism featuring in the present paper is more general, since we will associate different scaling factors with different coordinate axes. The fact that the presence of a small parameter is not needed in order to obtain stability results was already observed in [28] and [23] in the context of averaging theory. These references make use of, respectively, center manifold theory and generalized Lyapunov theorems. Related but independent work may be found in [13] and [25]. The first reference makes

use of Lyapunov theory; the second studies periodic differential equations (with period, say, 1) by means of nonlinear Floquet theory, the time-1-map being calculated by means of Lie and chronologico-algebraic tools.

2. Preliminary definitions. The state space of all dynamical systems in the present paper is $\mathbb{R}^n$ with $n \in \mathbb{N}$.

2.1. Homogeneity. Homogeneity refers to symmetry properties with respect to a family of dilation mappings. Homogeneity plays a prominent role in various aspects of nonlinear control theory. See, for example, [6, 8, 14, 20] for some applications in feedback control. In the present paper, homogeneity will play an important role in section 5, where the general theory of the paper is applied to particular examples.

Consider an n-tuple $r = (r_1, \ldots, r_n) \in ((0, \infty))^n$. We define the family of *dilation mappings* δ_λ^r $(\lambda \in (0, \infty))$ as

$$\delta_\lambda^r : \mathbb{R}^n \to \mathbb{R}^n : x \mapsto \delta_\lambda^r x = (\lambda^{r_1} x_1, \ldots, \lambda^{r_n} x_n). \tag{1}$$

A continuous function $h : \mathbb{R}^n \to \mathbb{R}$ is *r-homogeneous of degree $m \in [0, \infty)$* if $h(\delta_\lambda^r x) = \lambda^m h(x)$ for all $\lambda \in (0, \infty)$ and $x \in \mathbb{R}^n$. A continuous function $f : \mathbb{R}^n \to \mathbb{R}^n$ is *r-homogeneous of order $\tau \in [0, \infty)$* if $f(\delta_\lambda^r x) = \lambda^\tau \delta_\lambda^r f(x)$ for all $\lambda \in (0, \infty)$ and $x \in \mathbb{R}^n$. An *r-homogeneous norm* ρ is a continuous function $\rho : \mathbb{R}^n \to \mathbb{R}$ which is zero at the origin, strictly positive elsewhere, and r-homogeneous of degree 1.

Remark 1. A continuous function $f : \mathbb{R}^n \to \mathbb{R}^n$ which is r-homogeneous of order $\tau \geq 0$ will typically not be locally Lipschitz at the origin if $\tau < \max\{r_1, \ldots, r_n\} - \min\{r_1, \ldots, r_n\}$. When f is continuously differentiable on $\mathbb{R}^n \setminus \{0\}$, this follows from the homogeneity properties of the partial derivatives:

$$\frac{\partial f_i}{\partial x_j}(\delta_\lambda^r x) = \lambda^{\tau + r_i - r_j} \frac{\partial f_i}{\partial x_j}(x) \tag{2}$$

for all $x \in \mathbb{R}^n \setminus \{0\}$, $\lambda \in (0, \infty)$, and $i, j \in \{1, \ldots, n\}$. Indeed, if $\tau + r_i - r_j < 0$ for some pair of indices (i, j), then $\frac{\partial f_i}{\partial x_j}(\delta_\lambda^r x)$ may blow up as $\lambda \downarrow 0$ with x fixed. In the statements of the various results in the paper, we will therefore always take into account the possibility of the vectorfield being non-Lipschitz at the origin.

Remark 2. An r-homogeneous norm is not necessarily a norm in the topological sense since it might not satisfy the triangle inequality. For some considerations, however, an r-homogeneous norm ρ and the Euclidean norm $\|\cdot\|$ may be used interchangeably. For example, it is easy to see that

$$\rho(x) \to 0 \Leftrightarrow \|x\| \to 0, \tag{3}$$

$$\rho(x) \to \infty \Leftrightarrow \|x\| \to \infty. \tag{4}$$

In these cases it may be convenient to pass from an r-homogeneous norm to the Euclidean norm, since the Euclidean norm does satisfy the triangle inequality.

2.2. Dynamical systems and flows. All dynamical systems in the present paper are described by ordinary differential equations

$$\dot{x} = f(t, x),$$

where, by assumption, f is a continuous map from $\mathbb{R} \times \mathbb{R}^n$ to $\mathbb{R}^n$ and $\dot{x} = f(t, x)$ has the uniqueness property of solutions.[1] Recall that existence of solutions is guaranteed by

[1] That is, for every $(t_0, x_0) \in \mathbb{R} \times \mathbb{R}^n$, there is a solution $\xi : \mathrm{Dom}\,(\xi) \to \mathbb{R}^n$ of $\dot{x} = f(t, x)$ with $\xi(t_0) = x_0$ such that (i) $\mathrm{Dom}\,(\xi)$ is an open interval containing t_0 and (ii) for any other solution $\bar{\xi} : \mathrm{Dom}\,(\bar{\xi}) \to \mathbb{R}^n$ of $\dot{x} = f(t, x)$ with $\bar{\xi}(t_0) = x_0$, (a) $\mathrm{Dom}\,(\bar{\xi}) \subset \mathrm{Dom}\,(\xi)$ and (b) $\bar{\xi}$ and ξ coincide on $\mathrm{Dom}\,(\bar{\xi})$. We call ξ the *trajectory* of this system passing through state x_0 at time t_0.

continuity of f. These dynamical systems will be referred to as *admissible dynamical systems on* $\mathbb{R}^n$. We do not assume completeness of solutions; that is, we do not exclude finite escape times. Before we explicitly state sufficient conditions for uniqueness of solutions, we first introduce an interesting class of functions from $\mathbb{R} \times \mathbb{R}^n$ to $\mathbb{R}^n$.

DEFINITION 2.1 (class-$\mathcal{CLB}$ function). *A function* $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n : (t, x) \mapsto f(t, x)$ *is a class-*$\mathcal{CLB}$ *function if the following three conditions are all satisfied:*

1. f *is continuous in* (t, x)*;*
2. f *is locally Lipschitz in* $x \in \mathbb{R}^n \setminus \{0\}$ *uniformly with respect to* $t \in \mathbb{R}$*; that is, for each compact set* $K \subset \mathbb{R}^n \setminus \{0\}$*, there exists* $k \in [0, \infty)$ *such that*

$$\|f(t, x_1) - f(t, x_2)\| \leq k\|x_1 - x_2\| \qquad \forall t \in \mathbb{R} \quad \forall x_1, x_2 \in K;$$

3. f *is bounded in* $t \in \mathbb{R}$ *uniformly with respect to* x *in compact subsets of* $\mathbb{R}^n$*; that is, for each compact set* $K \subset \mathbb{R}^n$*, there exists* $M \in [0, \infty)$ *such that*

$$\|f(t, x)\| \leq M \qquad \forall t \in \mathbb{R} \quad \forall x \in K.$$

The following lemma provides sufficient conditions for uniqueness of solutions. It will be used in various proofs and may be of independent interest.

LEMMA 2.2. *Let* $r \in ((0, \infty))^n$ *and* $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n : (t, x) \mapsto f(t, x)$*. Assume that*

(a) $f \in \mathcal{CLB}$*;*

(b) $\delta^r_{1/\lambda} f(t, \delta^r_\lambda x)$ *remains bounded as* $\lambda \downarrow 0$*, uniformly with respect to* $t \in \mathbb{R}$ *and* x *in compact subsets of* $\mathbb{R}^n$*.*

Then the ordinary differential equation

$$\dot{x} = f(t, x) \tag{5}$$

has the uniqueness property of solutions.

Assumption (a) of the lemma implies uniqueness of solutions in the region $\mathbb{R}^n \setminus \{0\}$ of the state space, but does not exclude nonunique behavior at the origin. The possibility of nonuniqueness at the origin is ruled out by assumption (b). This may be seen as follows. First of all, notice that the origin is an equilibrium point by continuity of f and assumption (b). Suppose that the null solution is not unique, say, in forward time. That is, assume the existence of a solution $t \mapsto \zeta(t)$ that starts in the origin and leaves the origin in forward time. It is not difficult to see that this implies that $\sup\{||(\mathrm{d}/\mathrm{d}t)\delta^r_{1/\lambda}\zeta(t)|| : \delta^r_{1/\lambda}\zeta(t) \in K\} \to \infty$ as $\lambda \downarrow 0$, where K is a small neighborhood around the origin. This yields a contradiction, since $t \mapsto \delta^r_{1/\lambda}\zeta(t)$ is a solution of $\dot{x} = \delta^r_{1/\lambda} f(t, \delta^r_\lambda x)$, and the RHS of this equation is assumed to be bounded as $\lambda \downarrow 0$ as stated in assumption (b). A rigorous proof along these lines is given in Appendix A.

Remark 3. Lemma 2.2 is a modest extension of [14, Lemma 2]. In that reference, assumption (b) is replaced by the stronger assumption that f is r-homogeneous of order 0 in x. The present result includes this as a particular case but extends it, for example, to functions f which are r-homogeneous of order $\tau \geq 0$ in x, or to sums of functions which are r-homogeneous of various nonnegative orders.

Consider an admissible dynamical system on $\mathbb{R}^n$ and let $\phi(t, t_0, x_0)$ be the trajectory of this system passing through state x_0 at time t_0 evaluated at time t. The function $\phi : (t, t_0, x_0) \mapsto \phi(t, t_0, x_0)$ is the *flow* of this system. The domain of ϕ is open and ϕ is continuous on its domain [5, Chapter 5, Theorem 2.1].

Let $r \in ((0, \infty))^n$ and $\tau \in [0, \infty)$. An admissible dynamical system $\dot{x} = f(t, x)$ on $\mathbb{R}^n$ is *r-homogeneous of order τ* if f is r-homogeneous of order τ in x for all t. An admissible dynamical system $\dot{x} = f(t, x)$ on $\mathbb{R}^n$ and its corresponding flow ϕ are said to have the *(r, τ)-scaling property of trajectories* if

$$\phi(t, t_0, \delta_\lambda^r x_0) = \delta_\lambda^r \phi(\lambda^\tau t, \lambda^\tau t_0, x_0). \tag{6}$$

This means that, modulo a scaling of time, δ_λ^r maps trajectories to trajectories. As may be expected, there is a relationship between homogeneity properties of a vector-field and scaling properties of the corresponding trajectories. An admissible dynamical system $\dot{x} = f(t, x)$ on $\mathbb{R}^n$ that is r-homogeneous of order 0 has the $(r, 0)$-scaling property of trajectories, and an admissible dynamical system on $\mathbb{R}^n$ that is r-homogeneous of order $\tau > 0$ has the (r, τ)-scaling property of trajectories if the differential equation does not depend explicitly on time. These statements may be proven by verifying that the RHS of (6) viewed as a function of t is the solution of $\dot{x} = f(t, x)$ that passes through state $\delta_\lambda^r x_0$ at time t_0. It is important to notice that admissible dynamical systems on $\mathbb{R}^n$ that are r-homogeneous of order $\tau > 0$ in general do not have the (r, τ)-scaling property of trajectories if the differential equation depends explicitly on time.

3. Approximation of dynamical systems. We start this section with the main theorem of the paper: consider two admissible dynamical systems on $\mathbb{R}^n$ having an equilibrium point at the origin,

$$\dot{x} = f(t, x) \tag{7}$$

and

$$\dot{x} = g(t, x). \tag{8}$$

Assume that (8) has the (r, τ)-scaling property of trajectories for some $r \in ((0, \infty))^n$ and $\tau \geq 0$, and assume that the origin of (8) is locally uniformly asymptotically stable (LUAS). Theorem 3.1 states that, if trajectories of (7) that start near the origin are sufficiently close to trajectories of (8), then the origin of (7) is also LUAS. This theorem is the basis for all further result of the paper.

THEOREM 3.1. *Consider an admissible dynamical system $\dot{x} = g(t, x)$ on $\mathbb{R}^n$ with flow ψ having an equilibrium point at the origin. Consider also $r \in ((0, \infty))^n$, $\tau \geq 0$, and an r-homogeneous norm ρ. Assume that*

(a) *$\dot{x} = g(t, x)$ has the (r, τ)-scaling property of trajectories, and*

(b) *the origin of $\dot{x} = g(t, x)$ is LUAS.*

Then there exist $T \in [0, \infty)$ and $d \in (0, \infty)$ such that for every admissible dynamical system $\dot{x} = f(t, x)$ on $\mathbb{R}^n$ with flow ϕ having an equilibrium point at the origin, the following holds: if there exists $\sigma \in (0, \infty)$ such that for all $t_0 \in \mathbb{R}$ and $x_0 \in \mathbb{R}^n$ with $0 < \rho(x_0) \leq \sigma$

$$\begin{cases} \phi(t, t_0, x_0) \text{ exists} & \forall t \in [t_0, t_0 + \frac{T}{\rho(x_0)^\tau}], \\ \rho(\phi(t, t_0, x_0) - \psi(t, t_0, x_0)) < \rho(x_0) d & \forall t \in [t_0, t_0 + \frac{T}{\rho(x_0)^\tau}], \end{cases} \tag{9}$$

then the origin of $\dot{x} = f(t, x)$ is LUAS, and $\{x_0 \in \mathbb{R}^n : \rho(x_0) \leq \sigma\}$ is contained in its region of attraction.

Proof. We begin with studying the implications of assumptions (a) and (b) for the flow ψ of $\dot{x} = g(t, x)$. Since the origin of $\dot{x} = g(t, x)$ is assumed to be a LUAS

equilibrium point, there exist $c \in (0, \infty)$, $m \in [1, \infty)$, and $T \in (0, \infty)$ such that

$$\forall t_0 \in \mathbb{R}, \quad \forall x_0 \in \mathbb{R}^n \text{ with } \rho(x_0) = c, \quad \rho(\psi(t, t_0, x_0)) \leq \begin{cases} cm & \forall t \geq t_0, \\ c\frac{1}{2} & \forall t \geq t_0 + \frac{T}{c^\tau}. \end{cases} \tag{10}$$

(The particular choice for the constants cm, $c/2$, and T/c^τ in (10) is inspired by expression (11), which we are about to prove.) We show that, by the scaling property of trajectories of ψ, (10) implies

(11)

$$\forall t_0 \in \mathbb{R}, \quad \forall x_0 \in \mathbb{R}^n \text{ with } x_0 \neq 0, \quad \rho(\psi(t, t_0, x_0)) \leq \begin{cases} \rho(x_0)m & \forall t \geq t_0, \\ \rho(x_0)\frac{1}{2} & \forall t \geq t_0 + \frac{T}{\rho(x_0)^\tau}. \end{cases}$$

Indeed, for $x_0 \neq 0$,

$$\rho(\psi(t, t_0, x_0))$$

may be rewritten as

$$\rho\left(\psi(t, t_0, \delta^r_{\rho(x_0)/c}\delta^r_{c/\rho(x_0)}x_0)\right)$$

or, by the scaling property of trajectories of ψ, as

$$\rho\left(\delta^r_{\rho(x_0)/c}\psi\left((\rho(x_0)/c)^\tau t, (\rho(x_0)/c)^\tau t_0, \delta^r_{c/\rho(x_0)}x_0\right)\right)$$

and, since ρ is r-homogeneous of degree 1, as

$$\frac{\rho(x_0)}{c}\rho\left(\psi\left((\rho(x_0)/c)^\tau t, (\rho(x_0)/c)^\tau t_0, \delta^r_{c/\rho(x_0)}x_0\right)\right).$$

Since $\rho(\delta^r_{c/\rho(x_0)}x_0) = c$, we apply (10) and obtain

$$\begin{aligned} &\forall t_0 \in \mathbb{R}, \quad \forall x_0 \in \mathbb{R}^n \text{ with } x_0 \neq 0, \\ &\qquad \rho(\psi(t, t_0, x_0)) = \frac{\rho(x_0)}{c}\rho\left(\psi\left((\rho(x_0)/c)^\tau t, (\rho(x_0)/c)^\tau t_0, \delta^r_{c/\rho(x_0)}x_0\right)\right) \\ &\qquad\qquad \leq \begin{cases} \frac{\rho(x_0)}{c}cm & \forall\, (\rho(x_0)/c)^\tau t \geq (\rho(x_0)/c)^\tau t_0, \\ \frac{\rho(x_0)}{c}c\frac{1}{2} & \forall\, (\rho(x_0)/c)^\tau t \geq (\rho(x_0)/c)^\tau t_0 + \frac{T}{c^\tau}, \end{cases} \end{aligned}$$

yielding (11).

Next we derive some "triangle-like" inequalities for the r-homogeneous norm ρ. Let $d \in (0, \infty)$ be such that

$$\forall x_1, x_2 \in \mathbb{R}^n, \qquad \left.\begin{matrix} \rho(x_1) \leq \frac{1}{2} \\ \rho(x_2 - x_1) < d \end{matrix}\right\} \Rightarrow \rho(x_2) \leq \frac{3}{4}. \tag{12}$$

Equation (12) implies that for any $\lambda \in (0, \infty)$

$$\forall x_1, x_2 \in \mathbb{R}^n, \qquad \left.\begin{array}{r} \rho(x_1) \leq \lambda \frac{1}{2} \\ \rho(x_2 - x_1) < \lambda d \end{array}\right\} \Rightarrow \rho(x_2) \leq \lambda \frac{3}{4}. \tag{13}$$

Indeed, $\rho(x_1) \leq \lambda \frac{1}{2}$ iff $\rho\big(\delta^r_{1/\lambda} x_1\big) \leq \frac{1}{2}$ since ρ is r-homogeneous of degree 1. Similarly, $\rho(x_2 - x_1) < \lambda d$ iff $\rho\big(\delta^r_{1/\lambda}(x_2 - x_1)\big) < d$, which, by linearity of δ^r_λ, is equivalent to $\rho\big(\delta^r_{1/\lambda} x_2 - \delta^r_{1/\lambda} x_1\big) < d$. Therefore, it follows from (12) that $\rho\big(\delta^r_{1/\lambda} x_2\big) \leq \frac{3}{4}$. This, in turn, is equivalent to $\rho(x_2) \leq \lambda \frac{3}{4}$, from which (13) follows. Let $M \in (1, \infty)$ be such that

$$\forall x_1, x_2 \in \mathbb{R}^n, \qquad \left.\begin{array}{r} \rho(x_1) \leq m \\ \rho(x_1 - x_2) < d \end{array}\right\} \Rightarrow \rho(x_2) \leq M. \tag{14}$$

As above, this implies that for any $\lambda \in (0, \infty)$

$$\forall x_1, x_2 \in \mathbb{R}^n, \qquad \left.\begin{array}{r} \rho(x_1) \leq \lambda m \\ \rho(x_1 - x_2) < \lambda d \end{array}\right\} \Rightarrow \rho(x_2) \leq \lambda M. \tag{15}$$

Notice that the numbers $T \in [0, \infty)$ and $d \in (0, \infty)$ we have introduced so far are independent of $\dot{x} = f(t, x)$, as required in the statement of the theorem. We now turn our attention to $\dot{x} = f(t, x)$ and assume that there exists $\sigma \in (0, \infty)$ such that

$$\forall t_0 \in \mathbb{R}, \quad \forall x_0 \in \mathbb{R}^n \text{ with } 0 < \rho(x_0) \leq \sigma, \qquad \left\{\begin{array}{ll} \phi(t, t_0, x_0) \text{ exists} & \forall t \in [t_0,\, t_0 + \frac{T}{\rho(x_0)^\tau}], \\ \rho(\phi(t, t_0, x_0) - \psi(t, t_0, x_0)) < \rho(x_0) d & \forall t \in [t_0,\, t_0 + \frac{T}{\rho(x_0)^\tau}]. \end{array}\right. \tag{16}$$

We prove that this implies LUAS for the origin of $\dot{x} = f(t, x)$. Indeed, estimates (11), (13), (15), and (16) yield

$$\forall t_0 \in \mathbb{R}, \quad \forall x_0 \in \mathbb{R}^n \text{ with } 0 < \rho(x_0) \leq \sigma, \qquad \rho(\phi(t, t_0, x_0)) \leq \left\{\begin{array}{ll} \rho(x_0) M & \forall t \in [t_0,\, t_0 + \frac{T}{\rho(x_0)^\tau}], \\ \rho(x_0) \frac{3}{4} & \text{for } t = t_0 + \frac{T}{\rho(x_0)^\tau}. \end{array}\right. \tag{17}$$

The first inequality in (17) gives an upper bound for $\rho(\phi(t, t_0, x_0))$ on the interval $[t_0,\, t_0 + \frac{T}{\rho(x_0)^\tau}]$. The second inequality says that, after a time $\frac{T}{\rho(x_0)^\tau}$, the r-homogeneous norm ρ has at least decreased with a factor $\frac{3}{4}$ along trajectories of the system with initial state $0 < \rho(x_0) \leq \sigma$. We may then apply estimate (17) again with $\phi(t_0 + \frac{T}{\rho(x_0)^\tau}, t_0, x_0)$ as the new initial state, and so forth. This process of iterated applications of (17) may be formalized as follows: associated with a particular $t_0 \in \mathbb{R}$ and $x_0 \in \mathbb{R}^n$ satisfying $0 < \rho(x_0) \leq \sigma$ we introduce a sequence of times $t_0^{t_0,x_0} < t_1^{t_0,x_0} < t_2^{t_0,x_0} < \cdots \to \infty$ according to

$$\begin{aligned} t_0^{t_0,x_0} &= t_0, \\ t_i^{t_0,x_0} - t_{i-1}^{t_0,x_0} &= \tfrac{T}{\rho(x_0)^\tau} \big((\tfrac{4}{3})^\tau\big)^{i-1} \qquad \forall i \in \mathbb{N}. \end{aligned}$$

Then, applying (17) iteratively yields

$$\forall t_0 \in \mathbb{R}, \quad \forall x_0 \in \mathbb{R}^n \text{ with } 0 < \rho(x_0) \leq \sigma, \qquad \rho(\phi(t, t_0, x_0)) \leq \rho(x_0) (\tfrac{3}{4})^{i-1} M \qquad \forall t \in [t_{i-1}^{t_0,x_0},\, t_i^{t_0,x_0}] \quad \forall i \in \mathbb{N}. \tag{18}$$

This proves that the equilibrium point $x = 0$ of $\dot{x} = f(t, x)$ is LUAS and that $\{x_0 \in \mathbb{R}^n : \rho(x_0) \leq \sigma\}$ is contained in its region of attraction. ∎

Remark 4. If $\tau = 0$, then all the $t_i^{t_0,x_0}$ introduced in the proof above are equidistant with distance T. Estimate (18) then implies that the equilibrium point $x = 0$ of $\dot{x} = f(t, x)$ is locally uniformly exponentially stable with respect to the r-homogeneous norm ρ; that is, there exist $\mu \in [1, \infty)$, $\nu \in (0, \infty)$ such that

(19)
$$\forall t_0 \in \mathbb{R}, \quad \forall x_0 \in \mathbb{R}^n \text{ with } \rho(x_0) \leq \sigma, \quad \rho(\phi(t, t_0, x_0)) \leq \rho(x_0)\mu e^{-\nu(t-t_0)} \quad \forall t \geq t_0.$$

Notice that the number T in the statement of Theorem 3.1 depends on the dynamics of $\dot{x} = g(t, x)$ and thus may be hard to determine. Cases where the existence of σ is guaranteed for each $T \in [0, \infty)$ and $d \in (0, \infty)$ are therefore of particular interest. This leads to the following definition.

DEFINITION 3.2 ((r, τ)-approximation). *Consider admissible dynamical systems $\dot{x} = f(t, x)$ and $\dot{x} = g(t, x)$ on $\mathbb{R}^n$ with respective flows ϕ and ψ having an equilibrium point at the origin. Consider also $r \in ((0, \infty))^n$, $\tau \geq 0$, and an r-homogeneous norm ρ. System $\dot{x} = g(t, x)$ is an* (r, τ)-approximation *of $\dot{x} = f(t, x)$ if the following two conditions are both satisfied:*

Condition 1. *$\dot{x} = g(t, x)$ has the (r, τ)-scaling property of trajectories.*

Condition 2. *For each $T \in [0, \infty)$ satisfying $\{(t, t_0, x_0) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n : t \in [t_0, t_0 + T],\ \rho(x_0) = 1\} \subset \operatorname{Dom} \psi$ and for each $d \in (0, \infty)$, there exists $\sigma \in (0, \infty)$ such that for all $t_0 \in \mathbb{R}$ and $x_0 \in \mathbb{R}^n$ with $0 < \rho(x_0) \leq \sigma$*

$$\text{(20)} \qquad \begin{cases} \phi(t, t_0, x_0) \text{ exists} & \forall t \in [t_0, t_0 + \frac{T}{\rho(x_0)^\tau}], \\ \rho(\phi(t, t_0, x_0) - \psi(t, t_0, x_0)) < \rho(x_0)d & \forall t \in [t_0, t_0 + \frac{T}{\rho(x_0)^\tau}]. \end{cases}$$

Remark 5. An extra assumption on T is introduced in Condition 2 since solutions of $\dot{x} = g(t, x)$ need not be forward complete in general. This is in contrast with Theorem 3.1 where solutions of $\dot{x} = g(t, x)$ are guaranteed to be forward complete by assumption (b). Notice that, if $\{(t, t_0, x_0) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n : t \in [t_0, t_0 + T],\ \rho(x_0) = 1\} \subset \operatorname{Dom} \psi$, then $\{(t, t_0, x_0) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n : t \in [t_0, t_0 + \frac{T}{\rho(x_0)^\tau}],\ x_0 \neq 0\} \subset \operatorname{Dom} \psi$ by the (r, τ)-scaling property of trajectories, and the second expression in (20) indeed makes sense.

Remark 6. The definition of an (r, τ)-approximation is independent of the particular r-homogeneous norm ρ used. This may be proven based on the fact that for each two r-homogeneous norms ρ_1 and ρ_2, there exist $a, b \in (0, \infty)$ such that $a\rho_1(x) \leq \rho_2(x) \leq b\rho_1(x)$ for all $x \in \mathbb{R}^n$.

The definition of (r, τ)-approximation includes linear and homogeneous approximations as a particular case but is more general, as will be shown in section 5. It follows immediately from Theorem 3.1 that (r, τ)-approximations are useful for the purpose of stability analysis.

COROLLARY 3.3. *Consider an admissible dynamical system $\dot{x} = f(t, x)$ on $\mathbb{R}^n$ with an equilibrium point at the origin. This equilibrium point is LUAS if there exists, for some $r \in ((0, \infty))^n$ and $\tau \in [0, \infty)$, an (r, τ)-approximation of $\dot{x} = f(t, x)$ whose origin is a LUAS equilibrium point.*

Corollary 3.3 states that stability properties of a dynamical system—LUAS of the equilibrium point—may be inferred from stability properties of an (r, τ)-approximation. This may constitute an important simplification, since an (r, τ)-approximation is assumed to have the (r, τ)-scaling property of trajectories. In section 5 we will show

that Corollary 3.3 leads to alternative proofs for and generalizations of several existing stability results.

Remark 7. If $\tau = 0$, then the conclusion of Corollary 3.3 may be strengthened according to Remark 4. Consider an admissible dynamical system $\dot{x} = f(t,x)$ on $\mathbb{R}^n$ with an equilibrium point at the origin. This equilibrium point is locally uniformly *exponentially* stable with respect to an r-homogeneous norm if there exists, for some $r \in ((0,\infty))^n$, an $(r,0)$-approximation of $\dot{x} = f(t,x)$ whose origin is a LUAS equilibrium point.

Remark 8. It is instructive to notice that, by the (r,τ)-scaling property of trajectories, LUAS for the equilibrium point $x = 0$ of the (r,τ)-approximation actually implies global uniform asymptotic stability. Furthermore, if $\tau = 0$, then we actually have global uniform exponential stability with respect to an r-homogeneous norm. Both observations follow readily from estimate (11) in the proof of Theorem 3.1.

4. On the convergence analysis of trajectories. In the previous section, we have introduced (r,τ)-approximations and clarified their role in stability analysis. In order to apply this theory, we need to be able to verify Conditions 1 and 2 featuring in the definition of (r,τ)-approximation (Definition 3.2). In the present section, we discuss the second of these two conditions.

A possible approach to verify Condition 2 is based on the Gronwall lemma—this is the approach that will be taken in the present paper. It turns out, however, that a direct verification of Condition 2 based on the Gronwall lemma is complicated by the possible non-Lipschitz character of the vectorfields near the origin (see Remark 1). In order to avoid these complications, we introduce a rescaling mechanism that is adapted to the underlying family of dilation mappings. This rescaling mechanism "desingularizes" the non-Lipschitz behavior around the origin. Furthermore, this rescaling mechanism will reveal a close relationship between stability results and (r,τ)-approximations on the one hand and convergence results for trajectories of systems depending on a small parameter on the other hand.

Recall Condition 2 of Definition 3.2. As a first step, we reformulate this condition, introducing $\overline{x}_0$ and ε according to $\overline{x}_0 = \delta^r_{1/\rho(x_0)} x_0$ and $\varepsilon = \rho(x_0)$: for each $T \in [0,\infty)$ satisfying $\{(t,t_0,x_0) \in \mathbb{R}\times\mathbb{R}\times\mathbb{R}^n : t \in [t_0,\, t_0+T],\ \rho(x_0) = 1\} \subset \operatorname{Dom}\psi$ and for each $d \in (0,\infty)$, there exists $\sigma \in (0,\infty)$ such that for all $t_0 \in \mathbb{R}$ and $\overline{x}_0 \in \mathbb{R}^n$ with $\rho(\overline{x}_0) = 1$ and for all $\varepsilon \in (0,\sigma]$

$$\text{(21)}\qquad \begin{cases} \phi(t,t_0,\delta^r_\varepsilon \overline{x}_0) \text{ exists} & \forall t \in [t_0,\, t_0 + \frac{T}{\varepsilon^\tau}], \\ \rho(\phi(t,t_0,\delta^r_\varepsilon \overline{x}_0) - \psi(t,t_0,\delta^r_\varepsilon \overline{x}_0)) < \varepsilon d & \forall t \in [t_0,\, t_0 + \frac{T}{\varepsilon^\tau}]. \end{cases}$$

By homogeneity of ρ and linearity of $\delta^r_{1/\varepsilon}$, expression (21) is equivalent to

$$\text{(22)}\qquad \begin{cases} \delta^r_{1/\varepsilon}\phi(t,t_0,\delta^r_\varepsilon \overline{x}_0) \text{ exists} & \forall t \in [t_0,\, t_0 + \frac{T}{\varepsilon^\tau}], \\ \rho(\delta^r_{1/\varepsilon}\phi(t,t_0,\delta^r_\varepsilon \overline{x}_0) - \delta^r_{1/\varepsilon}\psi(t,t_0,\delta^r_\varepsilon \overline{x}_0)) < d & \forall t \in [t_0,\, t_0 + \frac{T}{\varepsilon^\tau}]. \end{cases}$$

Replacing the dummy variables t and t_0 by $\frac{t}{\varepsilon^\tau}$ and $\frac{t_0}{\varepsilon^\tau}$, respectively, this may be rewritten as

$$\text{(23)}\qquad \begin{cases} \delta^r_{1/\varepsilon}\phi(\frac{t}{\varepsilon^\tau}, \frac{t_0}{\varepsilon^\tau}, \delta^r_\varepsilon \overline{x}_0) \text{ exists} & \forall t \in [t_0,\, t_0 + T], \\ \rho(\delta^r_{1/\varepsilon}\phi(\frac{t}{\varepsilon^\tau}, \frac{t_0}{\varepsilon^\tau}, \delta^r_\varepsilon \overline{x}_0) - \delta^r_{1/\varepsilon}\psi(\frac{t}{\varepsilon^\tau}, \frac{t_0}{\varepsilon^\tau}, \delta^r_\varepsilon \overline{x}_0)) < d & \forall t \in [t_0,\, t_0 + T]. \end{cases}$$

Estimate (23) suggests the introduction of time-functions $\zeta^\varepsilon(\cdot, t_0, \overline{x}_0)$ and $\xi^\varepsilon(\cdot, t_0, \overline{x}_0)$ according to

$$\zeta^\varepsilon(t, t_0, \overline{x}_0) = \delta^r_{1/\varepsilon}\phi\left(\frac{t}{\varepsilon^\tau}, \frac{t_0}{\varepsilon^\tau}, \delta^r_\varepsilon \overline{x}_0\right), \tag{24}$$

$$\xi^\varepsilon(t, t_0, \overline{x}_0) = \delta^r_{1/\varepsilon}\psi\left(\frac{t}{\varepsilon^\tau}, \frac{t_0}{\varepsilon^\tau}, \delta^r_\varepsilon \overline{x}_0\right). \tag{25}$$

The function ζ^ε satisfies the differential relation

$$\frac{\partial}{\partial t}\zeta^\varepsilon(t, t_0, \overline{x}_0) = \frac{\partial}{\partial t}\delta^r_{1/\varepsilon}\phi\left(\frac{t}{\varepsilon^\tau}, \frac{t_0}{\varepsilon^\tau}, \delta^r_\varepsilon \overline{x}_0\right) \tag{26}$$

$$= \delta^r_{1/\varepsilon}\frac{\partial}{\partial t}\phi\left(\frac{t}{\varepsilon^\tau}, \frac{t_0}{\varepsilon^\tau}, \delta^r_\varepsilon \overline{x}_0\right) \tag{27}$$

$$= \delta^r_{1/\varepsilon} f\left(\frac{t}{\varepsilon^\tau}, \phi\left(\frac{t}{\varepsilon^\tau}, \frac{t_0}{\varepsilon^\tau}, \delta^r_\varepsilon \overline{x}_0\right)\right)\frac{1}{\varepsilon^\tau} \tag{28}$$

$$= \frac{1}{\varepsilon^\tau}\delta^r_{1/\varepsilon} f\left(\frac{t}{\varepsilon^\tau}, \delta^r_\varepsilon \zeta^\varepsilon(t, t_0, \overline{x}_0)\right) \tag{29}$$

and the initial condition

$$\zeta^\varepsilon(t_0, t_0, \overline{x}_0) = \delta^r_{1/\varepsilon}\phi\left(\frac{t_0}{\varepsilon^\tau}, \frac{t_0}{\varepsilon^\tau}, \delta^r_\varepsilon \overline{x}_0\right) \tag{30}$$

$$= \delta^r_{1/\varepsilon}\delta^r_\varepsilon \overline{x}_0 \tag{31}$$

$$= \overline{x}_0. \tag{32}$$

The function ξ^ε satisfies

$$\xi^\varepsilon(t, t_0, \overline{x}_0) = \psi(t, t_0, \overline{x}_0) \tag{33}$$

since ψ is assumed to have the (r, τ)-scaling property of trajectories.

Finally, measuring the distance between trajectories in terms of the Euclidean norm $\|\cdot\|$ instead of the r-homogeneous norm ρ (see Remark 2) and omitting the bar in the notation of the initial state, we conclude that Condition 2 of Definition 3.2 is equivalent to the following condition:

Condition 2bis. For each $T \in [0, \infty)$ satisfying $\{(t, t_0, x_0) \in \mathbb{R}\times\mathbb{R}\times\mathbb{R}^n : t \in [t_0, t_0 + T],\ \rho(x_0) = 1\} \subset \operatorname{Dom}\psi$, for each $d \in (0, \infty)$, there exists $\sigma \in (0, \infty)$ such that for all $t_0 \in \mathbb{R}$ and $x_0 \in \mathbb{R}^n$ with $\rho(x_0) = 1$ and for all $\varepsilon \in (0, \sigma]$

$$\begin{cases} \zeta^\varepsilon(t, t_0, x_0) \text{ exists} & \forall t \in [t_0, t_0 + T], \\ \|\zeta^\varepsilon(t, t_0, x_0) - \psi(t, t_0, x_0)\| < d & \forall t \in [t_0, t_0 + T], \end{cases} \tag{34}$$

where ζ^ε is the flow of

$$\dot{x} = \frac{1}{\varepsilon^\tau}\delta^r_{1/\varepsilon} f\left(\frac{t}{\varepsilon^\tau}, \delta^r_\varepsilon x\right) \tag{35}$$

and ψ the flow of

$$\dot{x} = g(t, x). \tag{36}$$

This shows that (r, τ)-approximations are related to convergence results for trajectories of systems depending on a small parameter ε: trajectories of (35) converge uniformly on compact time-intervals to trajectories of (36) as $\varepsilon \downarrow 0$.

5. Applications. In this section we give three particular examples of (r, τ)-approximations, leading to three corresponding stability results. As we have seen in the previous section, (r, τ)-approximations are related to convergence results for trajectories of systems depending on a small parameter. The three examples given here are, respectively, related to perturbation theory, averaging theory, and the theory of highly oscillatory systems [11, 27, 12]. This will lead to alternative proofs for several existing results, as well as to extensions of these results.

5.1. Zero-order approximations. Consider a dynamical system on $\mathbb{R}^n$ of the form

$$\dot{x} = X(t,x) + Y(t,x), \tag{37}$$

where X and Y are class-$\mathcal{CLB}$ functions. Assume that

(a) X is r-homogeneous of order zero in x for all t, and

(b) $\delta^r_{1/\lambda} Y(t, \delta^r_\lambda x) \to 0$ as $\lambda \downarrow 0$, uniformly with respect to $t \in \mathbb{R}$ and x in compact subsets of $\mathbb{R}^n$.

Assumption (b) is satisfied, for example, if Y is r-homogeneous of positive order in x, or if Y is a sum of functions which are r-homogeneous in x, possibly with different, positive orders. In other words, X is a zero-order approximation of $X+Y$ with respect to x near the origin. Consider the associated dynamical system on $\mathbb{R}^n$,

$$\dot{x} = X(t,x). \tag{38}$$

The assumptions imply that both (37) and (38) have the uniqueness property of solutions (see Lemma 2.2) and an equilibrium point at the origin.

THEOREM 5.1. *Let $r \in ((0,\infty))^n$. Consider systems* (37) *and* (38) *satisfying the assumptions introduced above. If the origin of* (38) *is LUAS, then the origin of* (37) *is locally uniformly exponentially stable with respect to an r-homogeneous norm.*

Theorem 5.1 generalizes several known stability results. For the particular case that X and Y are assumed to be time-invariant, we recover [7, Theorem 1] with $k = 1$. For the particular case that X and Y are assumed to be periodic in t, we recover a result that has been reported by Morin and Samson [21, Proposition 2, Part 2]. Also, the linearization principle corresponds to the special case where $r = (1, \ldots, 1)$ and X is assumed to be linear in x.

Remark 9. It is instructive to recall that LUAS for the origin of (38) actually implies that this equilibrium point is globally uniformly exponentially stable with respect to an r-homogeneous norm since (38) has the $(r, 0)$-scaling property of trajectories by assumption (a) of the theorem. Theorem 5.1 would not be true, in general, if the null-solution were only assumed to be (nonuniformly) exponentially stable with respect to an r-homogeneous norm. When X is linear in x, this issue is related to the notion of Lyapunov regularity, which plays a central role in the Lyapunov stability theory. It is known that the linearization principle requires a somewhat sophisticated assumption (Lyapunov regularity) to hold, if the null-solution of the linearized equation is only assumed to be (nonuniformly) exponentially stable [2, Theorem 1.1.2]. Lyapunov regularity, however, need not be assumed when the null-solution of the linearized equation is uniformly exponentially stable [2, Theorem 1.4.2]. Accordingly, Lyapunov regularity (or its appropriate generalization) need not be assumed in Theorem 5.1.

Theorem 5.1 is proven by showing that (38) is an $(r, 0)$-approximation of (37). Following the approach outlined in section 4, this turns out to be related to perturbation theory.

Proof. We prove that (38) is an $(r, 0)$-approximation of (37). As Condition 1 of Definition 3.2 is readily verified (see section 2.2), we focus upon Condition 2. Following the approach outlined in section 4 we prove the equivalent Condition 2bis instead. For the present particular case system (35) becomes

$$\begin{aligned} \dot{x} &= \delta^r_{1/\varepsilon} X(t, \delta^r_\varepsilon x) + \delta^r_{1/\varepsilon} Y(t, \delta^r_\varepsilon x) \\ &= X(t, x) + \delta^r_{1/\varepsilon} Y(t, \delta^r_\varepsilon x), \end{aligned} \tag{39}$$

since X is r-homogeneous of order 0. The second term in the RHS tends to zero as $\varepsilon \downarrow 0$ by assumption (b). In other words, system (39) is a perturbed version of system (38). We prove in Appendix B.1 that trajectories of (39) converge to trajectories of (38) as $\varepsilon \downarrow 0$ in the sense of Condition 2bis.[2] Theorem 5.1 now follows from Corollary 3.3 and Remark 7. ∎

5.2. Approximation and averaging. Consider a dynamical system on $\mathbb{R}^n$ of the form

$$\dot{x} = X(t, x) + Y(t, x), \tag{40}$$

where X and Y are class-$\mathcal{CLB}$ functions. Assume that

(a) X is r-homogeneous of order $\tau > 0$ in x for all t, and
(b) $\frac{1}{\lambda^\tau} \delta^r_{1/\lambda} Y(t, \delta^r_\lambda x) \to 0$ as $\lambda \downarrow 0$, uniformly with respect to $t \in \mathbb{R}$ and x in compact subsets of $\mathbb{R}^n$.

Assumption (b) is satisfied, for example, if Y is r-homogeneous of order $\tau' > \tau$ in x, or if Y is a sum of functions which are r-homogeneous in x with (possibly different) orders strictly larger than τ.

We introduce the average of X as the map

$$X_{\mathrm{av}} : \mathbb{R}^n \to \mathbb{R}^n : x \mapsto X_{\mathrm{av}}(x) = \lim_{T \to \infty} \frac{1}{T} \int_0^T X(t, x)\, \mathrm{d}t, \tag{41}$$

where it is assumed that this limit exists for all x. The assumptions on X imply that X_{av} is continuous, locally Lipschitz in $x \in \mathbb{R}^n \setminus \{0\}$, and r-homogeneous of order τ. Assume in addition that

(c) for each $T \in [0, \infty)$ and for each compact subset $K \in \mathbb{R}^n$,

$$\int_{t_0+\varsigma_1}^{t_0+\varsigma_2} \left\{ X\left(\frac{s}{\varepsilon^\tau}, x\right) - X_{\mathrm{av}}(x) \right\} \mathrm{d}s \to 0 \tag{42}$$

as $\varepsilon \downarrow 0$ uniformly with respect to $t_0 \in \mathbb{R}$, $\varsigma_1, \varsigma_2 \in [0, T]$, and $x \in K$.

Consider the associated averaged system on $\mathbb{R}^n$,

$$\dot{x} = X_{\mathrm{av}}(x). \tag{43}$$

The assumptions imply that both (40) and (43) have the uniqueness property of solutions (see Lemma 2.2) and an equilibrium point at the origin.

THEOREM 5.2. *Let $r \in ((0, \infty))^n$ and $\tau \in (0, \infty)$. Consider systems* (40) *and* (43) *satisfying the assumptions introduced above. If the origin of* (43) *is LUAS, then the origin of* (40) *is also LUAS.*

[2] Convergence results are available in the literature on perturbation theory, but, to the best of our knowledge, none of these results yields exactly the required convergence property Condition 2bis.

Remark 10. It is instructive to recall that LUAS for the origin of (43) actually implies that this equilibrium point is globally uniformly asymptotically stable since the assumptions of the theorem imply that (43) has the (r,τ)-scaling property of trajectories.

This theorem generalizes some known results from the literature. For the special case that X and Y are independent of t (and hence X_{av} coincides with X), we recover a well-known result by Hermes; see, for example, [7, Theorem 1] with $k > 1$. For the special case that Y vanishes identically and X is locally Lipschitz in x on the complete state space $\mathbb{R}^n$ (and hence also at the origin), we recover [23, Theorem 1]. Whereas the proof technique of [23] is based on generalized Lyapunov theorems, the present proof is based on closeness results for trajectories of fast time-varying systems via a rescaling mechanism. As already mentioned before, this rescaling mechanism enables us to relax the Lipschitz assumption at the origin.

Proof. We prove that (43) is an (r,τ)-approximation of (40). As Condition 1 of Definition 3.2 is readily verified (see section 2.2), we focus upon Condition 2. Following the approach outlined in section 4 we prove the equivalent Condition 2bis instead. For the present particular case system (35) becomes

$$\begin{aligned}\dot{x} &= \frac{1}{\varepsilon^\tau}\delta^r_{1/\varepsilon}X\left(\frac{t}{\varepsilon^\tau},\delta^r_\varepsilon x\right)+\frac{1}{\varepsilon^\tau}\delta^r_{1/\varepsilon}Y\left(\frac{t}{\varepsilon^\tau},\delta^r_\varepsilon x\right)\\ &= X\left(\frac{t}{\varepsilon^\tau},x\right)+\frac{1}{\varepsilon^\tau}\delta^r_{1/\varepsilon}Y\left(\frac{t}{\varepsilon^\tau},\delta^r_\varepsilon x\right),\end{aligned} \tag{44}$$

where we have used the homogeneity property of X. This is a fast time-varying differential equation since the RHS depends on time through $\frac{t}{\varepsilon^\tau}$, where ε is a small parameter. In addition the second term tends to zero as $\varepsilon \downarrow 0$ by assumption (b). We prove in Appendix B.2 that trajectories of (44) converge to trajectories of (43) as $\varepsilon \downarrow 0$ in the sense of Condition 2bis.[3] Theorem 5.2 now follows from Corollary 3.3. ∎

5.3. Approximation and Lie brackets. Consider a dynamical system on $\mathbb{R}^n$ of the form

$$\dot{x} = \sqrt{\gamma}\cos(\gamma t)X_1(x) + \sqrt{\gamma}\sin(\gamma t)X_2(x) + X_3(x), \tag{45}$$

where X_1, X_2, and X_3 are continuous on $\mathbb{R}^n$ and X_1 and X_2 (respectively, X_3) are of class C^2 (respectively, of class C^1) on $\mathbb{R}^n \setminus \{0\}$ and where γ is a strictly positive parameter. It is important to emphasize that γ is *not* assumed to be very large or very small. Let $\tau > 0$ and assume that

(a) X_1 and X_2 are r-homogeneous of order $\frac{\tau}{2}$,

(b) X_3 is r-homogeneous of order τ.

We introduce the *Lie bracket* of X_1 and X_2 as the map

$$[X_1,X_2]:\mathbb{R}^n\to\mathbb{R}^n:\quad x\mapsto[X_1,X_2](x)=\begin{cases}DX_2(x)\cdot X_1(x)-DX_1(x)\cdot X_2(x) & \forall x\in\mathbb{R}^n\setminus\{0\},\\ 0 & \text{for } x=0,\end{cases} \tag{46}$$

[3] Convergence results are available in the literature on averaging theory, but, to the best of our knowledge, none of these results yields exactly the required convergence property Condition 2bis.

where $DX_i(x)$ is the Jacobian of X_i evaluated at x and $\cdot$ indicates the matrix product. The assumptions on X_1 and X_2 imply that $[X_1, X_2]$ is continuous on $\mathbb{R}^n$, of class C^1 on $\mathbb{R}^n \setminus \{0\}$, and r-homogeneous of order τ. Consider the following system on $\mathbb{R}^n$:

$$\dot{x} = \frac{1}{2}[X_1, X_2](x) + X_3(x). \tag{47}$$

The assumptions imply that both (45) and (47) have the uniqueness property of solutions (see Lemma 2.2) and an equilibrium point at the origin.

THEOREM 5.3. *Let $r \in ((0, \infty))^n$ and $\tau \in (0, \infty)$. Consider systems* (45) *and* (47) *satisfying the assumptions introduced above. If the origin of* (47) *is LUAS, then the origin of* (45) *is also LUAS.*

Remark 11. It is instructive to recall that LUAS for the origin of (47) actually implies that this equilibrium point is globally uniformly asymptotically stable since, by the assumptions of the theorem, (47) has the (r, τ)-scaling property of trajectories.

Independently of the present research, closely related results have been reported by M'Closkey and Morin [13] and by Sarychev [25]. The approach in [13] is based on Lyapunov considerations; the approach in [25] is based on nonlinear Floquet theory. We also mention that this stability result may be applied to the constructive stabilization of driftless control affine systems; see [16].

Theorem 5.3 is proven by showing that (47) is an (r, τ)-approximation of (45). Following the approach outlined in section 4, this turns out to be related to the theory of highly oscillatory differential equations [11, 27, 12].

Proof. We prove that (47) is an (r, τ)-approximation of (45). As Condition 1 of Definition 3.2 is readily verified (see section 2.2), we focus upon Condition 2. Following the approach outlined in section 4 we prove the equivalent Condition 2bis instead. For the present particular case system (35) becomes

$$\begin{aligned}\dot{x} &= \frac{\sqrt{\gamma}}{\varepsilon^\tau}\delta^r_{1/\varepsilon}\cos\left(\gamma\frac{t}{\varepsilon^\tau}\right)X_1(\delta^r_\varepsilon x) + \frac{\sqrt{\gamma}}{\varepsilon^\tau}\delta^r_{1/\varepsilon}\sin\left(\gamma\frac{t}{\varepsilon^\tau}\right)X_2(\delta^r_\varepsilon x) + \frac{1}{\varepsilon^\tau}\delta^r_{1/\varepsilon}X_3(\delta^r_\varepsilon x)\\ &= \sqrt{\frac{\gamma}{\varepsilon^\tau}}\cos\left(\gamma\frac{t}{\varepsilon^\tau}\right)X_1(x) + \sqrt{\frac{\gamma}{\varepsilon^\tau}}\sin\left(\gamma\frac{t}{\varepsilon^\tau}\right)X_2(x) + X_3(x),\end{aligned} \tag{48}$$

where we have used the homogeneity properties of the X_i. Systems of the form (48) with ε a small parameter are studied in the literature on highly oscillatory systems. We prove in Appendix B.3 that trajectories of (48) converge to trajectories of (47) as $\varepsilon \downarrow 0$ in the sense of Condition 2bis.[4] Theorem 5.3 now follows from Corollary 3.3. ∎

In addition, we also prove the following semiglobal result, which is original.

THEOREM 5.4. *Consider the same data and assumptions of Theorem* 5.3. *If the origin of* (47) *is LUAS, then the origin of* (45) *is semiglobally uniformly asymptotically stable as $\gamma \to \infty$.*

Remark 12. A similar semiglobal result has been obtained in the context of averaging if, using the notation of section 5.2, $Y(t, x)$ vanishes identically and t is replaced by γt in (40); see [23].

Proof. The proof of Theorem 5.3, in particular, the proof of Condition 2bis given in Appendix B.3, reveals that for each $T \in [0, \infty)$ satisfying $\{(t, t_0, x_0) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n :$

[4] Convergence results are available in the literature on highly oscillatory systems but, to the best of our knowledge, none of these results yields exactly the required convergence property Condition 2bis.

$t \in [t_0,\, t_0+T],\ \rho(x_0)=1\} \subset \mathrm{Dom}\,\psi$, for each $d \in (0,\,\infty)$, there exists $\beta \in (0,\,\infty)$ such that for all $t_0 \in \mathbb{R}$ and $x_0 \in \mathbb{R}^n$ with $\rho(x_0)=1$ and for all $\varepsilon \in (0,\ \sqrt[\tau]{\gamma\beta}]$

$$\begin{cases} \zeta^\varepsilon(t,t_0,x_0) \text{ exists} & \forall t \in [t_0,\, t_0+T], \\ \rho(\zeta^\varepsilon(t,t_0,x_0)-\psi(t,t_0.x_0)) < d & \forall t \in [t_0,\, t_0+T], \end{cases} \tag{49}$$

with ζ^ε and ψ, respectively, the flow of (48) and (47), and where we have replaced the Euclidean norm by the r-homogeneous norm ρ in the second expression of (49)—see Remark 2. Following the manipulations of section 4, this may be restated as follows: for each $T \in [0,\,\infty)$ satisfying $\{(t,t_0,x_0) \in \mathbb{R}\times\mathbb{R}\times\mathbb{R}^n : t \in [t_0,\, t_0+T],\ \rho(x_0) = 1\} \subset \mathrm{Dom}\,\psi$, for each $d \in (0,\,\infty)$, there exists $\beta \in (0,\,\infty)$ such that for all $t_0 \in \mathbb{R}$ and $x_0 \in \mathbb{R}^n$ with $0 < \rho(x_0) \le \sqrt[\tau]{\gamma\beta}$

$$\begin{cases} \phi(t,t_0,x_0) \text{ exists} & \forall t \in [t_0,\, t_0+\frac{T}{\rho(x_0)^\tau}], \\ \rho(\phi(t,t_0,x_0)-\psi(t,t_0,x_0)) < \rho(x_0)d & \forall t \in [t_0,\, t_0+\frac{T}{\rho(x_0)^\tau}], \end{cases} \tag{50}$$

where ϕ is the flow of (45). In particular, there corresponds a β^* to the particular values of T and d associated with $\dot{x} = \frac{1}{2}[X_1,X_2](x)+X_3(x)$ according to Theorem 3.1. It then follows from Theorem 3.1 that the origin of (45) is LUAS with $\{x_0 \in \mathbb{R}^n : \rho(x_0) \le \sqrt[\tau]{\gamma\beta^*}\}$ contained in the region of attraction. The observation that $\sqrt[\tau]{\gamma\beta^*} \to \infty$ as $\gamma \to \infty$ completes the proof. □

As mentioned above, Theorems 5.3 and 5.4 are related to convergence results for trajectories of highly oscillatory systems. Here we have considered the particular case of one generated Lie bracket, but the theory of [27, 12] covers the general case of how to generate any number of (iterated) Lie brackets. We are therefore inclined to believe that, incorporating ideas from [27, 12], the present theory may be generalized to the case where several (iterated) Lie brackets are featuring in the local approximation.

6. Concluding remarks. The present paper has introduced a new method for proving local stability results for ordinary differential equations. The present approach is based on closeness results for trajectories on finite time intervals. This approach may serve as an alternative for other, Lyapunov-based techniques. A distinctive feature of the present approach is that it does not rely on converse Lyapunov theorems.

Instrumental for our approach is a rescaling mechanism. With this rescaling mechanism we avoid complications that would otherwise arise from the possible non-Lipschitz behavior of homogeneous vectorfields near the origin. This rescaling mechanism also reveals the close relationship between stability results and closeness results for trajectories of systems depending on a small parameter.

By means of several applications, we have shown that this approach enables us to recover and extend several existing stability results. Future research may focus on possible generalizations of this approach to the study of differential equations with delay.

Appendix A. Proof of Lemma 2.2. We start with the observation that the dynamical system obtained from (5) by restricting the state space to $\mathbb{R}^n \setminus \{0\}$ has the uniqueness property of solutions by assumption (a) of the lemma—see, for example, [26]. Furthermore, assumption (b) of the lemma implies that the origin is an equilibrium point of (5); that is, $f(t,0) = 0$. It therefore suffices to prove that no solution of (5) can leave the origin or reach the origin in finite time.

First we prove by contradiction that no solution of (5) starting at the origin can leave the origin in forward time. Indeed, if this is not true, then there exist a solution

ζ of (5) and time-instants $t_0 < t_1$ in the domain of ζ such that $\zeta(t_0) = 0$ and $\zeta(t) \neq 0$ for all $t \in (t_0, t_1]$. Let c be defined by

$$c = ||\zeta(t_1)||.$$

We introduce the family of time-functions ζ_λ defined by

$$\zeta_\lambda(t) = \delta^r_{1/\lambda}\zeta(t),$$

with $\lambda > 0$. Associated with these time-functions is the family of real numbers T_λ defined by

$$T_\lambda = \min\{t > t_0 : ||\zeta_\lambda(t)|| = c\}.$$

Based on the continuity of ζ and expression (1) for the dilation map, it is easy to see that T_λ is a well-defined number between t_0 and t_1, that $T_\lambda \downarrow t_0$ as $\lambda \downarrow 0$, and thus that $\max\{||\dot{\zeta}_\lambda(t)|| : t \in [t_0, T_\lambda]\} \to \infty$ as $\lambda \downarrow 0$. Since ζ_λ satisfies the differential equation

$$\begin{aligned}\dot{\zeta}_\lambda(t) &= \delta^r_{1/\lambda}\dot{\zeta}(t)\\ &= \delta^r_{1/\lambda} f(t, \zeta(t))\\ &= \delta^r_{1/\lambda} f(t, \delta^r_\lambda \zeta_\lambda(t)),\end{aligned}$$

we therefore conclude that necessarily

$$\max\{||\delta^r_{1/\lambda} f(t, \delta^r_\lambda x)|| : t \in \mathbb{R},\ ||x|| \leq c\} \to \infty$$

as $\lambda \downarrow 0$. This yields a contradiction with assumption (b) of the lemma.

We have thus proven that no solution starting in the origin can leave the origin in forward time. By means of similar arguments, it may be shown that no solution starting away from the origin can reach the origin in finite time. This concludes the proof of Lemma 2.2.

Appendix B. Convergence analysis of trajectories. Throughout Appendix B, ρ is an r-homogeneous norm that is assumed to be continuously differentiable on $\mathbb{R}^n \setminus \{0\}$—the particular choice is irrelevant by Remark 6. Also, we regard ζ^ε and ψ as functions of t and write $\zeta^\varepsilon(t)$ and $\psi(t)$ instead of $\zeta^\varepsilon(t, t_0, x_0)$ and $\psi(t, t_0, x_0)$ for notational convenience.

B.1. Proof of Condition 2bis: Perturbed systems. Let $T \in [0, \infty)$ be such that $\{(t, t_0, x_0) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n : t \in [t_0, t_0 + T],\ \rho(x_0) = 1\} \subset \operatorname{Dom}\psi$. Consider arbitrary $d \in (0, \infty)$, $t_0 \in \mathbb{R}$, and $x_0 \in \mathbb{R}^n$ with $\rho(x_0) = 1$. Condition 2bis is proven by showing the existence of $\sigma \in (0, \infty)$ independent of t_0 and x_0 such that for all $\varepsilon \in (0, \sigma]$

$$\left\{\begin{array}{ll}\zeta^\varepsilon(t) \text{ exists} & \forall t \in [t_0, t_0 + T],\\ \|\zeta^\varepsilon(t) - \psi(t)\| < d & \forall t \in [t_0, t_0 + T].\end{array}\right. \tag{51}$$

For that purpose, we analyze ζ^ε and ψ on the time-interval $[t_0, t_0 + T]$.

ζ^ε satisfies

$$\zeta^\varepsilon(t) = x_0 + \int_{t_0}^{t} X(s, \zeta^\varepsilon(s))\,\mathrm{d}s + \int_{t_0}^{t} \delta^r_{1/\varepsilon} Y(s, \delta^r_\varepsilon \zeta^\varepsilon(s))\,\mathrm{d}s \tag{52}$$

for all $t \in \operatorname{Dom}\zeta^\varepsilon \cap [t_0,\, t_0 + T]$. ψ satisfies

$$\psi(t) = x_0 + \int_{t_0}^{t} X(s, \psi(s))\,\mathrm{d}s \tag{53}$$

for all $t \in [t_0,\, t_0 + T]$. Subtracting (53) from (52) gives

$$\zeta^\varepsilon(t) - \psi(t) = \int_{t_0}^{t} \{X(s, \zeta^\varepsilon(s)) - X(s, \psi(s))\}\,\mathrm{d}s + \int_{t_0}^{t} \delta^r_{1/\varepsilon} Y(s, \delta^r_\varepsilon \zeta^\varepsilon(s))\,\mathrm{d}s \tag{54}$$

for all $t \in \operatorname{Dom}\zeta^\varepsilon \cap [t_0,\, t_0 + T]$.

A bound on $\zeta^\varepsilon(t) - \psi(t)$ will be obtained from the Gronwall lemma. First introduce two real numbers $0 < c_1 < c_2$ independent of t_0 and x_0 such that

$$0 < c_1 \le \|\psi(t)\| \le c_2 \qquad \forall t \in [t_0,\, t_0 + T]. \tag{55}$$

Assume for the time being the existence of these numbers. Next take $0 < d' < c_1$.

Let $[t_0,\, t_e]$ be the largest time-interval contained in $\operatorname{Dom}\zeta^\varepsilon \cap [t_0,\, t_0 + T]$ satisfying

$$0 < c_1 - d' \le \|\zeta^\varepsilon(t)\| \le c_2 + d' \qquad \forall t \in [t_0,\, t_e]. \tag{56}$$

Notice that in general the time t_e depends on t_0, x_0, and ε. Since $X \in \mathcal{CLB}$, there exists a Lipschitz constant $k \in [0,\, \infty)$ for X with respect to x on the set $\{(t, x) \in \mathbb{R} \times \mathbb{R}^n : c_1 - d' \le \|x\| \le c_2 + d'\}$. Notice that k is independent of t_0 and x_0.

Then by (54)

$$\|\zeta^\varepsilon(t) - \psi(t)\| \le \int_{t_0}^{t} k\|\zeta^\varepsilon(s) - \psi(s)\|\,\mathrm{d}s + \left\| \int_{t_0}^{t} \delta^r_{1/\varepsilon} Y(s, \delta^r_\varepsilon \zeta^\varepsilon(s))\,\mathrm{d}s \right\| \tag{57}$$

for all $t \in [t_0,\, t_e]$. Notice that the second term in the RHS of (57) is a function of $(\varepsilon, t, t_0, x_0)$ since it depends explicitly on ε, t, and t_0 and implicitly on t_0 and x_0 via ζ^ε. By (56) and assumption (b) of subsection 5.1, this term tends to zero as $\varepsilon \downarrow 0$ uniformly with respect to $t \in [t_0,\, t_e]$, $t_0 \in \mathbb{R}$, and $x_0 \in \mathbb{R}^n$ with $\rho(x_0) = 1$.

Hence, by the Gronwall lemma, there exists $\sigma \in (0,\, \infty)$ independent of t_0 and x_0 such that for all $\varepsilon \in (0,\, \sigma]$

$$\|\zeta^\varepsilon(t) - \psi(t)\| < \min\{d, d'\} \qquad \forall t \in [t_0,\, t_e]. \tag{58}$$

We show by contradiction that $t_e = t_0 + T$ if $\varepsilon \in (0,\, \sigma]$: indeed, from the definition of t_e it follows that if $t_e < t_0 + T$, then necessarily $\|\zeta^\varepsilon(t_e)\| = c_1 - d'$ or $\|\zeta^\varepsilon(t_e)\| = c_2 + d'$ and thus $\|\zeta^\varepsilon(t_e) - \psi(t_e)\| \ge d'$ by (55), which contradicts (58). We conclude that for all $\varepsilon \in (0,\, \sigma]$

$$\begin{cases} \zeta^\varepsilon(t) \text{ exists} & \forall t \in [t_0,\, t_0 + T], \\ \|\zeta^\varepsilon(t) - \psi(t)\| < d & \forall t \in [t_0,\, t_0 + T]. \end{cases} \tag{59}$$

The proof is completed by proving the existence of the real numbers c_1 and c_2 introduced above. For this purpose we study the evolution of the r-homogeneous

norm ρ along ψ. By uniqueness of solutions, $\psi(t) \in \mathbb{R}^n \setminus \{0\}$ for all $t \in [t_0,\, t_0 + T]$ and thus $\rho(\psi(\cdot))$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho(\psi(t)) = \alpha(t)\rho(\psi(t)) \tag{60}$$

for all $t \in [t_0,\, t_0 + T]$, where $\alpha : [t_0,\, t_0 + T] \to \mathbb{R}$ is a continuous function given by

$$\alpha(t) = \sum_{i=1}^{n} \frac{\partial \rho}{\partial x_i}(\delta^r_{1/\rho(\psi(t))}\psi(t)) \frac{1}{\rho(\psi(t))^{r_i}} X_i(t, \delta^r_{\rho(\psi(t))}\delta^r_{1/\rho(\psi(t))}\psi(t)). \tag{61}$$

By homogeneity of X, (61) may be simplified:

$$\alpha(t) = \sum_{i=1}^{n} \frac{\partial \rho}{\partial x_i}(\delta^r_{1/\rho(\psi(t))}\psi(t)) X_i(t, \delta^r_{1/\rho(\psi(t))}\psi(t)). \tag{62}$$

Observe that $\delta^r_{1/\rho(\psi(t))}\psi(t)$ belongs to the compact set $\{x \in \mathbb{R}^n : \rho(x) = 1\}$. Since $\frac{\partial \rho}{\partial x_i}$ is assumed to be continuous on $\mathbb{R}^n \setminus \{0\}$ and since X is assumed to be a class-$\mathcal{CLB}$ function, we conclude the existence of $M \in [0,\, \infty)$ independent of t_0 and x_0 such that $|\alpha(t)| \leq M$ for all $t \in [t_0,\, t_0 + T]$. Together with (60) this implies that for all $t \in [t_0,\, t_0 + T]$

$$0 < e^{-MT} \leq e^{-M(t-t_0)} \leq \rho(\psi(t)) \leq e^{M(t-t_0)} \leq e^{MT}. \tag{63}$$

Passing to the Euclidean norm $\|\cdot\|$—see Remark 2—we conclude the existence of c_1 and c_2 as required.

B.2. Proof of Condition 2bis: Fast time-varying systems. The proof is along the lines of the previous proof in Appendix B.1. Let $T \in [0,\, \infty)$ be such that $\{(t, t_0, x_0) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n : t \in [t_0,\, t_0 + T],\ \rho(x_0) = 1\} \subset \operatorname{Dom} \psi$. Consider arbitrary $d \in (0,\, \infty)$, $t_0 \in \mathbb{R}$, and $x_0 \in \mathbb{R}^n$ with $\rho(x_0) = 1$.

ζ^ε satisfies

$$\zeta^\varepsilon(t) = x_0 + \int_{t_0}^{t} X\left(\frac{s}{\varepsilon^\tau}, \zeta^\varepsilon(s)\right) \mathrm{d}s + \int_{t_0}^{t} \frac{1}{\varepsilon^\tau}\delta^r_{1/\varepsilon} Y\left(\frac{s}{\varepsilon^\tau}, \delta^r_\varepsilon \zeta^\varepsilon(s)\right) \mathrm{d}s \tag{64}$$

for all $t \in \operatorname{Dom} \zeta^\varepsilon \cap [t_0,\, t_0 + T]$. ψ satisfies

$$\psi(t) = x_0 + \int_{t_0}^{t} X_{\mathrm{av}}(\psi(s))\, \mathrm{d}s \tag{65}$$

for all $t \in [t_0,\, t_0 + T]$. Subtracting (65) from (64) and adding terms that cancel out gives

(66)

$$\begin{aligned}
\zeta^\varepsilon(t) - \psi(t) &= \int_{t_0}^{t} X\left(\frac{s}{\varepsilon^\tau}, \zeta^\varepsilon(s)\right) \mathrm{d}s - \int_{t_0}^{t} X_{\mathrm{av}}(\psi(s))\, \mathrm{d}s + \int_{t_0}^{t} \frac{1}{\varepsilon^\tau}\delta^r_{1/\varepsilon} Y\left(\frac{s}{\varepsilon^\tau}, \delta^r_\varepsilon \zeta^\varepsilon(s)\right) \mathrm{d}s \\
&= \int_{t_0}^{t} \left\{ X\left(\frac{s}{\varepsilon^\tau}, \zeta^\varepsilon(s)\right) - X\left(\frac{s}{\varepsilon^\tau}, \psi(s)\right) \right\} \mathrm{d}s \\
&\quad + \int_{t_0}^{t} \left\{ X\left(\frac{s}{\varepsilon^\tau}, \psi(s)\right) - X_{\mathrm{av}}(\psi(s)) \right\} \mathrm{d}s + \int_{t_0}^{t} \frac{1}{\varepsilon^\tau}\delta^r_{1/\varepsilon} Y\left(\frac{s}{\varepsilon^\tau}, \delta^r_\varepsilon \zeta^\varepsilon(s)\right) \mathrm{d}s
\end{aligned}$$

for all $t \in \operatorname{Dom} \zeta^\varepsilon \cap [t_0,\, t_0 + T]$.

A bound on $\zeta^\varepsilon(t) - \psi(t)$ will be obtained from the Gronwall lemma. First introduce two real numbers $0 < c_1 < c_2$ independent of t_0 and x_0 such that

$$0 < c_1 \leq \|\psi(t)\| \leq c_2 \qquad \forall t \in [t_0,\, t_0 + T]. \tag{67}$$

These numbers indeed exist since, by time-invariance of $\dot{x} = X_{\mathrm{av}}(x)$ and continuity of its flow, all states that are reached from initial states x_0 with $\rho(x_0) = 1$ in times $t \in [t_0,\, t_0 + T]$ form a compact set which, by uniqueness of solutions, does not contain the origin. Next take $0 < d' < c_1$.

Let $[t_0,\, t_e]$ be the largest time-interval contained in $\mathrm{Dom}\,\zeta^\varepsilon \cap [t_0,\, t_0 + T]$ satisfying

$$0 < c_1 - d' \leq \|\zeta^\varepsilon(t)\| \leq c_2 + d' \qquad \forall t \in [t_0,\, t_e]. \tag{68}$$

Since $X \in \mathcal{CLB}$, there exists a Lipschitz constant $k \in [0,\, \infty)$ for X with respect to x on the set $\{(t, x) \in \mathbb{R} \times \mathbb{R}^n : c_1 - d' \leq \|x\| \leq c_2 + d'\}$ and an upper bound $M \in [0,\, \infty)$ for $\|X\|$ on the set $\{(t, x) \in \mathbb{R} \times \mathbb{R}^n : c_1 - d' \leq \|x\| \leq c_2 + d'\}$. Notice that both k and M are independent of t_0 and x_0. It follows from the definition of X_{av} that k (respectively, M) is also a Lipschitz constant for X_{av} (respectively, an upper bound for $\|X_{\mathrm{av}}\|$) on the set $\{x \in \mathbb{R}^n : c_1 - d' \leq \|x\| \leq c_2 + d'\}$.

Then by (66)

$$\begin{aligned} \|\zeta^\varepsilon(t) - \psi(t)\| &\leq \int_{t_0}^{t} k\|\zeta^\varepsilon(s) - \psi(s)\|\,\mathrm{d}s \\ &+ \left\| \int_{t_0}^{t} \left\{ X\left(\frac{s}{\varepsilon^\tau}, \psi(s)\right) - X_{\mathrm{av}}(\psi(s)) \right\} \mathrm{d}s \right\| + \left\| \int_{t_0}^{t} \frac{1}{\varepsilon^\tau} \delta^r_{1/\varepsilon} Y\left(\frac{s}{\varepsilon^\tau}, \delta^r_\varepsilon \zeta^\varepsilon(s)\right) \mathrm{d}s \right\| \end{aligned} \tag{69}$$

for all $t \in [t_0,\, t_e]$. We now show that the second and the third terms in the RHS of this expression tend to zero as $\varepsilon \downarrow 0$. By (68) and assumption (b) of subsection 5.2

$$\left\| \int_{t_0}^{t} \frac{1}{\varepsilon^\tau} \delta^r_{1/\varepsilon} Y\left(\frac{s}{\varepsilon^\tau}, \delta^r_\varepsilon \zeta^\varepsilon(s)\right) \mathrm{d}s \right\| \to 0 \tag{70}$$

as $\varepsilon \downarrow 0$ uniformly with respect to $t \in [t_0,\, t_e]$, $t_0 \in \mathbb{R}$, and $x_0 \in \mathbb{R}^n$ with $\rho(x_0) = 1$. Bounding the second term using assumption (c) of subsection 5.2 is more complicated, since $\psi(s)$ featuring in the integrand of the second term varies with time s, whereas x featuring in the integrand of (42) is fixed. This problem may be overcome by sampling the trajectory ψ with sample period $\theta \in (0,\, \infty)$: define ψ_{sa} by

$$\psi_{\mathrm{sa}}(t) = \psi(t_0 + i\theta) \quad \forall t \in [t_0 + i\theta,\, t_0 + (i+1)\theta) \cap [t_0,\, t_0 + T] \quad \forall i \in \{0\} \cup \mathbb{N}. \tag{71}$$

Notice that $\psi_{\mathrm{sa}}(t)$ is related to $\psi(t)$ by the inequality

$$\|\psi_{\mathrm{sa}}(t) - \psi(t)\| \leq M\theta \tag{72}$$

for all $t \in [t_0,\, t_0 + T]$. Then

$$\begin{aligned} &\left\| \int_{t_0}^{t} \left\{ X\left(\frac{s}{\varepsilon^\tau}, \psi(s)\right) - X_{\mathrm{av}}(\psi(s)) \right\} \mathrm{d}s \right\| \leq \left\| \int_{t_0}^{t} \left\{ X\left(\frac{s}{\varepsilon^\tau}, \psi(s)\right) - X\left(\frac{s}{\varepsilon^\tau}, \psi_{\mathrm{sa}}(s)\right) \right\} \mathrm{d}s \right\| \\ &+ \left\| \int_{t_0}^{t} \left\{ X\left(\frac{s}{\varepsilon^\tau}, \psi_{\mathrm{sa}}(s)\right) - X_{\mathrm{av}}(\psi_{\mathrm{sa}}(s)) \right\} \mathrm{d}s \right\| + \left\| \int_{t_0}^{t} \{ X_{\mathrm{av}}(\psi_{\mathrm{sa}}(s)) - X_{\mathrm{av}}(\psi(s)) \}\, \mathrm{d}s \right\|. \end{aligned} \tag{73}$$

By (72) the first and the third terms in the RHS of (73) are both bounded by $TkM\theta$ and thus can be made as small as required by choosing the sample period θ sufficiently small. The second term in the RHS of (73) may be bounded as follows:

$$\begin{aligned}(74)\quad &\left\| \int_{t_0}^{t} \left\{ X\left(\frac{s}{\varepsilon^\tau}, \psi_{\mathrm{sa}}(s)\right) - X_{\mathrm{av}}(\psi_{\mathrm{sa}}(s)) \right\} \mathrm{d}s \right\| \\ &\qquad \le \left\| \int_{t_0}^{t_0+\theta} \left\{ X\left(\frac{s}{\varepsilon^\tau}, \psi_{\mathrm{sa}}(s)\right) - X_{\mathrm{av}}(\psi_{\mathrm{sa}}(s)) \right\} \mathrm{d}s \right\| \\ &\qquad + \left\| \int_{t_0+\theta}^{t_0+2\theta} \left\{ X\left(\frac{s}{\varepsilon^\tau}, \psi_{\mathrm{sa}}(s)\right) - X_{\mathrm{av}}(\psi_{\mathrm{sa}}(s)) \right\} \mathrm{d}s \right\| \\ &\qquad\qquad + \cdots + \left\| \int_{t_0+p\theta}^{t} \left\{ X\left(\frac{s}{\varepsilon^\tau}, \psi_{\mathrm{sa}}(s)\right) - X_{\mathrm{av}}(\psi_{\mathrm{sa}}(s)) \right\} \mathrm{d}s \right\|\end{aligned}$$

with $p \in \{0\} \cup \mathbb{N}$ defined by $t_0 + p\theta \le t < t_0 + (p+1)\theta$. For a given choice of θ the number of terms in the RHS of (74) is bounded for $t \in [t_0, t_0 + T]$, and thus by (67), (71), and assumption (c) of subsection 5.2 the RHS of (74) tends to zero as $\varepsilon \downarrow 0$ uniformly with respect to $t \in [t_0, t_0 + T]$, $t_0 \in \mathbb{R}$, and $x_0 \in \mathbb{R}^n$ with $\rho(x_0) = 1$. We finally conclude that

$$(75)\qquad \left\| \int_{t_0}^{t} \left\{ X\left(\frac{s}{\varepsilon^\tau}, \psi(s)\right) - X_{\mathrm{av}}(\psi(s)) \right\} \mathrm{d}s \right\| \to 0$$

as $\varepsilon \downarrow 0$ uniformly with respect to $t \in [t_0, t_0 + T]$, $t_0 \in \mathbb{R}$, and $x_0 \in \mathbb{R}^n$ with $\rho(x_0) = 1$.

Recall (69), (70), and (75). Hence, by the Gronwall lemma, there exists $\sigma \in (0, \infty)$ independent of t_0 and x_0 such that for all $\varepsilon \in (0, \sigma]$

$$(76)\qquad \|\zeta^\varepsilon(t) - \psi(t)\| < \min\{d, d'\} \qquad \forall t \in [t_0, t_e].$$

As in Appendix B.1, we conclude that for all $\varepsilon \in (0, \sigma]$

$$(77)\qquad \begin{cases} \zeta^\varepsilon(t) \text{ exists} & \forall t \in [t_0, t_0 + T], \\ \|\zeta^\varepsilon(t) - \psi(t)\| < d & \forall t \in [t_0, t_0 + T]. \end{cases}$$

B.3. Proof of Condition 2bis: Highly oscillatory systems. The proof is along the lines of the proofs in Appendices B.1 and B.2. Let $T \in [0, \infty)$ be such that $\{(t, t_0, x_0) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n : t \in [t_0, t_0 + T],\ \rho(x_0) = 1\} \subset \mathrm{Dom}\,\psi$. Consider arbitrary $d \in (0, \infty)$, $t_0 \in \mathbb{R}$, and $x_0 \in \mathbb{R}^n$ with $\rho(x_0) = 1$.

ζ^ε satisfies

$$\begin{aligned}(78)\quad \zeta^\varepsilon(t) = x_0 &+ \int_{t_0}^{t} \frac{1}{\sqrt{\mu}} \cos\left(\frac{s}{\mu}\right) X_1(\zeta^\varepsilon(s))\,\mathrm{d}s \\ &+ \int_{t_0}^{t} \frac{1}{\sqrt{\mu}} \sin\left(\frac{s}{\mu}\right) X_2(\zeta^\varepsilon(s))\,\mathrm{d}s + \int_{t_0}^{t} X_3(\zeta^\varepsilon(s))\,\mathrm{d}s\end{aligned}$$

for all $t \in \mathrm{Dom}\,\zeta^\varepsilon \cap [t_0, t_0 + T]$, where we have introduced the notation $\mu = \frac{\varepsilon^\tau}{\gamma}$.

Integrating the second and the third terms in the RHS of (78) by parts yields

(79)
$$\begin{aligned}\zeta^\varepsilon(t) &= x_0 + \sqrt{\mu}\sin\left(\frac{s}{\mu}\right) X_1(\zeta^\varepsilon(s))|_{s=t_0}^t \\ &\quad - \int_{t_0}^t \sqrt{\mu}\sin\left(\frac{s}{\mu}\right) DX_1\cdot\left\{\frac{1}{\sqrt{\mu}}\cos\left(\frac{s}{\mu}\right)X_1 + \frac{1}{\sqrt{\mu}}\sin\left(\frac{s}{\mu}\right)X_2 + X_3\right\}(\zeta^\varepsilon(s))\,\mathrm{d}s \\ &\quad - \sqrt{\mu}\cos\left(\frac{s}{\mu}\right) X_2(\zeta^\varepsilon(s))|_{s=t_0}^t \\ &\quad + \int_{t_0}^t \sqrt{\mu}\cos\left(\frac{s}{\mu}\right) DX_2\cdot\left\{\frac{1}{\sqrt{\mu}}\cos\left(\frac{s}{\mu}\right)X_1 + \frac{1}{\sqrt{\mu}}\sin\left(\frac{s}{\mu}\right)X_2 + X_3\right\}(\zeta^\varepsilon(s))\,\mathrm{d}s \\ &\quad + \int_{t_0}^t X_3(\zeta^\varepsilon(s))\,\mathrm{d}s.\end{aligned}$$

Applying the geometric identities $\cos^2(\phi) = \frac{1}{2} + \frac{1}{2}\cos(2\phi)$, $\sin^2(\phi) = \frac{1}{2} - \frac{1}{2}\cos(2\phi)$, and $\sin(\phi)\cos(\phi) = \frac{1}{2}\sin(2\phi)$ and rearranging terms gives

$$\zeta^\varepsilon(t) = x_0 + \int_{t_0}^t \left(\frac{1}{2}DX_2\cdot X_1 - \frac{1}{2}DX_1\cdot X_2 + X_3\right)(\zeta^\varepsilon(s))\,\mathrm{d}s + J_1 + J_2 + J_3 \tag{80}$$

with

$$\begin{aligned}J_1 &= \int_{t_0}^t \frac{1}{2}\sin\left(2\frac{s}{\mu}\right)\{DX_2\cdot X_2 - DX_1\cdot X_1\}(\zeta^\varepsilon(s))\,\mathrm{d}s, \\ J_2 &= \int_{t_0}^t \frac{1}{2}\cos\left(2\frac{s}{\mu}\right)\{DX_1\cdot X_2 + DX_2\cdot X_1\}(\zeta^\varepsilon(s))\,\mathrm{d}s, \\ J_3 &= \sqrt{\mu}\bigg(\int_{t_0}^t \left\{\cos\left(\frac{s}{\mu}\right)DX_2\cdot X_3(\zeta^\varepsilon(s)) - \sin\left(\frac{s}{\mu}\right)DX_1\cdot X_3(\zeta^\varepsilon(s))\right\}\mathrm{d}s \\ &\qquad + \sin\left(\frac{s}{\mu}\right)X_1(\zeta^\varepsilon(s))|_{s=t_0}^t - \cos\left(\frac{s}{\mu}\right)X_2(\zeta^\varepsilon(s))|_{s=t_0}^t\bigg).\end{aligned}$$

ψ satisfies

$$\begin{aligned}\psi(t) &= x_0 + \int_{t_0}^t \left(\frac{1}{2}[X_1, X_2] + X_3\right)(\psi(s))\,\mathrm{d}s \\ &= x_0 + \int_{t_0}^t \left(\frac{1}{2}DX_2\cdot X_1 - \frac{1}{2}DX_1\cdot X_2 + X_3\right)(\psi(s))\,\mathrm{d}s\end{aligned} \tag{81}$$

for all $t \in [t_0,\, t_0 + T]$, where we used the definition of $[X_1, X_2]$ taking into account that $\psi(s) \in \mathbb{R}^n \setminus \{0\}$ by the uniqueness property of solutions. Subtracting (81) from (80) gives

$$\begin{aligned}\zeta^\varepsilon(t) - \psi(t) = \int_{t_0}^t \bigg\{ &\left(\frac{1}{2}DX_2\cdot X_1 - \frac{1}{2}DX_1\cdot X_2 + X_3\right)(\zeta^\varepsilon(s)) \\ &- \left(\frac{1}{2}DX_2\cdot X_1 - \frac{1}{2}DX_1\cdot X_2 + X_3\right)(\psi(s))\bigg\}\,\mathrm{d}s + J_1 + J_2 + J_3\end{aligned} \tag{82}$$

for all $t \in \operatorname{Dom}\zeta^\varepsilon \cap [t_0,\, t_0 + T]$.

A bound on $\zeta^\varepsilon(t)-\psi(t)$ will be obtained from the Gronwall lemma. First introduce two real numbers $0 < c_1 < c_2$ independent of t_0 and x_0 such that

$$0 < c_1 \le \|\psi(t)\| \le c_2 \qquad \forall t \in [t_0,\, t_0+T]. \tag{83}$$

These numbers indeed exist since, by time-invariance of $\dot{x} = \frac{1}{2}[X_1, X_2](x)+X_3(x)$ and continuity of its flow, all states that are reached from initial states x_0 with $\rho(x_0) = 1$ in times $t \in [t_0,\, t_0+T]$ form a compact set which, by uniqueness of solutions, does not contain the origin. Next take $0 < d' < c_1$.

Let $[t_0,\, t_e]$ be the largest time-interval contained in $\operatorname{Dom}\zeta^\varepsilon \cap [t_0,\, t_0+T]$ satisfying

$$0 < c_1 - d' \le \|\zeta^\varepsilon(t)\| \le c_2 + d' \qquad \forall t \in [t_0,\, t_e]. \tag{84}$$

By the smoothness assumptions on the X_i, there exists a Lipschitz constant $k \in [0,\, \infty)$ for $\frac{1}{2}DX_2\cdot X_1 - \frac{1}{2}DX_1\cdot X_2 + X_3$ on the set $\{x \in \mathbb{R}^n : c_1 - d' \le \|x\| \le c_2 + d'\}$. Notice that k is independent of t_0 and x_0.

Then by (82)

$$\|\zeta^\varepsilon(t) - \psi(t)\| \le \int_{t_0}^{t} k\|\zeta^\varepsilon(s) - \psi(s)\|\,\mathrm{d}s + \|J_1\| + \|J_2\| + \|J_3\| \tag{85}$$

for all $t \in [t_0,\, t_e]$. Notice that the J_i are functions of (μ, t, t_0, x_0) since J_i depends explicitly on μ, t, and t_0 and implicitly on t_0 and x_0 via $\zeta^\varepsilon(s)$. We now show that the $\|J_i\|$ converge to zero as $\mu \downarrow 0$, uniformly with respect to $t \in [t_0,\, t_e]$, $t_0 \in \mathbb{R}$, and $x_0 \in \mathbb{R}^n$ with $\rho(x_0) = 1$. This is easily verified for $\|J_3\|$: J_3 is the product of $\sqrt{\mu}$ with a factor that is bounded on $\{(\mu, t, t_0, x_0) : t \in [t_0,\, t_e],\ \rho(x_0) = 1\}$ by the smoothness assumptions on the X_i. Next we focus on J_1: integration by parts yields

$$\begin{aligned}
J_1 &= -\frac{\mu}{4}\cos\left(2\frac{s}{\mu}\right)\{DX_2\cdot X_2 - DX_1\cdot X_1\}(\zeta^\varepsilon(s))|_{s=t_0}^{t} \\
&\quad + \int_{t_0}^{t} \frac{\mu}{4}\cos\left(2\frac{s}{\mu}\right) D\{DX_2\cdot X_2 - DX_1\cdot X_1\}\cdot \\
&\qquad\qquad \left\{\frac{1}{\sqrt{\mu}}\cos\left(\frac{s}{\mu}\right)X_1 + \frac{1}{\sqrt{\mu}}\sin\left(\frac{s}{\mu}\right)X_2 + X_3\right\}(\zeta^\varepsilon(s))\,\mathrm{d}s \\
&= \mu\bigg(-\frac{1}{4}\cos\left(2\frac{s}{\mu}\right)\{DX_2\cdot X_2 - DX_1\cdot X_1\}(\zeta^\varepsilon(s))|_{s=t_0}^{t} \\
&\qquad + \int_{t_0}^{t}\frac{1}{4}\cos\left(2\frac{s}{\mu}\right) D\{DX_2\cdot X_2 - DX_1\cdot X_1\}\cdot X_3(\zeta^\varepsilon(s))\,\mathrm{d}s\bigg) \\
&\quad + \sqrt{\mu}\bigg(\int_{t_0}^{t}\frac{1}{4}\cos\left(2\frac{s}{\mu}\right) D\{DX_2\cdot X_2 - DX_1\cdot X_1\} \\
&\qquad\qquad \cdot\left\{\cos\left(\frac{s}{\mu}\right)X_1 + \sin\left(\frac{s}{\mu}\right)X_2\right\}(\zeta^\varepsilon(s))\,\mathrm{d}s\bigg).
\end{aligned} \tag{86}$$

Both factors between brackets in the second RHS of (86) are bounded on the set $\{(\mu, t, t_0, x_0) : t \in [t_0,\, t_e],\ \rho(x_0) = 1\}$ by the smoothness assumptions on the X_i. We thus see that $\|J_1\|$ also converges to zero as $\mu \downarrow 0$, uniformly with respect to $t \in [t_0,\, t_e]$, $t_0 \in \mathbb{R}$, and $x_0 \in \mathbb{R}^n$ with $\rho(x_0) = 1$. Finally, a similar argument shows that $\|J_2\|$ also converges to zero as $\mu \downarrow 0$, uniformly with respect to $t \in [t_0,\, t_e]$, $t_0 \in \mathbb{R}$, and $x_0 \in \mathbb{R}^n$ with $\rho(x_0) = 1$.

Hence, by the Gronwall lemma, there exists $\beta \in (0, \infty)$ independent of t_0 and x_0 such that for all $\mu \in (0, \beta]$

$$\|\zeta^\varepsilon(t) - \psi(t)\| < \min\{d, d'\} \qquad \forall t \in [t_0, t_e]. \tag{87}$$

As in Appendix B.1, we conclude that for all $\mu \in (0, \beta]$ or, equivalently since $\mu = \frac{\varepsilon^\tau}{\gamma}$ for all $\varepsilon \in (0, \sqrt[\tau]{\gamma\beta}]$,

$$\begin{cases} \zeta^\varepsilon(t) \text{ exists} & \forall t \in [t_0, t_0 + T], \\ \|\zeta^\varepsilon(t) - \psi(t)\| < d & \forall t \in [t_0, t_0 + T]. \end{cases} \tag{88}$$

Acknowledgment. The authors would like to thank the anonymous reviewers for their constructive comments that have helped to improve the paper. In particular we thank one of the reviewers for providing an alternative proof of Lemma 2.2, which has been incorporated in the present version of the paper.

REFERENCES

[1] D. Aeyels and J. Peuteman, *On exponential stability of nonlinear time-varying differential equations*, Automatica J. IFAC, 35 (1999), pp. 1091–1100.

[2] L. Barreira and Y. B. Pesin, *Lyapunov Exponents and Smooth Ergodic Theory*, Univ. Lecture Ser. 23, AMS, Providence, RI, 2002.

[3] S.-N. Chow and J. Mallet-Paret, *Integral averaging and bifurcation*, J. Differential Equations, 26 (1977), pp. 112–159.

[4] M. M. Hapaev, *Averaging in Stability Theory: A Study of Resonance Multi-Frequency Systems*, Math. Appl. 79, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993.

[5] P. Hartman, *Ordinary Differential Equations*, 2nd ed., Birkhäuser Boston, Cambridge, MA, 1982.

[6] H. Hermes, *Nilpotent and high-order approximations of vector field systems*, SIAM Rev., 33 (1991), pp. 238–264.

[7] H. Hermes, *Asymptotic stabilization via homogeneous approximations*, in Geometry of Feedback and Optimal Control, Monogr. Textbooks Pure Appl. Math. 207, B. Jakubczyk and W. Respondek, eds., Marcel Dekker, New York, 1998, pp. 205–218.

[8] M. Kawski, *Homogeneous stabilizing feedback laws*, Control-Theory and Advanced Technology (C-TAT), 6 (1990), pp. 497–516.

[9] H. Khalil, *Nonlinear Systems*, 2nd ed., Prentice–Hall, Upper Saddle River, NJ, 1996.

[10] P. Kokotović, H. K. Khalil, and J. O'Reilly, *Singular Perturbation Methods in Control: Analysis and Design*, Academic Press, London, 1986.

[11] J. Kurzweil and J. Jarník, *Limit processes in ordinary differential equations*, Z. Angew. Math. Phys., 38 (1987), pp. 241–256.

[12] W. Liu, *An approximation algorithm for nonholonomic systems*, SIAM J. Control Optim., 35 (1997), pp. 1328–1365.

[13] R. M'Closkey and P. Morin, *Time-varying homogeneous feedback: Design tools for the exponential stabilization of systems with drift*, Internat. J. Control, 71 (1998), pp. 837–869.

[14] R. M'Closkey and R. Murray, *Exponential stabilization of driftless nonlinear control systems using homogeneous feedback*, IEEE Trans. Automat. Control, 42 (1997), pp. 614–628.

[15] L. Moreau and D. Aeyels, *Local approximations and stability: A trajectory-based approach*, in Proceedings of the 38th IEEE Conference on Decision and Control (CDC), 1999, pp. 734–739.

[16] L. Moreau and D. Aeyels, *A systematic design tool for asymptotic stabilization of driftless control affine systems*, in Proceedings of the 38th IEEE Conference on Decision and Control (CDC), 1999, pp. 861–862.

[17] L. Moreau and D. Aeyels, *Asymptotic methods in the stability analysis of parametrized homogeneous flows*, Automatica J. IFAC, 36 (2000), pp. 1213–1218.

[18] L. Moreau and D. Aeyels, *Practical stability and stabilization*, IEEE Trans. Automat. Control, 45 (2000), pp. 1554–1558.

[19] L. Moreau, W. Michiels, D. Aeyels, and D. Roose, *Robustness of nonlinear delay equations with respect to input perturbations: A trajectory based approach*, Math. Control Signals Systems, 15 (2002), pp. 316–335.

[20] P. Morin, J.-B. Pomet, and C. Samson, *Design of homogeneous time-varying stabilizing control laws for driftless controllable systems via oscillatory approximation of Lie brackets in closed loop*, SIAM J. Control Optim., 38 (1999), pp. 22–49.

[21] P. Morin and C. Samson, *Time-varying exponential stabilization of a rigid spacecraft with two control torques*, IEEE Trans. Automat. Control, 42 (1997), pp. 528–534.

[22] J. Peuteman and D. Aeyels, *Exponential stability of nonlinear time-varying differential equations and parital averaging*, Math. Control Signals Systems, 15 (2002), pp. 42–70.

[23] J. Peuteman and D. Aeyels, *Averaging results and the study of uniform asymptotic stability of homogeneous differential equations that are not fast time-varying*, SIAM J. Control Optim., 37 (1999), pp. 997–1010.

[24] J. Sanders and F. Verhulst, *Averaging Methods in Nonlinear Dynamical Systems*, Appl. Math. Sci. 59, Springer–Verlag, New York, 1985.

[25] A. V. Sarychev, *Lie- and chronologico-algebraic tools for studying stability of time-varying systems*, Systems Control Lett., 43 (2001), pp. 59–76.

[26] E. Sontag, *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, 2nd ed., Texts Appl. Math. 6, Springer–Verlag, New York, 1998.

[27] H. Sussmann and W. Liu, *Limits of highly oscillatory controls and the approximation of general paths by admissible trajectories*, in Proceedings of the 30th IEEE Conference on Decision and Control, 1991, pp. 437–442.

[28] A. Teel, R. Murray, and G. Walsh, *Nonholonomic control systems: From steering to stabilization with sinusoids*, Internat. J. Control, 62 (1995), pp. 849–870.

[29] A. R. Teel, L. Moreau, and D. Nešić, *A unified framework for input-to-state stability in systems with two time scales*, IEEE Trans. Automat. Control, submitted.

[30] F. Verhulst, *Nonlinear Differential Equations and Dynamical Systems*, 2nd ed., Springer–Verlag, Berlin, 1996.

SIAM J. CONTROL OPTIM.
Vol. 41, No. 6, pp. 1946–1979

A DIFFUSION MODEL FOR OPTIMAL DIVIDEND DISTRIBUTION FOR A COMPANY WITH CONSTRAINTS ON RISK CONTROL*

TAHIR CHOULLI†, MICHAEL TAKSAR‡, AND XUN YU ZHOU§

Abstract. This paper investigates a model of a corporation which faces constant liability payments and which can choose a production/business policy from an available set of control policies with different expected profits and risks. The objective is to find a business policy and a dividend distribution scheme so as to maximize the expected present value of the total dividend distributions. The main feature of this paper is that there are constraints on business activities such as inability to completely eliminate risk (even at the expense of reducing the potential profit to zero) or when such a risk cannot exceed a certain level. The case in which there is no restriction on the dividend pay-out rates is dealt with. This gives rise to a mixed regular-singular stochastic control problem. First the value function is analyzed in great detail and in particular is shown to be a viscosity solution of the corresponding Hamilton–Jacobi–Bellman (HJB) equation. Based on this it is further proved that the value function must be twice continuously differentiable. Then a delicate analysis is carried out on the HJB equation, leading to an explicit expression of the value function as well as the optimal policies.

Key words. diffusion model, dividend distribution, risk control, optimal stochastic control, HJB equation, viscosity solution, Skorohod problem

AMS subject classifications. 91B70, 93E20

PII. S0363012900382667

1. Introduction.

Recently there has been an upsurge of interest in diffusion models for optimal dividend optimization and/or risk control techniques (see Jeanblanc–Piqué and Shiryaev [11], Asmussen and Taksar [2], Radner and Shepp [16], Boyle, Elliott, and Yang [3], Højgaard and Taksar [8], [9], [10], Paulsen and Gjessing [13], and Taksar and Zhou [18]). In those models the liquid assets of the company are modeled by a Brownian motion with constant drift and diffusion coefficients. The drift term corresponds to the expected (potential) profit per unit time, while the diffusion term is interpreted as risk. The larger the diffusion coefficient the greater the business risk the company takes on. If the company wants to decrease the risk from its business activities, it also faces a decrease in its potential profit. In other words, different business activities in this model correspond to changing *simultaneously* the drift and the diffusion coefficients of the underlying process. This sets a scene for an optimal stochastic control model where the controls affect not only the drift but also the diffusion part of the dynamic of the system.

Another important feature of our paper is dividend distribution. Dividends are paid from the liquid reserve of the company and distributed to the shareholders.

*Received by the editors December 19, 2000; accepted for publication (in revised form) September 7, 2002; published electronically March 13, 2003.

http://www.siam.org/journals/sicon/41-6/38266.html

†Mathematical and Statistical Sciences Department, University of Alberta, Edmonton, AB, T6G2G1 Canada. This author wishes to gratefully acknowledge the financial support and hospitality of the Department of Systems Engineering and Engineering Management at the Chinese University of Hong Kong—where the main part of this work was done—and the Pacific Institute for Mathematical Sciences.

‡Department of Mathematics, University of Missouri, Columbia, MO 65211 (taksar@math.missouri.edu). This author is supported by the National Science Foundation grant DMS 9705011.

§Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong (xyzhou@se.cuhk.edu.hk). This author is supported by the RGC earmarked grant CUHK 4054/98E.

In the control model the dividend distribution plan is represented by an increasing functional, C_t, whose meaning is the cumulative amount of dividends paid out up to time t. In this paper the dividend pay-out rate is unbounded which, together with the risk control part, leads to a mixed regular-singular stochastic control model. The risk control/dividend distribution policy determines uniquely the dynamics of the liquid reserve. The company is bankrupt when its liquid assets vanish. The objective is to find the policy which maximizes the expected cumulative discounted dividend pay-outs up to the time of the bankruptcy.

Insurance is one of the natural areas where those models become widely applied. The risk control in insurance takes on the form of reinsurance. Specifically, if at any fixed time both the drift and diffusion coefficients of the controlled stochastic process are multiples of one and the same control parameter a, $0 \leq a \leq 1$, then this would be the case of the so-called *proportional reinsurance*, which is employed by a *cedent* in order to reduce the insurance risks. Other types of reinsurance schemes result in different types of drift/diffusion control models (see, e.g., [1], [17], [4]).

In this paper we consider a company whose business activities are modeled by a control process a_t, $t \geq 0$, which takes on values in the interval $[\alpha, \beta]$, $0 < \alpha < \beta < +\infty$, with risk and potential profit at any time t proportional to a_t. The restriction $\alpha > 0$ reflects the fact that there are institutional or statutory reasons (e.g., the company is public) that its business activities cannot be reduced to zero, unless the company faces bankruptcy. In addition, in our model the company has a constant rate of liability payments, such as mortgage payments on its property or amortization of bonds. In the case of an insurance company, when the control parameter a_t lies within $[0, 1]$ this problem was considered by Taksar and Zhou [18]. In this regard, the model treated in [18] can be viewed as a limiting case of $\alpha \to 0+$ and $\beta = 1$. However, the strictly positive lower bound treated in this paper renders the argument in [18] invalid and imposes a great difficulty for the problem. It is interesting to observe that the analytic expression for the optimal return function (value function) obtained in this paper, in the limiting case of $\alpha \to 0+$, $\beta = 1$, looks completely different from that in [18]. However, we will show via a detailed analysis that these are two analytic expressions for one and the same function.

We start our analysis with the value function v of the underlying stochastic control model. We first show that v is a viscosity solution of the corresponding Hamilton–Jacobi–Bellman (HJB) equation, which is interesting in its own right as the underlying stochastic control model is of a mixed regular-singular type. Based on this fact, along with the concavity of the value function, we prove a priori that v must be twice continuously differentiable (and hence, as a by-product, must be a classical solution to the HJB equation). The proof is very general and should be applicable to a large set of problems whenever concavity can be proved in advance. Afterwards we perform a delicate analysis on the HJB equation, which leads to explicit expressions of the value function for all the possible parameter values. Once this is done, optimal risk control and dividend policies are constructed via the verification theorem and the solution to a Skorohod problem.

The paper is structured as follows. In the next section we give a rigorous mathematical formulation of the problem and analyze the structure of the value function. We show that it is concave and is twice continuously differentiable. In section 3 we analyze the case without liability, which is interesting in its own right and inspiring in treating more general cases. In section 4 we extend the results to the case of a constant liability payments. Section 5 is devoted to the construction of optimal policies

based on the results of the preceding sections. The last section is devoted to economic interpretation of the obtained results, along with some concluding remarks.

2. Properties of the optimal return function. We start with a filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$ and a one-dimensional standard Brownian motion W_t (with $W_0 = 0$) on it, adapted to the filtration $\mathcal{F}_t$. We denote by R_t^π the reserve of the company at time t under a control policy $\pi = (a_t^\pi, C_t^\pi; t \geq 0)$ (to be specified below). The dynamics of the reserve process R_t^π is described by

$$dR_t^\pi = (a_t^\pi \mu - \delta)dt + a_t^\pi \sigma dW_t - dC_t^\pi, \tag{2.1}$$

with initial condition

$$R_{0-}^\pi = x, \tag{2.2}$$

where μ is the expected profit per unit time (profit rate) and σ is the volatility rate of the reserve process in the absence of any risk control, δ represents the amount of money the company has to pay per unit time (the debt rate) irrespective of what business activities it chooses, and x is the initial reserve. It should be noted that a dividend distribution (see below) may take place at the initial time, hence the notation R_{0-}^π represents the initial reserve level *before* such a distribution has ever occurred.

The control in this model is described by a pair of $\mathcal{F}_t$-adapted measurable processes $\pi = (a_t^\pi, C_t^\pi; t \geq 0)$. A control $\pi = (a_t^\pi, C_t^\pi; t \geq 0)$ is admissible if $\alpha \leq a_t^\pi \leq \beta$ for all $t \geq 0$, and C_t^π is nondecreasing, right continuous having left limits process, where $0 < \alpha < \beta < +\infty$ are given scalars. In addition it is required that the state process R_t^π is nonnegative for all $t \leq \tau$, where τ is the time of bankruptcy given by (2.3) below. The last condition describes the requirement that one cannot have a negative reserve even at the time of bankruptcy. Thus if at time t a lump sum payment of dividends is made, then it cannot be in excess of the reserve at hand. We denote the set of all admissible controls by $\mathcal{A}$. The control component a_t^π represents one of the possible business activities available for the company at time t, and the component C_t^π corresponds to the total amount of dividends paid out by the company up to time t. For any admissible C^π and any $t < 0$ we set $C_t^\pi = 0$. Thus $C_{0-}^\pi = 0$ and if $C_0^\pi > 0$, then $R_0 = x - C_0^\pi$. The latter corresponds to the policy π, which pays a lump sum dividend of C_0^π at time 0.

Given a control policy π, the time of bankruptcy is defined as

$$\tau^\pi = \inf\{t \geq 0 : R_t^\pi = 0\}. \tag{2.3}$$

We make the convention that

$$R_t^\pi = 0 \qquad \forall t \geq \tau^\pi. \tag{2.4}$$

The *performance functional* associated with each control π is

$$J_x(\pi) = E\left(\int_0^{\tau^\pi} e^{-\gamma t} dC_t^\pi\right), \tag{2.5}$$

where $\gamma > 0$ is an a priori given discount factor (used in calculating the present value of the future dividends), and the subscript x denotes the initial state in the right-hand side of (2.2). To simplify notation, in what follows we will omit the superscript π in τ^π, a^π, etc. when it is clear from the context which policy we are dealing with.

The integral in (2.5) is understood as an integral on $[0, \tau]$ with respect to the measure whose distribution function is C_t. In particular, if $C_0 > 0$, this measure has an atomic mass of C_0 at 0 and this quantity is included in the integral. Likewise if $C_{\tau-} < C_\tau$, then $e^{-c\tau}(C_\tau - C_{\tau-})$ is included in the integral in the right-hand side of (2.5). The objective is to find the *value function* (also known as *optimal return function*)

$$v(x) = \sup_{\pi \in \mathcal{A}} J_x(\pi) \tag{2.6}$$

and the optimal policy π^* such that

$$J_x(\pi^*) = v(x). \tag{2.7}$$

It is worthwhile to mention that from the requirement of nonnegativity of the reserve one can deduce that C_t does not exceed the running maximum of a Brownian motion with drift coefficient $\beta\mu - \delta$ and diffusion coefficient $\beta\sigma$. From this it is easy to see that for each policy π

$$J_x(\pi) \le \int_0^\infty e^{-\gamma t} \max_{0 \le u \le t} [x + (\beta\mu - \delta)u + \beta\sigma W_u] dt.$$

From the above inequality one can deduce that $v(x)$ is finite. However, more elaborate arguments are not needed since the finiteness of v will be also implied from the results of sections 3 and 4. The exogenous parameters of the problem are $\mu, \sigma, \delta, \alpha, \beta$, and γ. The aim of this paper is to obtain the value function v and the optimal policy *explicitly* in terms of these parameters.

A few remarks on the control component a_t are in order. The way this quantity enters into the dynamics (2.1) clearly shows that it reduces or increases the risk simultaneously reducing or increasing the expected profit rate at the same scale. In other words, the diffusion coefficient of the system (2.1) depends on the control component a_t. In [18], the problem is formulated in the context of an insurance company where $1 - a_t$ signifies the reinsurance fraction and the constraint $0 \le a_t \le 1$ is imposed, which is a limiting case of $\alpha \to 0+$ and $\beta = 1$. (Note that while in our analysis below we require $\alpha > 0$, the solution we obtain does have a limit when $\alpha \to 0+$ and this limit coincides with the solution in [18]. In this sense the model in [18] is indeed a special case of the model presented here.) It is certainly meaningful to relax this constraint to one with any arbitrary upper and lower bounds. For example, for the insurance company case, $\beta > 1$ would mean that the company can take an extra insurance business from other companies (that is, act as a reinsurer for other cedents). Moreover, our formulation can model risk control problems for companies other than insurance ones. On the other hand, the two general bounds α and β add a new, nontrivial feature to this model, as will be evident in what follows.

The main tools for solving the problem are the dynamic programming and the HJB equation (see Fleming and Rishel [6], Fleming and Soner [7], and Yong and Zhou [19], as well as relevant discussions in [2], [9], and [18]). To analyze the value function, we need the following lemma.

LEMMA 1. *Let X_t be Ito's process on a positive half line,*

$$X_t = x + \int_0^t m(u)du + \int_0^t s(u)dW_u, \tag{2.8}$$

where

$$0 < d \le s(u) \le g, \qquad b \le m(u) \le c, \tag{2.9}$$

for some constants $b, c, d,$ and g. Let $h > 0$ and $\zeta_h = \inf\{t \geq 0 : X_t = h\}$. Then for any fixed $t > 0$,

$$P(\zeta_0 < \zeta_h \wedge t) \to 1, \tag{2.10}$$

$$E\left(\max_{0\leq s\leq \zeta_0\wedge t} X_s\right) \to 0 \tag{2.11}$$

as $x \downarrow 0$ uniformly over all the processes X_t with the drift and diffusion terms subject to (2.9).

Proof. Let $\tilde{t}(r) = \int_0^r s(u)^2 du$ and $\hat{t}(v) = \tilde{t}^{-1}(v), Y(v) = X(\hat{t}(v))$. Then, in view of [14, Theorem 3.6, Corollary 3.7],

$$dY(v) = \hat{m}(v)dv + d\hat{W}_v,$$

where $\hat{W}$ is a standard Brownian motion. On the other hand, (2.9) shows $\frac{1}{d^2} \geq \frac{d\hat{t}(v)}{dv} \geq \frac{1}{g^2}$. Thus $|\hat{m}(v)| \leq \frac{1}{d^2}(|b| \vee |c|) \equiv K$,

$$X(s) = Y(\tilde{t}(s)), \qquad Y(s) \leq x + Ks + \hat{W}_s,$$

and

$$\max_{0\leq s\leq t} X(s) = \max_{0\leq v\leq \tilde{t}(t)} Y(v) \leq \max_{0\leq v\leq \tilde{t}(t)} (x + Kv + \hat{W}_v) \leq \max_{0\leq v\leq g^2 t} (x + Kv + \hat{W}_v). \tag{2.12}$$

Therefore, if $0 < x < h$, then

$$\tilde{t}(\zeta_0) \leq \hat{\zeta}_0, \qquad \tilde{t}(\zeta_h) \geq \hat{\zeta}_h, \tag{2.13}$$

where $\hat{\zeta}_h$ is the first hitting time of h by the process $x + Kv + \hat{W}_v$. As a result,

$$P(\zeta_0 < \zeta_h \wedge t) \geq P(\hat{\zeta}_0 < \hat{\zeta}_h \wedge d^2 t)$$

and

$$\max_{0\leq s\leq \zeta_0\wedge t} X_s \leq \max_{0\leq v\leq \hat{\zeta}_0\wedge g^2 t} (x + Kv + \hat{W}_v).$$

Thus the statement of the proposition follows from a similar statement for a standard Brownian motion with a constant drift. The latter is trivial since $x + Kv + \hat{W}_v$ decreases as $x \downarrow 0$ and the law of iterated logarithm (see [14, section II, Theorem 1.9]) implies that $\hat{\zeta}_0 \to 0$, wherefrom (2.10) follows. On the other hand, since $x + Kv + \hat{W}_v$ is a continuous process, $\max_{0\leq v\leq \hat{\zeta}_0\wedge g^2 t}(x + Kv + \hat{W}_v) \downarrow 0$ as $x \downarrow 0$. Moreover, this maximum is majorized by $\max_{0\leq v\leq g^2 t}(h + Kv + \hat{W}_v)$; hence (2.11) follows. ∎

Now we show that the value function v has the following basic properties.

PROPOSITION 1. *The value function v is a continuous, nondecreasing function subject to*

$$v(0+) = 0. \tag{2.14}$$

Proof. If $y > x$, then for any $\pi = (a_t, C_t)$, we can put $\hat{\pi} = (a_t, C_t + y - x)$. The policy $\hat{\pi}$ corresponds to instantaneously paying dividends in the amount of $y - x$,

thus instantaneously changing the initial reserve from y into x and then following the policy π. If R_t is the process satisfying (2.1), (2.2) and $\hat{R}_t$ is the process satisfying (2.1) and (2.2) with π replaced by $\hat{\pi}$ and x replaced by y in (2.2), then $R_t = \hat{R}_t$ for all $t > 0$. Obviously $J_y(\hat{\pi}) = (y - x) + J_x(\pi)$. This shows that

$$v(y) \geq (y - x) + v(x) \tag{2.15}$$

and, in particular, v is nondecreasing.

To prove (2.14), let $\pi = (a_t, C_t)$ be an arbitrary policy and R_t be given by (2.1), (2.2) while $X_t = R_t + C_t$. Thus X is the process governed by (2.8) with $m(u) = \mu a_u - \delta$ and $s(u) = \sigma a_u$. From the condition $\alpha \leq a_u \leq \beta$ follows (2.9). Fix $t > 0$. In view of Lemma 1, for any $0 < \varepsilon < 1$ choose x such that the expression in the left-hand side of (2.10) is greater than $1 - \varepsilon$ and the expression in the left-hand side of (2.11) is less than ε.

Fix $h > 0$ and let $\eta = \tau \wedge \zeta_h \wedge t$. Since $R_t \leq X_t$, we see that $\tau \leq \zeta_0$ and $P(\tau < \zeta_h \wedge t) > P(\zeta_0 < \zeta_h \wedge t) > 1 - \varepsilon$. Due to the requirement that $R_s = X_s - C_s \geq 0$ for all $s \leq \tau$, we have $C_\eta \leq X_\eta \leq \max_{0 \leq u \leq \eta} X_u$. Therefore $EC_\eta < \varepsilon$. As a result

$$\begin{aligned} J_x(\pi) &= E \int_0^\tau e^{-\gamma s} dC_s = E \int_0^\eta e^{-\gamma s} dC_s + E \left(1_{\tau > \eta} \int_\eta^\tau e^{-\gamma s} dC_s \right) \\ &\leq EC_\eta + E \left(1_{\tau > \eta} E \left(\int_\eta^\tau e^{-\gamma s} dC_s | \mathcal{F}_\eta \right) \right) \leq \varepsilon + E \left(1_{\tau > \eta} e^{-\gamma \eta} v(R_\eta) \right). \end{aligned} \tag{2.16}$$

The last inequality in (2.16) is due to the definition of the value function v. Since $R_\eta \leq X_\eta \leq h$, and the function v is increasing, $v(R_\eta) \leq v(h)$. Consequently,

$$J_x(\pi) \leq \varepsilon + E(1_{\tau > \eta} e^{-\gamma \eta} v(h)) \leq \varepsilon + v(h) P(\tau > \eta) \leq \varepsilon(1 + v(h)). \tag{2.17}$$

In view of arbitrariness of ε, we conclude the validity of (2.14).

Finally, we prove that v is continuous at any $y > 0$. Let $\pi = (a_t, C_t)$ be a control admissible from y and R_t is the corresponding process with the initial position y (that is, $R_{0-} = y$). Let $0 < x < y$ be such that the expression in the left-hand side of (2.10) is greater than $1 - \varepsilon$ and the expression in the left-hand side of (2.11) is less than ε. Suppose $\xi \geq 0$ is any stopping time such that $R_\xi \leq x$. Then, following exactly the same line of proof as for (2.17), we can show that

$$E \left(\int_\xi^\tau e^{-\gamma t} dC_t \right) \leq \varepsilon(1 + v(h)), \tag{2.18}$$

where $h > 0$ is fixed. Let $\hat{y} = y - x$ and $\hat{\pi} = (\hat{a}_t, \hat{C}_t)$ be the control which makes the resulting state process $\hat{R}$ "trail" the process R at the constant distance of x until the bankruptcy of the tracing process. That is, $(\hat{a}_t, \hat{C}_t) = (a_t, C_t)$, $t < \zeta$, where $\zeta = \inf\{t \geq 0 : \hat{R}_t \leq 0\}$ and

$$\hat{R}_t = y - x + \int_0^t (a_s \mu - \delta) ds + \int_0^t a_s \sigma dW_s - C_t.$$

We set $\hat{C}_\zeta = C_{\zeta -} + \hat{R}_{\zeta -}$. Obviously $\hat{R}_t = R_t - x$ for $t < \zeta$ and $\hat{R}_\zeta = 0$. Since $\hat{R}_{\zeta -} = R_{\zeta -} - x$ and $\hat{R}_{\zeta -} - (C_\zeta - C_{\zeta -}) \leq 0$, we conclude that

$$R_\zeta = R_{\zeta -} - (C_\zeta - C_{\zeta -}) \leq x. \tag{2.19}$$

On the other hand, since $R_\zeta = R_{\zeta-} - (C_\zeta - C_{\zeta-}) = \hat{R}_{\zeta-} + x - (C_\zeta - C_{\zeta-}) \geq 0$, we also have

$$(C_\zeta - C_{\zeta-}) - \hat{R}_{\zeta-} \leq x. \tag{2.20}$$

Then

$$\begin{aligned} J_y(\pi) &= E\left(\int_0^\tau e^{-\gamma t} dC_t\right) = E\left(\int_0^\zeta e^{-\gamma t} dC_t + \int_\zeta^\tau e^{-\gamma t} dC_t\right) \\ &= E\left[\int_0^{\zeta-} e^{-\gamma t} dC_t + e^{-\gamma\zeta}(C_\zeta - C_{\zeta-}) + \int_\zeta^\tau e^{-\gamma t} dC_t\right] \\ &= E\left\{\int_0^{\zeta-} e^{-\gamma t} dC_t + e^{-\gamma\zeta}\hat{R}_{\zeta-} + e^{-\gamma\zeta}[(C_\zeta - C_{\zeta-}) - \hat{R}_{\zeta-}] + \int_\zeta^\tau e^{-\gamma t} dC_t\right\} \\ &= J_{y-x}(\hat{\pi}) + E\left\{e^{-\gamma\zeta}[(C_\zeta - C_{\zeta-}) - \hat{R}_{\zeta-}] + \int_\zeta^\tau e^{-\gamma t} dC_t\right\}. \end{aligned}$$

In view of (2.19) we see that (2.18) is true with $\xi = \zeta$. This, together with (2.20), yields

$$J_y(\pi) \leq J_{y-x}(\hat{\pi}) + x + \varepsilon(1 + v(h)).$$

Thus

$$0 \leq v(y) - v(y-x) \leq x + \varepsilon(1 + v(h)),$$

which shows continuity of v. □

Since the value function has been proved to be continuous, the following dynamic programming principle holds:

$$v(x) = \sup_{\pi\in\mathcal{A}}\left[E\int_0^{\tau\wedge\theta} e^{-\gamma t} dC_t + Ee^{-\gamma(\tau\wedge\theta)} v(R_{\tau\wedge\theta})\right] \tag{2.21}$$

for every $x \geq 0$ and $\mathcal{F}_t$-stopping time θ (which may depend on the policy π); see [7, p. 333] for details.

PROPOSITION 2. *The value function v is a concave function.*

Proof. First note that the dynamic programming principle (2.21) implies that if $0 < y < x$, then

$$v(x) = \sup_{\pi\in\mathcal{A}} E_x\left(\int_0^{\chi_y} e^{-\gamma t} dC_t + e^{-\gamma\chi_y} v(R_{\chi_y})\right), \tag{2.22}$$

where $\chi_y = \inf\{t \geq 0 : R_t \leq y\}$ with R_t the reserve process corresponding to π. Since $x > y$ and the process R_t can have only the downward jumps, we have

$$R_{\chi_y} \leq y \tag{2.23}$$

and $R_{\chi_y} \equiv R_{\chi_y-} - \Delta C_{\chi_y} < y$ only if $\Delta C_{\chi_y} = C_{\chi_y} - C_{\chi_y-} > 0$. Put $\Delta' C_{\chi_y} = R_{\chi_y-} - y$ and $\Delta'' C_{\chi_y} = \Delta C_{\chi_y} - \Delta' C_{\chi_y}$. In view of (2.23) we have $\Delta' C_{\chi_y} \leq \Delta C_{\chi_y}$, $\Delta'' C_{\chi_y} \geq 0$ and

$$R_{\chi_y-} - \Delta' C_{\chi_y} = y, \qquad y - \Delta'' C_{\chi_y} = R_{\chi_y}. \tag{2.24}$$

Taking into account (2.24), we have for any $\pi \in \mathcal{A}$

$$
\begin{aligned}
& E_x\left(\textstyle\int_0^{\chi_y} e^{-\gamma t} dC_t + e^{-\gamma\chi_y} v(R_{\chi_y})\right) \\
= \; & E_x\left(\textstyle\int_0^{\chi_y -} e^{-\gamma t} dC_t + e^{-\gamma\chi_y}\Delta C_{\chi_y} + e^{-\gamma\chi_y}[v(y) + v(R_{\chi_y}) - v(y)]\right) \\
= \; & E_x\Big(\textstyle\int_0^{\chi_y -} e^{-\gamma t} dC_t + e^{-\gamma\chi_y}(\Delta' C_{\chi_y} + \Delta'' C_{\chi_y}) \\
& \qquad + e^{-\gamma\chi_y}[v(y) + v(y - \Delta'' C_{\chi_y}) - v(y)]\Big) \\
= \; & E_x\Big(\textstyle\int_0^{\chi_y -} e^{-\gamma t} dC_t + e^{-\gamma\chi_y}\Delta' C_{\chi_y} + e^{-\gamma\chi_y} v(y) \\
& \qquad + e^{-\gamma\chi_y}[v(y - \Delta'' C_{\chi_y}) - v(y) + \Delta'' C_{\chi_y}]\Big).
\end{aligned} \tag{2.25}
$$

In view of (2.15),

$$v(y - \Delta'' C_{\chi_y}) - v(y) \le -\Delta'' C_{\chi_y};$$

therefore (2.25) implies

$$
\begin{aligned}
& E_x\left(\int_0^{\chi_y} e^{-\gamma t} dC_t + e^{-\gamma\chi_y} v(R_{\chi_y})\right) \\
& \le E_x\left(\int_0^{\chi_y -} e^{-\gamma t} dC_t + e^{-\gamma\chi_y}\Delta' C_{\chi_y} + e^{-\gamma\chi_y} v(y)\right) \\
& = E_x\left(\int_0^{\chi_y} e^{-\gamma t} dC'_t + e^{-\gamma\chi_y} v(y)\right),
\end{aligned} \tag{2.26}
$$

where

$$
\begin{aligned}
C'_s &= C_s \qquad \forall s < \chi_y, \\
C'_{\chi_y} &= C_{\chi_y -} + \Delta' C_{\chi_y}.
\end{aligned}
$$

Thus if $(a_t, C_t) \in \mathcal{A}$, then by changing C_t into C'_t (that is, not changing control C_t until χ_y and only changing the jump at the time χ_y from ΔC_{χ_y} to $\Delta' C_{\chi_y}$) we can only increase the left-hand side of (2.25). Therefore the supremum in the right-hand side of (2.22) can be taken over only those controls in $\mathcal{A}$ for which ΔC_{χ_y} is replaced by $\Delta' C_{\chi_y}$. Note that an easy calculation shows that under the control (a_t, C'_t) the corresponding reserve process *always* satisfies

$$R_{\chi_y} = y. \tag{2.27}$$

Consequently, for the right-hand side of the dynamic programming equation (2.22) we need only consider those controls π for which (2.27) holds.

For $h > 0$, let $\mathcal{A}^h$ be the set of controls (a_t, C_t) such that

$$\int_0^\zeta \mu a(s) ds + \int_0^\zeta \sigma a(s) dW_s - \delta\zeta - C_\zeta = -h$$

on the set $\{\zeta < \infty\}$, where

$$\zeta = \inf\left\{t \ge 0 : \int_0^t \mu a(s) ds + \int_0^t \sigma a(s) dW_s - \delta t - C_t \le -h\right\}.$$

From (2.22) and the argument above, we can write (putting $h = x - y$)

$$v(x) = \sup_{(a_t, C_t) \in \mathcal{A}^h} E_x \left(\int_0^\zeta e^{-\gamma t} dC_t + e^{-\gamma \zeta} v(y) \right). \tag{2.28}$$

Thus, by putting $y = x - h$ in (2.28), we have

$$v(x) - v(x-h) = \sup_{(a_t, C_t) \in \mathcal{A}^h} E_x \left(\int_0^\zeta e^{-\gamma t} dC_t + (e^{-\gamma \zeta} - 1) v(x-h) \right). \tag{2.29}$$

Since $e^{-\gamma\zeta} - 1 \leq 0$ and $v(x)$ is a nondecreasing function as shown earlier, the right-hand side of (2.29) is a nonincreasing function of x. Thus $v(x) - v(x-h)$ is a nonincreasing function of x, which shows the concavity of v. □

Remark 1. Normally a proof for the concavity of the value function is straightforward if the dynamics of the underlying stochastic control model is linear and is on, say, an infinite time horizon. In such a case there is no need to employ the dynamic programming approach. In the present case, however, the random time horizon terminated by a stopping time, the constraint that a_t must be *strictly* positive as well as the presence of $\delta > 0$ render the normal approach invalid. Here, we use the dynamic programming principle to overcome the difficulty in proving the concavity of the value function, which is new according to our best knowledge.

THEOREM 1. *The value function v is a viscosity solution of the HJB equation*

$$\begin{aligned} &\max\left(\max_{\alpha \leq a \leq \beta} \left(\tfrac{1}{2}\sigma^2 a^2 V''(x) + (a\mu - \delta) V'(x) - \gamma V(x) \right),\ 1 - V'(x) \right) = 0, \quad x > 0, \\ &V(0) = 0. \end{aligned} \tag{2.30}$$

That is,

(i) *for any $x_0 > 0$ and any C^2 function f such that $f(x_0) = v(x_0)$ and $f(x) \geq v(x)$ for all x in the neighborhood of x_0,*

$$\max\left(\max_{\alpha \leq a \leq \beta} \left(\frac{1}{2}\sigma^2 a^2 f''(x_0) + (a\mu - \delta) f'(x_0) - \gamma f(x_0) \right),\ 1 - f'(x_0) \right) \geq 0; \tag{2.31}$$

(ii) *for any $x_0 > 0$ and any C^2 function f such that $f(x_0) = v(x_0)$ and $f(x) \leq v(x)$ for all x in the neighborhood of x_0,*

$$\max\left(\max_{\alpha \leq a \leq \beta} \left(\frac{1}{2}\sigma^2 a^2 f''(x_0) + (a\mu - \delta) f'(x_0) - \gamma f(x_0) \right),\ 1 - f'(x_0) \right) \leq 0. \tag{2.32}$$

Moreover, if v is twice differentiable at the point x_0, then v satisfies (2.30) *at x_0.*

Proof. For any $a \in [\alpha, \beta]$ denote the operator

$$L^a = \frac{1}{2}\sigma^2 a^2 \frac{d^2}{dx^2} + (a\mu - \delta)\frac{d}{dx} - \gamma. \tag{2.33}$$

Fix $x_0 > 0$. Dynamic programming principle yields

$$v(x_0) = \sup_{\pi \in \mathcal{A}} \left[E \int_0^{\tau \wedge \tau'} e^{-\gamma t} dC_t + E e^{-\gamma(\tau \wedge \tau')} v(R_{\tau \wedge \tau'}) \right] \tag{2.34}$$

for any $\mathcal{F}_t$-stopping time τ' (which may depend on the policy π).

For any C^2 function f such that $f(x_0) = v(x_0)$ and $f(x) \le v(x)$ for all x in the neighborhood of x_0, let $O_\varepsilon(x_0) = [x_0-\varepsilon, x_0+\varepsilon]$ be the small interval, where $\varepsilon > 0$, such that $f(x) \le v(x)$ and $x > 0$ for all $x \in O_\varepsilon(x_0)$. Fix a policy $\pi = (a_t, C_t; t \ge 0) \in \mathcal{A}$, where $C_t = \begin{cases} 0 \text{ if } t = 0-, \\ \eta \text{ if } t \ge 0, \end{cases}$ with $0 \le \eta < \varepsilon$. Let θ be the exit time of the corresponding reserve process R_t from $O_\varepsilon(x_0)$, namely, $\theta = \inf\{t \ge 0 : R_t \notin O_\varepsilon(x_0)\}$. By the choice of $O_\varepsilon(x_0)$ above, $\theta \le \tau$. It is also easily seen that the corresponding reserve process R_t has at most one jump at $t = 0$ while remaining continuous on $(0, \theta]$ (hence staying in $O_\varepsilon(x_0)$ at or before θ). Now, taking $\tau' = \theta \wedge h$ in (2.34), where $h > 0$ is a deterministic quantity, and noting that $R_{\theta\wedge h} \in O_\varepsilon(x_0)$, we have

$$\begin{aligned} f(x_0) = v(x_0) &\ge E[\textstyle\int_0^{\theta\wedge h} e^{-\gamma s} dC_s + e^{-\gamma(\theta\wedge h)} v(R_{\theta\wedge h})] \\ &\ge E[\textstyle\int_0^{\theta\wedge h} e^{-\gamma s} dC_s + e^{-\gamma(\theta\wedge h)} f(R_{\theta\wedge h})] \qquad \forall h > 0. \end{aligned} \tag{2.35}$$

Applying a generalized Ito formula (see Dellacherie and Meyer [5, Theorem VIII.27]) to the process $e^{-\gamma t} f(R_t)$, we get (below C_t^c stands for the continuous part of the increasing process C_t)

$$\begin{aligned} e^{-\gamma(\theta\wedge h)} f(R_{\theta\wedge h}) &= f(x_0) + \int_0^{\theta\wedge h} e^{-\gamma s} \sigma a_s f'(R_s) dW_s \\ &\quad + \int_0^{\theta\wedge h} e^{-\gamma s} L^{a_s} f(R_s) ds - \int_0^{\theta\wedge h} e^{-\gamma s} f'(R_s) dC_s \\ &\quad + \sum_{s \le \theta\wedge h} e^{-\gamma s} [f(R_s) - f(R_{s-}) - f'(R_{s-})(R_s - R_{s-})] \\ &= f(x_0) + \int_0^{\theta\wedge h} e^{-\gamma s} \sigma a_s f'(R_s) dW_s + \int_0^{\theta\wedge h} e^{-\gamma s} L^{a_s} f(R_s) ds \\ &\quad - \int_0^{\theta\wedge h} e^{-\gamma s} f'(R_s) dC_s^c + \sum_{s \le \theta\wedge h} e^{-\gamma s} [f(R_s) - f(R_{s-})], \end{aligned} \tag{2.36}$$

where we used the equality $R_s - R_{s-} = -(C_s - C_{s-})$. Since f is C^2 and R_s stays in the bounded region $O_\varepsilon(x_0)$ for $s \le \theta$, we conclude that $f'(R_s)$ is bounded on $[0, \theta \wedge h]$ and hence the stochastic integral in (2.36) is a square integrable martingale whose expectation vanishes. Taking the expectation of (2.36) and substituting it into (2.35), we obtain

$$\begin{aligned} &E\textstyle\int_0^{\theta\wedge h} e^{-\gamma s} dC_s + E\int_0^{\theta\wedge h} e^{-\gamma s} L^{a_s} f(R_s) ds \\ &\quad - E\textstyle\int_0^{\theta\wedge h} e^{-\gamma s} f'(R_s) dC_s^c + E\sum_{s \le \theta\wedge h} e^{-\gamma s} [f(R_s) - f(R_{s-})] \le 0 \qquad \forall h > 0. \end{aligned} \tag{2.37}$$

In the above, taking $a_t \equiv a$ and $C_t \equiv 0$ (i.e., $\eta = 0$), where $a \in [\alpha, \beta]$, we have $E\int_0^{\theta\wedge h} e^{-\gamma s} L^a f(R_s) ds = E\int_0^h e^{-\gamma s} L^a f(R_s) 1_{s \le \theta} ds \le 0$ for all $h > 0$. Note that in this case R_s is continuous at $s = 0$ and the integrand, $e^{-\gamma s} L^a f(R_s) 1_{s \le \theta}$, converges to $L^a f(x_0)$ almost surely as $s \to 0$. Dividing the above inequality by h and sending h to zero, we conclude by the uniform boundedness of the above integrand that

$$L^a f(x_0) \le 0 \qquad \forall a \in [\alpha, \beta]. \tag{2.38}$$

On the other hand, in (2.37) taking $a_t \equiv a$ and $\eta > 0$, we obtain

$$E\int_0^{\theta\wedge h} e^{-\gamma s} L^a f(R_s) ds + \eta + f(x_0 - \eta) - f(x_0) \le 0 \qquad \forall h > 0 \ \ \forall 0 < \eta < \varepsilon.$$

First letting $h \to 0$, then dividing by η and finally sending η to 0, we obtain

$$1 - f'(x_0) \leq 0. \tag{2.39}$$

Combining (2.38) and (2.39) we arrive at (2.32).

To prove (2.31), let a C^2 function f be such that $f(x_0) = v(x_0)$ and $f(x) \geq v(x)$ for all x in the neighborhood of $x_0 > 0$. If (2.31) is not true, then there is $A > 0$ such that

$$\max\left(\max_{\alpha\leq a\leq\beta}\left(\frac{1}{2}\sigma^2a^2f''(x) + (a\mu - \delta)f'(x) - \gamma f(x)\right),\ 1 - f'(x)\right) \leq -A < 0 \tag{2.40}$$

for all $x \in O_\varepsilon(x_0) = [x_0 - \varepsilon, x_0 + \varepsilon]$, where $\varepsilon > 0$. As before we can make ε sufficiently small so that $f(x) \geq v(x)$ and $x > 0$ for all $x \in O_\varepsilon(x_0)$. Given a policy $\pi \in \mathcal{A}$, let $\theta = \inf\{t \geq 0 : R_t \notin O_\varepsilon(x_0)\}$. Note that it is possible that $\theta = 0+$, meaning that there is a jump at $t = 0$ getting the reserve process out of $O_\varepsilon(x_0)$ instantly. However, by definition R_s stays in $O_\varepsilon(x_0)$ so long as $s < \theta$. Once again, $\theta \leq \tau$. (If there is a jump at $t = 0$ making the reserve process zero instantly, then $\theta = \tau = 0+$.) Now, the generalized Ito formula yields

$$\begin{aligned} &Ee^{-\gamma\theta}f(R_{\theta-}) - f(x_0) + E\textstyle\int_0^{\theta-} e^{-\gamma s}dC_s \\ &\quad = E\textstyle\int_0^{\theta-} e^{-\gamma s}L^{a_s}f(R_s)ds - E\int_0^{\theta-} e^{-\gamma s}f'(R_s)dC_s^c \\ &\qquad + E\textstyle\sum_{s<\theta} e^{-\gamma s}[f(R_s) - f(R_{s-})] + E\int_0^{\theta-} e^{-\gamma s}dC_s. \end{aligned} \tag{2.41}$$

First we have

$$\begin{aligned} f(R_s) - f(R_{s-}) &= (R_s - R_{s-})\textstyle\int_0^1 f'(R_{s-} + z(R_s - R_{s-}))dz \\ &= -(C_s - C_{s-})\int_0^1 f'(R_{s-} + z(R_s - R_{s-}))dz \\ &\leq -(1 + A)(C_s - C_{s-}), \end{aligned} \tag{2.42}$$

where the last inequality is due to (2.40). Going back to (2.41), noting (2.40) and (2.42), we have

$$\begin{aligned} &Ee^{-\gamma\theta}f(R_{\theta-}) - f(x_0) + E\textstyle\int_0^{\theta-} e^{-\gamma s}dC_s \\ \leq\ &-AE\textstyle\int_0^\theta e^{-\gamma s}ds - (1 + A)E\int_0^{\theta-} e^{-\gamma s}dC_s^c - (1 + A)E\sum_{s<\theta} e^{-\gamma s}(C_s - C_{s-}) \\ &+E\textstyle\int_0^{\theta-} e^{-\gamma s}dC_s \\ =\ &-AE\textstyle\int_0^\theta e^{-\gamma s}ds - (1 + A)E\int_0^{\theta-} e^{-\gamma s}dC_s + E\int_0^{\theta-} e^{-\gamma s}dC_s \\ =\ &-AE\textstyle\int_0^\theta e^{-\gamma s}ds - AE\int_0^{\theta-} e^{-\gamma s}dC_s. \end{aligned} \tag{2.43}$$

To proceed, note that $R_\theta \leq R_{\theta-} \in O_\varepsilon(x_0)$ while R_θ may be out of $O_\varepsilon(x_0)$. However, one can always find a point x_λ on the boundary of $O_\varepsilon(x_0)$ such that

$$x_\lambda = R_{\theta-} + \lambda(R_\theta - R_{\theta-}) \equiv R_{\theta-} - \lambda(C_\theta - C_{\theta-}) \in \{x_0 - \varepsilon, x_0 + \varepsilon\}, \tag{2.44}$$

where $\lambda \in [0, 1]$ is a random variable. Clearly we have $R_\theta \leq x_\lambda \leq R_{\theta-}$, and $x_\lambda = x_0 - \epsilon$ if $R_\theta \notin O_\epsilon(x_0)$. Now, similar to (2.42), we have

$$f(R_{\theta-}) - f(x_\lambda) \geq (1 + A)(R_{\theta-} - x_\lambda) = \lambda(1 + A)(C_\theta - C_{\theta-}). \tag{2.45}$$

Moreover, since $R_\theta \le x_\lambda$ we have by (2.15) that

$$v(x_\lambda) \ge x_\lambda - R_\theta + v(R_\theta) = (1-\lambda)(C_\theta - C_{\theta-}) + v(R_\theta). \tag{2.46}$$

Combining (2.43), (2.45), and (2.46), we derive

(2.47)
$$\begin{aligned} v(x_0) = f(x_0) &\ge E\int_0^{\theta-} e^{-\gamma s}dC_s + Ee^{-\gamma\theta}f(R_{\theta-}) \\ &\quad + A\big[E\int_0^{\theta} e^{-\gamma s}ds + E\int_0^{\theta-} e^{-\gamma s}dC_s\big] \\ &\ge E\int_0^{\theta-} e^{-\gamma s}dC_s + Ee^{-\gamma\theta}f(x_\lambda) + \lambda(1+A)Ee^{-\gamma\theta}(C_\theta - C_{\theta-}) \\ &\quad + A\big[E\int_0^{\theta} e^{-\gamma s}ds + E\int_0^{\theta-} e^{-\gamma s}dC_s\big] \\ &\ge E\int_0^{\theta-} e^{-\gamma s}dC_s + Ee^{-\gamma\theta}v(x_\lambda) + \lambda(1+A)Ee^{-\gamma\theta}(C_\theta - C_{\theta-}) \\ &\quad + A\big[E\int_0^{\theta} e^{-\gamma s}ds + E\int_0^{\theta-} e^{-\gamma s}dC_s\big] \\ &\ge E\int_0^{\theta-} e^{-\gamma s}dC_s + Ee^{-\gamma\theta}v(R_\theta) + (1-\lambda)Ee^{-\gamma\theta}(C_\theta - C_{\theta-}) \\ &\quad + \lambda(1+A)Ee^{-\gamma\theta}(C_\theta - C_{\theta-}) \\ &\quad + A\big[E\int_0^{\theta} e^{-\gamma s}ds + E\int_0^{\theta-} e^{-\gamma s}dC_s\big] \\ &= E\int_0^{\theta} e^{-\gamma s}dC_s + Ee^{-\gamma\theta}v(R_\theta) \\ &\quad + A\big[E\int_0^{\theta} e^{-\gamma s}ds + E\int_0^{\theta-} e^{-\gamma s}dC_s + \lambda Ee^{-\gamma\theta}(C_\theta - C_{\theta-})\big]. \end{aligned}$$

Next we are going to show that there is a constant $k_0 > 0$ such that

$$E\int_0^{\theta} e^{-\gamma s}ds + E\int_0^{\theta-} e^{-\gamma s}dC_s + \lambda Ee^{-\gamma\theta}(C_\theta - C_{\theta-}) \ge k_0 \qquad \forall \pi \in \mathcal{A}. \tag{2.48}$$

To this end, define a C^2 function

$$w(x) = K_0(|x - x_0|^2 - \varepsilon^2), \tag{2.49}$$

where

(2.50)
$$K_0 = \min\left\{ \frac{1}{\sup_{|x-x_0|\le\delta,\alpha\le a\le\beta}\left[\sigma^2a^2 + \gamma\varepsilon^2 + 2|(a\mu-\delta)(x-x_0)|\right]}, \frac{1}{2\varepsilon} \right\} > 0.$$

From the definition of $w(\cdot)$ it is easy to verify that

$$L^a w(x) \le 1 \quad \text{and} \quad |w'(x)| \le 1 \qquad \forall x \in O_\varepsilon(x_0)\ \ \forall a \in [\alpha,\beta]. \tag{2.51}$$

Now applying the generalized Ito formula, we have

$$\begin{aligned} &E[e^{-\gamma\theta}w(R_{\theta-}) - w(x_0)] \\ &= E\int_0^{\theta-} e^{-\gamma s}L^{a_s}w(R_s)ds - E\int_0^{\theta-} e^{-\gamma s}w'(R_s)dC_s^c \\ &\qquad + E\textstyle\sum_{s<\theta} e^{-\gamma s}[w(R_s) - w(R_{s-})] \\ &\le E\int_0^{\theta} e^{-\gamma s}ds + E\int_0^{\theta-} e^{-\gamma s}dC_s, \end{aligned} \tag{2.52}$$

where the last inequality is due to (2.51). However, since $w'(x) \ge -1$, we have

$$w(R_{\theta-}) - w(x_\lambda) \ge -(R_{\theta-} - x_\lambda) = -\lambda(C_\theta - C_{\theta-}). \tag{2.53}$$

Substituting (2.53) into (2.52) we obtain

(2.54)
$$E\int_0^{\theta} e^{-\gamma s}ds + E\int_0^{\theta-} e^{-\gamma s}dC_s \geq E[e^{-\gamma\theta}w(x_\lambda) - w(x_0)] - \lambda E e^{-\gamma\theta}(C_\theta - C_{\theta-}).$$

Noting that $w(x_\lambda) = 0$ and $w(x_0) = -K_0\varepsilon^2$, we prove (2.48) with $k_0 := K_0\varepsilon^2 > 0$.

Thus, (2.47) gives

$$v(x_0) \geq E\int_0^{\theta} e^{-\gamma s}dC_s + Ee^{-\gamma\theta}v(R_\theta) + Ak_0, \tag{2.55}$$

which holds true for any $\pi \in \mathcal{A}$. Taking supremum over $\pi \in \mathcal{A}$ on both sides and appealing to (2.34), we have

$$v(x_0) \geq v(x_0) + Ak_0,$$

which is a contradiction. This proves that (2.40) is invalid.

Finally, if v is twice differentiable at x_0, then by definition we have $(v'(x_0), v''(x_0)) \in D^{2,+}v(x_0)$, the latter being the second-order superdifferential of v at x_0. Hence, by [19, p. 193, Lemma 5.4], there is a C^2 function f such that $v - f$ attains a strict maximum at x_0 and $(v(x_0), v'(x_0), v''(x_0)) = (f(x_0), f'(x_0), f''(x_0))$. By (i) proved above, f satisfies (2.31) at x_0, hence so does v at x_0. On the other hand, we also have $(v'(x_0), v''(x_0)) \in D^{2,-}v(x_0)$, the latter being the second-order subdifferential of v at x_0. Thus a symmetric reasoning leads to that v satisfies (2.32) at x_0. This proves the last claim of the theorem. □

Remark 2. It is proved in Fleming and Soner [7, chapter VIII] that the value function of a *pure* singular control problem is a viscosity solution of the corresponding HJB equation. The proof there is very involved as it is for a multidimensional problem. Our proof here is for a mixed regular-singular problem and is relatively simple by greatly exploiting the special structure of the single dimensionality.

Remark 3. Technically speaking the boundary condition in (2.30) should be $V(0+) = 0$ as (2.14) shows, rather than $V(0) = 0$. However, here and in what follows we will always adopt a convention that the solution of the HJB equation is extended to 0 by continuity and the value of V at 0 is its limit from the right.

Based on the fact that the value function is a concave viscosity solution of the HJB equation, we are able to show that it is in fact C^2. To this end we need the following lemma.

LEMMA 2. *Suppose g is a concave function such that $g(x) = g(x_0) + a(x - x_0)$ for $x \leq x_0$ and $g(x) = g(x_0) + b(x - x_0)$ for $x \geq x_0$, where $a > b$. Then for each sufficiently small $\varepsilon > 0$ there exists a concave C^2 function $f \geq g$ such that $f(x_0) = g(x_0)$, $f'(x) = a$, $x < x_0 - \varepsilon$, $f'(x) = b$, $x \geq x_0 + \varepsilon$, $f'(x_0) = (b + a)/2$, and $f''(x_0) \leq -\varepsilon^{-1}$.*

Proof. Choose $F(x)$ to be a nonincreasing continuously differentiable function, such that $F(x) = a$ for $x \leq x_0 - \varepsilon$, $F(x) = b$ for $x \geq x_0 + \varepsilon$, and $F(x) = -\frac{2}{\varepsilon}(x - x_0) + (b + a)/2$ for $-\varepsilon^2/2 < x - x_0 < \varepsilon^2/2$. Then $f(x) := g(x_0) + \int_{x_0}^{x} F(y)dy$ is the desired function. □

THEOREM 2. *The value function $v(x)$ is twice continuously differentiable for all $x > 0$.*

Proof. We divide the proof into several steps.

Step 1°. In view of concavity, the function v is a continuous function on $(0,\infty)$. First we show that v is a C^1 function. Let $D_+v(x)$ and $D_-v(x)$ be the right and the left derivatives of v at x. In view of concavity of v, the derivatives $D_+v(x)$ and $D_-v(x)$ exist for any $x>0$ and $D_-v(x)\geq D_+v(x)$. By virtue of (2.15) $D_+v(x)\geq 1$ for all $x>0$. Suppose x_0 is such that $D_-v(x_0)>D_+v(x_0)\geq 1$. Then $v(x)\leq g(x)$, where $g(x)=v(x_0)+D_+v(x_0)(x-x_0)$ for $x\geq x_0$ and $g(x)=v(x_0)+D_-v(x_0)(x-x_0)$ for $x\leq x_0$. Lemma 2 guarantees an existence of a concave C^2 function f such that $f(x)\geq g(x)\geq v(x)$, $f(x_0)=g(x_0)=v(x_0)$, $f'(x_0)=(D_+v(x_0)+D_-v(x_0))/2>1$, and $f''(x_0)\leq -\varepsilon^{-1}$. Theorem 1 shows that f must satisfy (2.31). Since $f'(x_0)>1$ we must have

$$\frac{1}{2}\sigma^2a^2f''(x_0)+(a\mu-\delta)f'(x_0)-\gamma f(x_0)\geq 0 \tag{2.56}$$

for some $\alpha\leq a\leq\beta$. However, $|(a\mu-\delta)f'(x_0)-\gamma f(x_0)|\leq|(a\mu-\delta)f'(x_0)|+\gamma v(x_0)\leq(\beta\mu-\delta)D_-v(x_0)+\gamma v(x_0)$ while $\frac{1}{2}\sigma^2a^2f''(x_0)<-\frac{1}{2}\alpha^2\sigma^2\frac{1}{\varepsilon}$. Thus choosing ε sufficiently small, we can see that (2.56) is violated. This shows that $D_-v(x_0)=D_+v(x_0)$.

Step 2°. Inequality (2.15) shows that $v'(x)\geq 1$ for all $x>0$. Let $x_1=\min\{x: v'(x)=1\}$. Since v' is a nonincreasing function, we have $v'(x)>1$ for all $x<x_1$. If f is subject to condition (i) of Theorem 1, then $f'(x_0)=v'(x_0)>1$ for $x_0<x_1$. Therefore (2.31) implies

$$\max_{\alpha\leq a\leq\beta}\left(\frac{1}{2}\sigma^2a^2f''(x_0)+(a\mu-\delta)f'(x_0)-\gamma f(x_0)\right)\geq 0. \tag{2.57}$$

Since v is concave, existence of the derivative of v everywhere implies that v' is a continuous nonincreasing function. Let $\mathcal{B}$ be the set of points x where $v''(x)$ exists. In view of [15, chapter 5, Theorem 3] the Lebesgue measure of the complement of $\mathcal{B}$ is zero. Thus $\mathcal{B}$ is an everywhere dense set.

Suppose $y_n\in\mathcal{B}, y_n<x_1$ and $y_n\to x_0$. Then by Theorem 1,

$$\max_{\alpha\leq a\leq\beta}\left(\frac{1}{2}\sigma^2a^2v''(y_n)+(a\mu-\delta)v'(y_n)-\gamma v(y_n)\right)=0. \tag{2.58}$$

Let $a^*(y_n)$ be the maximizer of the left-hand side of (2.58). Since $v(y_n)\to v(x_0)$, $v'(x_n)\to v'(x_0)$, and $\beta\geq a^*(y_n)\geq\alpha>0$ we see that $v''(y_n)$ is a bounded sequence. Choosing a subsequence if necessary, we can assume that $v''(y_n)\to q$. The expression which is maximized in the left-hand side of (2.58) is a quadratic polynomial of a on the interval $[\alpha,\beta]$ with convergent coefficients. Therefore we can pass to a limit and conclude

$$\max_{\alpha\leq a\leq\beta}\left(\frac{1}{2}\sigma^2a^2q+(a\mu-\delta)v'(x_0)-\gamma v(x_0)\right)=0. \tag{2.59}$$

If there exists another sequence $z_n\in\mathcal{B}, z_n<x_1$, such that $z_n\to x_0$ and $v''(z_n)\to q_1\neq q$, then the same arguments show that

$$\max_{\alpha\leq a\leq\beta}\left(\frac{1}{2}\sigma^2a^2q_1+(a\mu-\delta)v'(x_0)-\gamma v(x_0)\right)=0. \tag{2.60}$$

However,

$$\begin{aligned}&\left|\frac{1}{2}\sigma^2a^2q+(a\mu-\delta)v'(x_0)-\gamma v(x_0)-\frac{1}{2}\sigma^2a^2q_1-(a\mu-\delta)v'(x_0)-\gamma v(x_0)\right|\\ &=\frac{1}{2}\sigma^2a^2|q-q_1|\geq\frac{1}{2}\sigma^2\alpha^2|q-q_1|>0\end{aligned}$$

for all $a \in [\alpha, \beta]$. This inequality shows that (2.59) and (2.60) cannot hold simultaneously. Therefore $q = q_1$. As a result $v''(x)$ is a continuous function on $\mathcal{B}$ which can be extended to a continuous function on $(0, x_1)$. Consequently, v' is continuously differentiable on $(0, x_1)$.

Step 3°. The function v' is nonincreasing and $v'(x) \geq 1$ for all $x > 0$, while $v'(x) \leq 1$ for all $x \geq x_1$. Thus, $v'(x) = 1$ for all $x \geq x_1$. Therefore $v''(x) = 0$ for $x > x_1$. It is left only to show that $v''(x_1)$ exists.

Using the same arguments as in Step 2°, we can prove that $q = \lim_{y \uparrow x_1} v''(y)$ exists and (2.59) is true if we replace x_0 by x_1 in the left-hand side. Since v is concave, $q \leq 0$. Suppose $q < 0$. Then from (2.59) follows (recall that $v'(x_1) = 1$)

$$\begin{aligned}
&\max_{\alpha \leq a \leq \beta}[(a\mu - \delta) - \gamma v(x_1)] \\
&= \max_{\alpha \leq a \leq \beta}[(a\mu - \delta)v'(x_1) - \gamma v(x_1)] \\
&\geq \max_{\alpha \leq a \leq \beta}[\tfrac{1}{2}\sigma^2 a^2 q + (a\mu - \delta)v'(x_1) - \gamma v(x_1)] - \max_{\alpha \leq a \leq \beta}[\tfrac{1}{2}\sigma^2 a^2 q] \\
&= -\tfrac{1}{2}\sigma^2 \alpha^2 q \equiv K > 0.
\end{aligned}$$

Continuity of v implies that there exists $y > x_1$ such that $\gamma(v(y) - v(x_1)) < K/2$. Consequently,

$$\max_{\alpha \leq a \leq \beta}\left[\frac{1}{2}\sigma^2 a^2 v''(y) + (\mu a - \delta)v'(y) - \gamma v(y)\right] = \max_{\alpha \leq a \leq \beta}[(\mu a - \delta) - \gamma v(y)] > K/2 > 0.$$

The above inequality contradicts the last statement of Theorem 1. This contradiction proves that $q = v''(x_1-) = 0$. Therefore v'' can be continuously extended to the point x_1 and as a result v has a second derivative at x_1. □

In view of the last statement of Theorem 1, the following corollary is straightforward.

COROLLARY 1. *The value function v is the classical (i.e., C^2) solution to the HJB equation* (2.30).

Note that we do not know a priori whether the HJB equation has any C^2 solution other than the value function. However, the following verification theorem, which says that *any* concave solution V to the HJB equation (2.30) whose derivative is finite at 0 majorizes the performance functional for any policy π, is sufficient for us to identify optimal policies.

THEOREM 3. *Let V be a concave, twice continuously differentiable solution of* (2.30)*, such that $V'(0+) < +\infty$. Then for any policy $\pi = (a_t, C_t; t \geq 0)$,*

$$V(x) \geq J_x(\pi). \tag{2.61}$$

Proof. Let R_t be the reserve process given by (2.1) and (2.2) corresponding to a given policy π. Then applying the generalized Ito formula to the process $e^{-\gamma t}V(R_t)$, we get

(2.62)

$$\begin{aligned}
e^{-\gamma(t\wedge\tau)}V(R_{t\wedge\tau}) = V(x) &+ \int_0^{t\wedge\tau} e^{-\gamma s}\sigma a_s V'(R_s)dW_s + \int_0^{t\wedge\tau} e^{-\gamma s}L^{a_s}V(R_s)ds \\
&- \int_0^{t\wedge\tau} e^{-\gamma s}V'(R_s)dC_s^c + \sum_{s \leq t\wedge\tau} e^{-\gamma s}[V(R_s) - V(R_{s-})].
\end{aligned}$$

In view of the HJB equation (2.30), the quantity $L^{a_s}V(R_s)$ is always nonpositive. Taking expectations of both sides of (2.62), we get

$$E(e^{-\gamma(t\wedge\tau)}V(R_{t\wedge\tau})) \le V(x) - E\int_0^{t\wedge\tau} e^{-\gamma s}V'(R_s)dC_s^c + E\sum_{s\le t\wedge\tau} e^{-\gamma s}[V(R_s) - V(R_{s-})]. \tag{2.63}$$

Since $V'(x) \ge 1$, the mean-value theorem implies $V(R_s) - V(R_{s-}) \le R_s - R_{s-} = -(C_s - C_{s-})$. Thus in (2.63), we can replace $\sum_{s\le t\wedge\tau} e^{-\gamma s}[V(R_s) - V(R_{s-})]$ by $-\sum_{s\le t\wedge\tau} e^{-\gamma s}(C_s - C_{s-})$ with inequality preserved. Also the right-hand side of (2.63) will not decrease if we replace $V'(R_s)$ in the first integral by 1 because $V'(x) \ge 1$. As a result we get

$$\begin{aligned} E(e^{-\gamma(t\wedge\tau)}V(R_{t\wedge\tau})) &\le V(x) - E\int_0^{t\wedge\tau} e^{-\gamma s}dC_s^c - E\sum_{s\le t\wedge\tau} e^{-\gamma s}(C_s - C_{s-}) \\ &= V(x) - E\int_0^{t\wedge\tau} e^{-\gamma s}dC_s, \end{aligned}$$

or

$$E(e^{-\gamma(t\wedge\tau)}V(R_{t\wedge\tau})) + E\int_0^{t\wedge\tau} e^{-\gamma s}V'(R_s)dC_s \le V(x). \tag{2.64}$$

Note that in view of boundedness of V',

$$e^{-\gamma(t\wedge\tau)}V(R_{t\wedge\tau}) \le e^{-\gamma t}K(1 + R_{t\wedge\tau}) \le e^{-\gamma t}K(1 + |R_t|)$$

for some constant K. Since R_t is a diffusion process with uniformly bounded drift and diffusion coefficient, standard arguments yield $E|R_t| \le x + K_1 t$ for some constant K_1. Therefore

$$Ee^{-\gamma(t\wedge\tau)}V(R_{t\wedge\tau}) \to 0 \tag{2.65}$$

as $t \to \infty$. Thus taking limit in (2.64) as $t \to \infty$ we arrive at

$$V(x) \ge E\int_0^{\tau} e^{-\gamma s}V'(R_s)dC_s \ge J_\pi(x). \qquad \square$$

The idea of solving the original optimization problem is to first find a concave, smooth solution to the HJB equation (2.30) and then construct a control policy (via solving a Skorohod problem; for details see section 5) whose performance functional can be shown to coincide with the bounded solution to (2.30). Then, the above verification theorem establishes the optimality of the constructed control policy. As a by-product, we have a proof that there is no concave solution to (2.30) other than the value function.

3. Case of no liability. In this section we study the case where there is no debt liability, namely, $\delta = 0$. While being part of a more general case, it is interesting in its own right and will provide some valuable insights into the resolution of the general problem.

In this case, the HJB equation reads

(3.1)
$$\max\left(\max_{\alpha\le a\le\beta}\left(\tfrac{1}{2}\sigma^2a^2V''(x)+a\mu V'(x)-\gamma V(x)\right),\ 1-V'(x)\right)=0,\quad x>0,$$
$$V(0)=0.$$

As mentioned above, the key is to find a concave, smooth function V satisfying (3.1). While we could have presented such a solution immediately without any explanation (one would need only to check if it does satisfy (3.1), which is a relatively easy task), we believe that it is better to unfold the entire process of finding the solution for the benefit of the readers. Therefore, what we are going to present below is indeed the original process of tracking down the solution. Suppose such a solution, V, to the HJB equation (3.1) is found. Then due to concavity, V' is a nonincreasing function. Let $x_1=\inf\{x\ge 0\ :\ V'(x)\le 1\}$. Suppose $x_1>0$. Since V is concave we have $V'(x)>1$ for all $x<x_1$. In view of (3.1), V satisfies

$$\max_{\alpha\le a\le\beta}\left(\frac{1}{2}\sigma^2a^2V''(x)+a\mu V'(x)-\gamma V(x)\right)=0\qquad\forall x<x_1. \tag{3.2}$$

Note that if V satisfies (3.2) and $V''(x)=0$ on an open interval, then the maximum in the right-hand side of (3.2) is attained at $a=\beta$. Therefore on this interval the function V satisfies the first-order linear differential equation whose solution is $C\exp(\frac{\gamma}{\beta\mu}x)$, which contradicts concavity. Let

$$a(x)\equiv-\frac{\mu V'(x)}{\sigma^2V''(x)}>0,\qquad x<x_1, \tag{3.3}$$

be the maximizer of $\frac{1}{2}\sigma^2a^2V''(x)+a\mu V'(x)-\gamma V(x)$ over all $a\ge 0$, which is defined for those x for which $V''(x)\neq 0$. If $\alpha<a(x)<\beta$, then we can substitute the expression for $a(x)$ given by (3.3) back into (3.2) to get

$$-\frac{\mu^2(V'(x))^2}{2\sigma^2V''(x)}-\gamma V(x)=0.$$

Replacing $-\frac{\mu V'(x)}{\sigma^2V''(x)}$ in the above equation by $a(x)$ again, we get $\mu a(x)V'(x)/2-\gamma V(x)=0$ or

$$a(x)=\frac{2\gamma V(x)}{\mu V'(x)}.$$

For a concave nondecreasing function V the numerator of the above expression is a nondecreasing function, while the denominator is nonincreasing. Therefore $a(x)$ should be a nondecreasing function of x on the set where $\alpha<a(x)<\beta$. On the other hand continuity of V, V', and V'' implies that $a(x)$ is a continuous function of x for all $x>0$. Therefore, if $a(\bar{x})=\alpha$ for some $\bar{x}$, then $a(x)\le\alpha$ for all $x<\bar{x}$. Indeed, suppose $a(\bar{y})>\alpha$ for some $\bar{y}<\bar{x}$. Define $\hat{y}=\inf\{y>\bar{y}:a(y)\le\alpha\}\le\bar{x}$. Since $a(x)$ is continuous, $\alpha<a(z)<\beta$ whereas $a(z)$ is not nondecreasing for z in a left neighborhood of $\hat{y}$, leading to a contradiction. Likewise if $a(\bar{x})=\beta$ for some $\bar{x}$, then $a(x)\ge\beta$ for all $x>\bar{x}$.

Since $V''(x)\le 0$, the expression $\frac{1}{2}\sigma^2a^2V''(x)+a\mu V'(x)-\gamma V(x)$ as a function of a increases on $[0,\ a(x)]$ and decreases on $[a(x),\ \infty)$. As was just shown there exist

x_α and x_β, $0 \le x_\alpha < x_\beta \le +\infty$, such that $a(x) \le \alpha$ for all $x \le x_\alpha$ and $a(x) \ge \beta$ for all $x \ge x_\beta$; and $a(x)$ is nondecreasing from α to β on $[x_\alpha,\ x_\beta]$. Our next step is to show that $x_\alpha > 0$. To this end it is sufficient to analyze $a(0)$.

PROPOSITION 3. *The following holds:*

$$a(0) = \frac{1}{2}\alpha. \tag{3.4}$$

Proof. Put

$$\phi(x,\ a) = \frac{1}{2}\sigma^2 a^2 V''(x) + a\mu V'(x) - \gamma V(x), \qquad x \ge 0,\ a \ge 0. \tag{3.5}$$

It follows from (3.2) that $\max_{\alpha \le a \le \beta} \phi(0,\ a) = 0$. Let $\tilde{a} \in [\alpha,\ \beta]$ be such that $\phi(0,\ \tilde{a}) = \max_{\alpha \le a \le \beta} \phi(0,\ a) = 0$. Since $V(0) = 0$ we conclude

$$\tilde{a} = \frac{-2\mu V'(0)}{\sigma^2 V''(0)} \equiv 2a(0).$$

On the other hand, $\phi(0, a) = a[\frac{1}{2}\sigma^2 a V''(0) + \mu V'(0)]$, where $\frac{1}{2}\sigma^2 V''(0) \le 0$. Hence the maximum of $\phi(0, \cdot)$ is attained at the lower end of the interval $[\alpha, \beta]$, namely, $\tilde{a} = \alpha$. This proves (3.4). □

Continuity of $a(x)$ and the inequality $a(0) < \alpha$ show that for all x in a right neighborhood of 0, the maximum over $a \in [\alpha, \beta]$ in (3.2) is attained at α. Substituting $a = \alpha$ into (3.2) and solving the resulting second-order linear ordinary differential equation (ODE), we get

$$V(x) = k_1(\alpha,\ \beta)\left(e^{r_+(\alpha)x} - e^{r_-(\alpha)x}\right), \tag{3.6}$$

where $k_1(\alpha,\ \beta)$ is a free constant to be determined, and

$$r_+(z) \equiv \frac{-\mu + [\mu^2 + 2\sigma^2\gamma]^{1/2}}{z\sigma^2} > 0, \quad r_-(z) \equiv \frac{-\mu - [\mu^2 + 2\sigma^2\gamma]^{1/2}}{z\sigma^2} < 0 \qquad \forall z > 0. \tag{3.7}$$

From (3.6) and (3.3) it follows that

$$a'(x) = \frac{-\mu r_-(\alpha) r_+(\alpha) k_1^2(\alpha,\ \beta) e^{(r_+(\alpha)+r_-(\alpha))x} \left(r_+(\alpha) - r_-(\alpha)\right)^2}{(\sigma V''(x))^2} > 0$$

for all x in the right neighborhood of 0. Therefore in the right neighborhood of 0 the function $a(x)$ increases. Let x_α be such that $a(x_\alpha) = \alpha$. From (3.3) and (3.6), we obtain

$$x_\alpha = \frac{1}{r_+(\alpha) - r_-(\alpha)} \ln\left(-\frac{r_-(\alpha)}{r_+(\alpha)}\right) > 0. \tag{3.8}$$

PROPOSITION 4. *For each $x \ge x_\alpha$,*

$$a(x) \ge \alpha. \tag{3.9}$$

Proof. Suppose that there exists $x_0 > x_\alpha$ such that $a(x_0) < \alpha$. Then there exists $\varepsilon > 0$ such that $a(x) < \alpha$ for each x with $|x - x_0| < \varepsilon$. Let $x' = \sup\{x < x_0 : a(x) = \alpha\}$.

Then $x_\alpha \le x' < x_0 < x_0 + \varepsilon$ and $a(x') = \alpha$. Since $a(x) \le \alpha$ for all $x \in [x',\ x_0 + \varepsilon)$, the function V satisfies (3.2) with the maximum there attained at $a = \alpha$. Therefore

$$V(x) = K_1 e^{r_+(\alpha)(x-x')} + K_2 e^{r_-(\alpha)(x-x')} \qquad \forall x \in [x',\ x_0 + \varepsilon). \tag{3.10}$$

From (3.10) and (3.3), the equation $a(x') = \alpha$ can be rewritten as

$$K_1 r_+(\alpha) = -K_2 r_-(\alpha) \frac{\mu + \alpha\sigma^2 r_-(\alpha)}{\mu + \alpha\sigma^2 r_+(\alpha)},$$

which establishes a relation between the constants K_1 and K_2. Using this relation, we calculate

(3.11)

$$\begin{aligned} a(x) &= \frac{-\mu V'(x)}{\sigma^2 V''(x)} \\ &= \frac{-\mu \left(e^{(r_+(\alpha)-r_-(\alpha))(x-x')} - \frac{\mu+\alpha\sigma^2 r_+(\alpha)}{\mu+\alpha\sigma^2 r_-(\alpha)} \right)}{\sigma^2 \left(r_+(\alpha) e^{(r_+(\alpha)-r_-(\alpha))(x-x')} - r_-(\alpha) \frac{\mu+\alpha\sigma^2 r_+(\alpha)}{\mu+\alpha\sigma^2 r_-(\alpha)} \right)} \qquad \forall x \in [x',\ x_0 + \varepsilon). \end{aligned}$$

However, we have $a(x) < \alpha$ for $x > x'$, which after a simple algebraic transformation of (3.11) is equivalent to $e^{(r_+(\alpha)-r_-(\alpha))(x-x')} < 1$. This leads to a contradiction. Therefore (3.9) holds. □

In view of $a(x_\alpha) = \alpha < \beta$ and (3.9), we have

$$\alpha \le a(x) < \beta$$

in the right neighborhood of x_α. Therefore

$$\phi(x,\ a(x)) = \max_{\alpha \le a \le \beta} \phi(x,\ a) = 0. \tag{3.12}$$

From (3.3), we have $V''(x) = -\frac{\mu V'(x)}{\sigma^2 a(x)}$. Substituting this expression for V'' into (3.12), we get

$$\mu a(x) V'(x)/2 = \gamma V(x). \tag{3.13}$$

Differentiating this equation and again using $V''(x) = -\frac{\mu V'(x)}{\sigma^2 a(x)}$, we arrive at

$$a'(x) = \frac{\mu^2 + 2\sigma^2\gamma}{\mu\sigma^2}.$$

Integrating this equation results in (recall that $a(x_\alpha) = \alpha$)

$$a(x) = \frac{\mu^2 + 2\sigma^2\gamma}{\mu\sigma^2}(x - x_\alpha) + \alpha. \tag{3.14}$$

Let

$$x_\beta = \frac{\mu\sigma^2}{\mu^2 + 2\sigma^2\gamma}(\beta - \alpha) + x_\alpha, \tag{3.15}$$

which is obtained by setting $a(x_\beta) = \beta$. Then $\alpha \le a(x) < \beta$ for all $x \in [x_\alpha,\ x_\beta)$. It follows from (3.13) that (noting $a(x_\alpha) = \alpha$)

$$\frac{V(x_\alpha)}{V'(x_\alpha)} = \frac{\mu\alpha}{2\gamma} \equiv \frac{y_\alpha}{(1-\Gamma)}, \tag{3.16}$$

where

$$0 < \Gamma \equiv \frac{\mu^2}{\mu^2 + 2\sigma^2\gamma} < 1, \qquad y_\alpha \equiv \frac{\mu\sigma^2\alpha}{\mu^2 + 2\sigma^2\gamma}. \tag{3.17}$$

Substituting the expression (3.14) for $a(x)$ into (3.3) and then solving the resulting equation for $V(x)$ on $[x_\alpha,\ x_\beta)$, while taking into account (3.16), we get

$$V(x) = \frac{\mu\alpha}{2\gamma} V'(x_\alpha) \left(\frac{x - x_\alpha + y_\alpha}{y_\alpha} \right)^{1-\Gamma}, \qquad x_\alpha \le x < x_\beta, \tag{3.18}$$

where the free constant $V'(x_\alpha)$ can be determined by

$$V'(x_\alpha) = k_1(\alpha,\beta)(r_+(\alpha)e^{r_+(\alpha)x_\alpha} - r_-(\alpha)e^{r_-(\alpha)x_\alpha}), \tag{3.19}$$

in view of (3.6) and a smooth fit at $x = x_\alpha$. Straightforward computations show that the function V defined by (3.6) and (3.18) is continuous with continuous first and second derivatives at x_α.

So far, we have obtained the forms of V on two intervals, $[0, x_\alpha)$ and $[x_\alpha, x_\beta)$, by (3.6) and (3.18), respectively. Now we proceed to the interval beyond x_β. To this end, we first have the following.

PROPOSITION 5. *For all $x \ge x_\beta$*

$$a(x) \ge \beta.$$

Proof. By virtue of Proposition 4, $a(x) \ge \alpha$ for all $x \ge x_\beta$. Suppose there exists $x' > x_\beta$ such that $a(x') < \beta$. Then there exists $\varepsilon > 0$ such that $a(x) < \beta$ for all $x < x' + \varepsilon$. Let $\bar{x} = \sup\{x < x' \ :\ a(x) = \beta\}$. Then $x_\beta \le \bar{x} < x'$ and $a(\bar{x}) = \beta$. In addition, $\alpha \le a(x) < \beta$ for all $\bar{x} < x \le x'$. Repeating the same arguments as in the proof of (3.9) we get $a(x) = \frac{\mu^2 + 2\sigma^2\gamma}{\mu\sigma^2}(x - \bar{x}) + \beta > \beta$ for each $x > \bar{x}$, which is a contradiction. □

The above proposition implies that the maximum in (3.2) is obtained at $a = \beta$ for $x \ge x_\beta$. The resulting equation of (3.2) then becomes a second-order linear ODE, whose solution is of the form

$$V(x) = k_1(\beta)e^{r_+(\beta)(x - x_1)} + k_2(\beta)e^{r_-(\beta)(x-x_1)}, \qquad x_\beta \le x < x_1, \tag{3.20}$$

where $x_1 > x_\beta$, as defined earlier, is also the first point such that $V''(x_1) = 0$ (see Proposition 6 below).

PROPOSITION 6. *Let $x_1 > x_\beta$ be the first point where V'' vanishes. Then $V'(x_1) = 1$.*

Proof. Suppose $V'(x_1) > 1$. Then there exists $\varepsilon > 0$ such that $V'(x) > 1$ for all $x_1 \le x \le x_1 + \varepsilon$. Therefore on the interval $[x_1, x_1 + \varepsilon]$ the function V satisfies (3.2). In view of Proposition 5, $a(x) \ge \beta$, and hence on the interval $[x_1, x_1 + \varepsilon]$, V is of the form given by (3.20). Since $V''(x_1) = 0$, we conclude $k_1(\beta)r_+^2(\beta) = -k_2(\beta)r_-^2(\beta) > 0$. (The positivity of $k_1(\beta)$ follows from $V' > 0$.) Thus $V''(x) =$

$k_1(\beta)r_+^2(\beta)\left(e^{(r_+(\beta)-r_-(\beta))(x-x_1)}-1\right)$. This expression is positive for each $x > x_1$. This contradicts the concavity of V. □

The following corollary is straightforward in view of the above proposition and the inequality $V'(x) \geq 1$.

COROLLARY 2. *Under the assumption of Proposition* 6,

$$V'(x) = 1 \qquad \forall x \geq x_1.$$

The analysis so far shows that the function V is of the following form:

$$V(x) = \begin{cases} k_1(\alpha,\ \beta)(e^{r_+(\alpha)x} - e^{r_-(\alpha)x}), & 0 \leq x < x_\alpha, \\ \frac{\mu\alpha}{2\gamma}V'(x_\alpha)\left(\frac{x-x_\alpha+y_\alpha}{y_\alpha}\right)^{1-\Gamma}, & x_\alpha \leq x < x_\beta, \\ k_1(\beta)e^{r_+(\beta)(x-x_1)} + k_2(\beta)e^{r_-(\beta)(x-x_1)}, & x_\beta \leq x < x_1, \\ k_1(\beta) + k_2(\beta) + x - x_1, & x_1 \leq x, \end{cases} \tag{3.21}$$

where $r_+(\alpha)$, $r_-(\alpha)$, $r_+(\beta)$, $r_-(\beta)$, x_α and x_β, and Γ and y_α are given by (3.7), (3.8), (3.15), and (3.17), respectively.

The next step is to determine the remaining constants in (3.21). To do so we use the principle of smooth fit at the points x_β and x_1. Namely, we have to choose the unknown constants $k_1(\beta), k_2(\beta), k_1(\alpha,\beta)$, and x_1 in such a way that the function V and its first and second derivatives are continuous at these points. To this end, first for V of the form (3.20) the condition

$$V'(x_1) = 1, \qquad V''(x_1) = 0$$

can be written as

$$k_1(\beta)r_+(\beta) + k_2(\beta)r_-(\beta) = 1, \qquad k_1(\beta)r_+^2(\beta) + k_2(\beta)r_-^2(\beta) = 0.$$

As a result,

$$k_1(\beta) = \frac{-r_-(\beta)}{r_+(\beta)(r_+(\beta)-r_-(\beta))}, \qquad k_2(\beta) = \frac{r_+(\beta)}{r_-(\beta)(r_+(\beta)-r_-(\beta))}. \tag{3.22}$$

Next, let $\Delta = x_\beta - x_1$; then we can calculate V' and V'' at x_β as

$$\begin{aligned} V'(x_\beta) &= k_1(\beta)r_+(\beta)e^{r_+(\beta)\Delta} + k_2(\beta)r_-(\beta)e^{r_-(\beta)\Delta}, \\ V''(x_\beta) &= k_1(\beta)r_+^2(\beta)e^{r_+(\beta)\Delta} + k_2(\beta)r_-^2(\beta)e^{r_-(\beta)\Delta}. \end{aligned} \tag{3.23}$$

Recall that $V''(x_\beta) = \frac{-\mu V'(x_\beta)}{\sigma^2 a(x_\beta)} = \frac{-\mu V'(x_\beta)}{\sigma^2\beta}$, which results in

$$x_\beta - x_1 \equiv \Delta = \frac{\beta\sigma^2}{[\mu^2+2\sigma^2\gamma]^{1/2}} \log\left(\frac{-\mu+[\mu^2+2\sigma^2\gamma]^{1/2}}{\mu+[\mu^2+2\sigma^2\gamma]^{1/2}}\right) < 0. \tag{3.24}$$

On the other hand, a smooth fit, in terms of $V'(x_\beta)$, for (3.18) and (3.23) with (3.19) taken into consideration yields

$$\begin{aligned} &\frac{\alpha\mu}{2\gamma y_\alpha}k_1(\alpha,\ \beta)\left(r_+(\alpha)e^{r_+(\alpha)x_\alpha} - r_-(\alpha)e^{r_-(\alpha)x_\alpha}\right)\left(\frac{\beta}{\alpha}\right)^{-\Gamma} \\ &\quad = k_1(\beta)r_+(\beta)e^{r_+(\beta)\Delta} + k_2(\beta)r_-(\beta)e^{r_-(\beta)\Delta}. \end{aligned} \tag{3.25}$$

Formulas (3.24) and (3.25) determine x_1 and $k_1(\alpha,\beta)$. Note that from (3.22) and (3.24) it follows that

$$\frac{V(x_\beta)}{V'(x_\beta)} = \frac{k_1(\beta)e^{r_+(\beta)\Delta} + k_2(\beta)e^{r_-(\beta)\Delta}}{k_1(\beta)r_+(\beta)e^{r_+(\beta)\Delta} + k_2(\beta)r_-(\beta)e^{r_-(\beta)\Delta}} = \frac{e^{(r_+(\beta)-r_-(\beta))\Delta} - \frac{r_-(\beta)}{r_+(\beta)}}{r_+(\beta)e^{(r_+(\beta)-r_-(\beta))\Delta} - r_-(\beta)} = \frac{\mu\beta}{2\gamma}.$$

Therefore $\lim_{x\to x_\beta,\ x<x_\beta} V(x) = \frac{\mu\alpha}{2\gamma}V'(x_\alpha)(\frac{\beta}{\alpha})^{1-\Gamma} = \frac{\mu\beta}{2\gamma}V'(x_\beta) = V(x_\beta)$, which proves the continuity of V at x_β.

These calculations enable us to formulate the main result of this section.

THEOREM 4. *Let $r_+(\alpha)$, $r_-(\alpha)$, $r_+(\beta)$ and $r_-(\beta)$, Γ and y_α, x_α, x_β, x_1, $k_1(\beta)$ and $k_2(\beta)$, and $k_1(\alpha,\ \beta)$ be given by* (3.7), (3.17), (3.8), (3.15), (3.24), (3.22), *and* (3.25), *respectively. Then $V(x)$ given by* (3.21) *is a concave, twice continuously differentiable solution of the HJB equation* (3.1).

Proof. From the way we constructed V, it must be a twice continuously differentiable solution to the HJB equation (3.1). What remains to show is the concavity. From (3.21), we deduce that

$$V'''(x) = k_1(\alpha,\ \beta)\left(r_+^3(\alpha)e^{r_+(\alpha)x} - r_-^3(\alpha)e^{r_-(\alpha)x}\right) > 0 \qquad \forall 0 \le x < x_\alpha,$$

due to $r_-(\alpha) < 0 < k_1(\alpha,\ \beta)$. Hence on this interval V'' is increasing and

$$V''(x) < V''(x_\alpha) = k_1(\alpha,\ \beta)\left(r_+^2(\alpha)e^{r_+(\alpha)x_\alpha} - r_-^2(\alpha)e^{r_-(\alpha)x_\alpha}\right) < 0,$$

due to $\frac{r_-(\alpha)}{r_+(\alpha)} = e^{(r_+(\alpha)-r_-(\alpha))x_\alpha}$ and $|r_-(\alpha)| > r_+(\alpha)$.

For $x_\alpha \le x < x_\beta$, $V''(x) = \frac{-\mu V'(x)}{\sigma^2 a(x)} < 0$. For $x_\beta \le x <\ x_1$,

$$V'''(x) = k_1(\beta)r_+^3(\beta)e^{r_+(\beta)(x-x_1)} + k_2(\beta)r_-^3(\beta)e^{r_-(\beta)(x-x_1)} > 0,$$

since $k_2(\beta)$ and $r_-(\beta)$ are of the same signs. Thus $V''(x) < V''(x_1) = 0$ for all $x_\beta \le x <\ x_1$. Finally, $V''(x) = 0$ for all $x \ge x_1$. This establishes the concavity of V. ∎

4. Case with nonzero liability. This section deals with the general model (2.1) where $\delta > 0$. In this case the HJB equation is given by (2.30). Again we are looking for a smooth concave function that solves this equation. As before, suppose that such a solution V exists and consider $x_1 = \inf\{x \ge 0\ :\ V'(x) \le 1\}$. Then it is obvious that $x_1 = 0$ if and only if $V(x) = x$ for all $x \ge 0$. Our first step is to characterize the existence of such a trivial solution to (2.30).

THEOREM 5. *$V(x) = x$ for all $x \ge 0$ if and only if*

$$\beta\mu \le \delta. \tag{4.1}$$

Proof. Suppose $V(x) = x$ for each $x \ge 0$. Then in view of (2.30)

$$\max_{\alpha\le a\le\beta}\left(\frac{1}{2}\sigma^2a^2V''(0) + (a\mu-\delta)V'(0) - \gamma V(0)\right) \equiv \beta\mu - \delta \le 0.$$

Conversely, if $\beta\mu \le \delta$, then due to concavity

$$\max_{\alpha\le a\le\beta}\left(\frac{1}{2}\sigma^2a^2V''(x) + (a\mu-\delta)V'(x) - \gamma V(x)\right) \le -\gamma V(x) < 0 \qquad \forall x > 0.$$

Thus, (2.30) is satisfied only if $V'(x) = 1$ for all $x > 0$. □

In the rest of this section we assume $\beta\mu > \delta$. In view of (2.30)

$$0 = \max_{\alpha \le a \le \beta} \left(\frac{1}{2}\sigma^2 a^2 V''(x) + (a\mu - \delta)V'(x) - \gamma V(x) \right) \qquad \forall x < x_1. \tag{4.2}$$

For each $x \ge 0$ and $a \ge 0$ define

$$\phi(x,\ a) = \frac{1}{2}\sigma^2 a^2 V''(x) + (a\mu - \delta)V'(x) - \gamma V(x). \tag{4.3}$$

The maximizer of the function $\phi(x,\ a)$ over $a \ge 0$ is given by

$$a(x) = -\frac{\mu V'(x)}{\sigma^2 V''(x)} > 0, \qquad x \ge 0. \tag{4.4}$$

Following the same scheme as in the no-liability case, we will prove that there exist $x_\alpha \le x_\beta < x_1$ such that $a(x) \le \alpha$ for all $x \le x_\alpha$, and $a(x) \ge \beta$ for all $x \ge x_\beta$, and the function $a(x)$ increases from α to β on the interval $[x_\alpha,\ x_\beta]$.

As before we start with analyzing $a(0)$.

PROPOSITION 7. *Let $\beta\mu > \delta$. Then*

(i) $\frac{2\delta}{\mu} < \alpha$ *if and only if* $a(0) < \alpha$. *In this case* $a(0) = \frac{\mu\alpha^2}{2(\mu\alpha - \delta)}$.

(ii) $\alpha \le \frac{2\delta}{\mu} < \beta$ *if and only if* $\alpha \le a(0) < \beta$. *In this case* $a(0) = 2\frac{\delta}{\mu}$.

(iii) $\beta \le \frac{2\delta}{\mu}$ *if and only if* $a(0) \ge \beta$. *In this case* $a(0) = \frac{\mu\beta^2}{2(\mu\beta - \delta)}$.

Proof. Let $\tilde{a} \in [\alpha,\ \beta]$ be such that

$$0 = \max_{\alpha \le a \le \beta} \left(\frac{1}{2}\sigma^2 a^2 V''(0) + (a\mu - \delta)V'(0) \right) = \frac{1}{2}\sigma^2 \tilde{a}^2 V''(0) + (\tilde{a}\mu - \delta)V'(0). \tag{4.5}$$

Comparing (4.5) with (4.4) we obtain

$$\tilde{a}^2 - 2a(0)\tilde{a} + \frac{2\delta}{\mu}a(0) = 0. \tag{4.6}$$

From (4.6), it follows that $a(0) \ge \frac{2\delta}{\mu}$. Moreover, by definition, $a(0) \in [\alpha,\ \beta]$ is equivalent to $\tilde{a} = a(0)$, which is further equivalent to $a(0) = \frac{2\delta}{\mu} \in [\alpha,\ \beta]$. Thus we conclude:

(i) If $a(0) < \alpha$, then $\frac{2\delta}{\mu} \le a(0) < \alpha$. Conversely, suppose $\frac{2\delta}{\mu} < \alpha$. If $a(0) \in [\alpha,\ \beta]$, then by the above $a(0) = \frac{2\delta}{\mu} < \alpha$, which is a contradiction. Thus either $a(0) < \alpha$ or $a(0) > \beta$. Suppose $a(0) > \beta$; then $\tilde{a} = \beta$ and by (4.6), $a(0) = \frac{\mu\beta^2}{2(\mu\beta - \delta)} < \beta$ (due to $\frac{2\delta}{\mu} < \alpha < \beta$). This is again a contradiction. Hence we have $a(0) < \alpha$. Then $\tilde{a} = \alpha$ and in view of (4.6), we get $a(0) = \frac{\mu\alpha^2}{2(\alpha\mu - \delta)}$.

(ii) Suppose $\alpha \le \frac{2\delta}{\mu} < \beta$. Then due to (i) we have $a(0) \ge \alpha$. Now we proceed to prove that $a(0) \le \frac{2\delta}{\mu} < \beta$. Suppose $a(0) > \frac{2\delta}{\mu}$. Then $a(0) > \beta \equiv \tilde{a}$. On the other hand, in view of (4.6) we have $a(0) = \frac{\mu\beta^2}{2(\beta\mu - \delta)}$; thus $\frac{\mu\beta^2}{2(\beta\mu - \delta)} \ge \beta$, which is equivalent to $2\frac{\delta}{\mu} \ge \beta$. This, however, is a contradiction and therefore $a(0) = \frac{2\delta}{\mu} \in [\alpha,\ \beta)$. Conversely, if $a(0) \in [\alpha,\ \beta)$, then $a(0) = \frac{2\delta}{\mu} \in [\alpha,\ \beta)$.

(iii) Suppose $\beta \le \frac{2\delta}{\mu}$. Then $a(0) \ge \frac{2\delta}{\mu} \ge \beta$, leading to $\tilde{a} = \beta$ and $a(0) = \frac{\mu\beta^2}{2(\beta\mu - \delta)} \ge \beta$. Conversely, if $a(0) \ge \beta$, then $\tilde{a} = \beta$ and $a(0) = \frac{\mu\beta^2}{2(\mu\beta - \delta)} \ge \beta$, which is equivalent to $\frac{2\delta}{\mu} \ge \beta$. □

As it will be seen in what follows, the structure of the solution to our original optimization problem depends on three cases specified by (i), (ii), and (iii) above. Accordingly in the rest of the section we will analyze these three cases.

4.1. Case of $\frac{2\delta}{\mu} < \alpha$. We begin our analysis with an observation that in this case, in view of Proposition 7(i), $a(x) < \alpha$ for all x in the right neighborhood of 0. Substituting $a = \alpha$ in (4.2) and solving the resulting second-order linear ODE, we obtain

$$V(x) = k_1(\alpha,\ \beta)(e^{r_+(\alpha)x} - e^{r_-(\alpha)x}), \tag{4.7}$$

where $k_1(\alpha,\ \beta)$ is a free constant to be determined, and

$$\begin{aligned} r_+(z) &= \frac{-(z\mu-\delta) + [(z\mu-\delta)^2 + 2\sigma^2 z^2\gamma]^{1/2}}{\sigma^2 z^2}, \\ r_-(z) &= \frac{-(z\mu-\delta) - [(z\mu-\delta)^2 + 2\sigma^2 z^2\gamma]^{1/2}}{\sigma^2 z^2}, \qquad z > 0. \end{aligned} \tag{4.8}$$

Due to (4.4) and (4.7),

$$\begin{aligned} a'(x) &= \frac{-\mu}{\sigma^2}\frac{(V''(x))^2 - V'(x)V^{(3)}(x)}{(V''(x))^2} \\ &= \frac{-\mu r_+(\alpha)r_-(\alpha)e^{(r_+(\alpha)+r_-(\alpha))x}\left(r_+(\alpha) - r_-(\alpha)\right)^2}{\sigma^2(V''(x))^2} > 0 \end{aligned}$$

for each x in the right neighborhood of 0. Therefore $a(x)$ increases and reaches α at the point x_α given by

$$x_\alpha = \frac{1}{r_+(\alpha) - r_-(\alpha)} \log\left(\frac{r_-(\alpha)\left(\mu + \alpha\sigma^2 r_-(\alpha)\right)}{r_+(\alpha)\left(\mu + \alpha\sigma^2 r_+(\alpha)\right)}\right) > 0. \tag{4.9}$$

PROPOSITION 8. *For each $x \in [x_\alpha, x_1]$,*

$$a(x) \geq \alpha.$$

Proof. Suppose there exists $x' > x_\alpha$ such that $a(x) < \alpha$ and let $\bar{x} = \sup\{x < x' \ :\ a(x) = \alpha\}$. Then $x_\alpha \leq \bar{x} < x'$, $a(\bar{x}) = \alpha$, and $a(x) < \alpha$ for $\bar{x} < x \leq x'$. Substituting $a = \alpha$ into (4.2) and solving the resulting second-order linear ODE, we get $V(x) = k_1 e^{r_+(\alpha)(x-\bar{x})} + k_2 e^{r_-(\alpha)(x-\bar{x})}$. Therefore

$$a(x) = \frac{-\mu V'(x)}{\sigma^2 V''(x)} = -\frac{\mu}{\sigma^2}\frac{k_1 r_+(\alpha)e^{(r_+(\alpha)-r_-(\alpha))(x-\bar{x})} + k_2 r_-(\alpha)}{k_1 r_+^2(\alpha)e^{(r_+(\alpha)-r_-(\alpha))(x-\bar{x})} + k_2 r_-^2(\alpha)}, \qquad \bar{x} < x \leq x'.$$

Since $a(\bar{x}) = \alpha$, we have

$$k_1 r_+(\alpha)\left(1 + \frac{\alpha\sigma^2 r_+(\alpha)}{\mu}\right) = -k_2 r_+(\alpha)\left(1 + \frac{\alpha\sigma^2 r_-(\alpha)}{\mu}\right).$$

Thus, for $\bar{x} < x \leq x'$ the inequality $a(x) < \alpha$ is equivalent to $e^{(r_+(\alpha)-r_-(\alpha))(x-\bar{x})} < 1$, which is a contradiction. □

By virtue of Proposition 8, $\alpha \leq a(x) < \beta$ in the right neighborhood of x_α. In this case we have

$$\phi(x,\ a(x)) = \max_{\alpha \leq a \leq \beta} \phi(x,\ a) = 0. \tag{4.10}$$

Substituting

$$V''(x) = \frac{-\mu V'(x)}{\sigma^2 a(x)} \tag{4.11}$$

into (4.10), differentiating the resulting equation, and substituting $V''(x) = \frac{-\mu V'(x)}{\sigma^2 a(x)}$ once more, we arrive at $\frac{\mu a'(x)}{2} + \frac{\mu\delta}{\sigma^2 a(x)} = \frac{\mu^2+2\gamma\sigma^2}{2\sigma^2}$. As a result

$$a'(x) = \frac{\mu^2 + 2\gamma\sigma^2}{\mu\sigma^2}\left(1 - \frac{c}{a(x)}\right) \tag{4.12}$$

with

$$c \equiv 2\delta\mu/(\mu^2 + 2\gamma\sigma^2). \tag{4.13}$$

Integrating (4.12), we get $G(a(x)) \equiv \frac{\mu^2+2\gamma\sigma^2}{\mu\sigma^2}(x - x_\alpha) + G(\alpha)$, where

$$G(u) = u + c\log(u - c). \tag{4.14}$$

Therefore

$$a(x) = G^{-1}\left(\frac{\mu^2 + 2\gamma\sigma^2}{\mu\sigma^2}(x - x_\alpha) + G(\alpha)\right). \tag{4.15}$$

Thus $a(x)$ is increasing and $a(x_\beta) = \beta$ for

$$x_\beta \equiv \frac{\mu\sigma^2}{\mu^2 + 2\gamma\sigma^2}[G(\beta) - G(\alpha)] + x_\alpha = \frac{\mu\sigma^2}{\mu^2 + 2\gamma\sigma^2}(\beta - \alpha) + \frac{\mu\sigma^2 c}{\mu^2 + 2\gamma\sigma^2}\log\left(\frac{\beta - c}{\alpha - c}\right). \tag{4.16}$$

Solving (4.11), we obtain

$$V(x) = V(x_\alpha) + V'(x_\alpha)\int_{x_\alpha}^{x} \exp\left(-\frac{\mu}{\sigma^2}\int_{x_\alpha}^{y}\frac{du}{a(u)}\right)dy, \qquad x_\alpha \le x < x_\beta, \tag{4.17}$$

where $V(x_\alpha)$ and $V'(x_\alpha)$ are free constants. Choosing $V(x_\alpha)$ and $V'(x_\alpha)$ as the value and the derivative, respectively, of the right-hand side of (4.7) at x_α, we can ensure that the function V given by (4.7) and (4.17) is continuous with its first and second derivatives at the point x_α no matter what the choice of $k(\alpha, \beta)$ is. (Note that due to the HJB equation, continuity of V and its first derivative at x_α automatically implies continuity of the second derivative as well.)

Next we are to simplify (4.17). First, changing variables $a(u) = \theta$ we get

$$\begin{aligned}\int_{x_\alpha}^{x} &\exp\left(-\frac{\mu}{\sigma^2}\int_{x_\alpha}^{y}\frac{du}{a(u)}dy\right)\\ &= \frac{\mu\sigma^2}{\mu^2 + 2\gamma\sigma^2}\int_{\alpha}^{a(x)}\left(1 + \frac{c}{\theta - c}\right)\left(\frac{\theta - c}{\alpha - c}\right)^{-\Gamma}d\theta, \qquad x_\alpha \le x < x_\beta.\end{aligned}$$

On the other hand, relations (4.9) and (4.7) imply

$$V(x_\alpha) = \frac{\alpha\mu - 2\delta}{2\gamma}V'(x_\alpha).$$

Simple algebraic transformations yield

$$\left(\frac{\mu\sigma^2}{\mu^2+2\gamma\sigma^2}\right)\left(\frac{c}{\Gamma}-\frac{z-c}{1-\Gamma}\right)=\frac{z\mu-2\delta}{2\gamma}\qquad \forall z>0, \tag{4.18}$$

where c is given by (4.13) and

$$\Gamma=\frac{\mu^2}{\mu^2+2\gamma\sigma^2}. \tag{4.19}$$

Therefore

$$V(x)=V'(x_\alpha)\frac{\mu a(x)-2\delta}{2\gamma}\left(\frac{a(x)-c}{\alpha-c}\right)^{-\Gamma},\qquad x_\alpha\le x<x_\beta. \tag{4.20}$$

Now, we proceed to the next piece of V on the interval beyond x_β.

PROPOSITION 9. *For each $x\in[x_\beta,x_1]$,*

$$a(x)\ge\beta.$$

Proof. Suppose that there exists $x'>x_\beta$ such that $a(x)<\beta$. Since $x'\ge x_\alpha$, we have $\beta>a(x)\ge\alpha$. Denote $\bar{x}=\sup\{x<x' \ : \ a(x)=\beta\}$. Then $x_\beta\le\bar{x}<x'$, $a(\bar{x})=\beta$, and $\alpha\le a(x)<\beta$ for $\bar{x}<x\le x'$. Thus $a(x)$ satisfies (4.12) for $\bar{x}<x\le x'$ and

$$a(x)=G^{-1}\left(\frac{\mu^2+2\gamma\sigma^2}{\mu\sigma^2}(x-\bar{x})+G(\beta)\right)>\beta.$$

This is a contradiction. □

In view of the above proposition,

$$\phi(x,\ \beta)=\max_{\alpha\le a\le\beta}\phi(x,\ a)=0,\qquad x_\beta\le x<x_1, \tag{4.21}$$

where x_1, which is defined earlier, is also the first point such that $V''(x_1)=0$. This results in

$$V(x)=k_1(\beta)e^{r_+(\beta)(x-x_1)}+k_2(\beta)e^{r_-(\beta)(x-x_1)},\qquad x_\beta\le x<x_1, \tag{4.22}$$

where $k_1(\beta)$ and $k_2(\beta)$ are two free constants also to be determined.

PROPOSITION 10. *For $x\ge x_1$,*

$$V'(x)=1.$$

Proof. Suppose that there exists $x'\ge x_1$ such that $V''(x')<0$. Since $x'\ge x_\beta$, inequality $a(x)\ge\beta$ holds. Let $\bar{x}=\sup\{x<x' \ : \ V''(x)=0\}$. For any $x\in(\bar{x},\ x']$, we have $V(x)=K_1'e^{r_+(\beta)(x-\bar{x})}+K_2'e^{r_-(\beta)(x-\bar{x})}$. Equality $V''(\bar{x})=0$ results in $0<K_1'r_+^2(\beta)=-K_2'r_-^2(\beta)$. Consequently $V''(x)=K_1'r_+^2(\beta)e^{r_+(\beta)(x-\bar{x})}+K_2'r_-^2(\beta)e^{r_-(\beta)(x-\bar{x})}=K_1'r_+^2(\beta)\left(e^{r_+(\beta)(x-\bar{x})}-e^{r_-(\beta)(x-\bar{x})}\right)>0$. This contradicts the concavity of the function V. □

From Proposition 10 it follows that

$$V(x)=x-x_1+k_1(\beta)+k_2(\beta),\qquad x\ge x_1.$$

To compute the free constants $k_1(\beta)$ and $k_2(\beta)$, we use the relationship

$$V'(x_1) = 1, \qquad V''(x_1) = 0.$$

From (4.22) it follows that

$$k_1(\beta)r_+(\beta) + k_2(\beta)r_-(\beta) = 1, \qquad k_1(\beta)r_+^2(\beta) + k_2(\beta)r_-^2(\beta) = 0.$$

As a result

$$k_1(\beta) = \frac{-r_-(\beta)}{r_+(\beta)(r_+(\beta) - r_-(\beta))} > 0, \tag{4.23}$$
$$k_2(\beta) = \frac{r_+(\beta)}{r_-(\beta)(r_+(\beta) - r_-(\beta))} < 0.$$

To determine the remaining unknown constants we apply the principle of smooth fit at the point x_β. Let $\Delta = x_\beta - x_1$. By (4.22) we have

$$V'(x_\beta) = k_1(\beta)r_+(\beta)e^{r_+(\beta)\Delta} + k_2(\beta)r_-(\beta)e^{r_-(\beta)\Delta}, \tag{4.24}$$
$$V''(x_\beta) = k_1(\beta)r_+^2(\beta)e^{r_+(\beta)\Delta} + k_2(\beta)r_-^2(\beta)e^{r_-(\beta)\Delta}.$$

However, the relation (4.11) (recall that $a(x_\beta) = \beta$) yields $V''(x_\beta) = \frac{-\mu V'(x_\beta)}{\sigma^2\beta}$. This leads to

$$x_\beta - x_1 = \Delta = \frac{1}{r_+(\beta) - r_-(\beta)} \log\left(\frac{\frac{1}{r_-(\beta)} + \frac{\sigma^2\beta}{\mu}}{\frac{1}{r_+(\beta)} + \frac{\sigma^2\beta}{\mu}}\right) < 0, \tag{4.25}$$

which determines x_1 and, in turn, determines $V'(x_\beta)$ via (4.24). To proceed, simple but tedious algebraic transformations show that from (4.20) and (4.17) it follows that

$$V'(x_\alpha) = V'(x_\beta)\left(\frac{\beta - c}{\alpha - c}\right)^\Gamma. \tag{4.26}$$

As in the previous section this implies the continuity of V at x_β. Finally, the continuity of V at x_α gives rise to

$$k_1(\alpha,\ \beta) = \frac{V'(x_\beta)\left(\frac{\beta-c}{\alpha-c}\right)^\Gamma}{r_+(\alpha)e^{r_+(\alpha)x_\alpha} - r_-(\alpha)e^{r_-(\alpha)x_\alpha}}. \tag{4.27}$$

This enables us to establish the main result of this section.

THEOREM 6. *Suppose* $\frac{2\delta}{\mu} < \alpha$. *Let* $k_1(\alpha,\ \beta)$, $r_+(\alpha)$, $r_-(\alpha)$, $r_+(\beta)$, $r_-(\beta)$, x_α, x_β, x_1 $k_1(\beta)$, $k_2(\beta)$, $a(x)$, c, Γ, *and* $V'(x_\alpha)$ *be given by* (4.27), (4.8), (4.9), (4.16), (4.25), (4.23), (4.15), (4.13), (4.19), *and* (4.26), *respectively. Then*

$$V(x) = \begin{cases} k_1(\alpha,\ \beta)\left(e^{r_+(\alpha)x} - e^{r_-(\alpha)x}\right), & 0 \le x < x_\alpha, \\ V'(x_\alpha)\frac{\mu a(x) - 2\delta}{2\gamma}\left(\frac{a(x)-c}{\alpha-c}\right)^{-\Gamma}, & x_\alpha \le x < x_\beta, \\ k_1(\beta)e^{r_+(\beta)(x-x_1)} + k_2(\beta)e^{r_-(\beta)(x-x_1)}, & x_\beta \le x < x_1, \\ k_1(\beta) + k_2(\beta) + x - x_1, & x \ge x_1, \end{cases} \tag{4.28}$$

is a concave, twice continuously differentiable solution of the HJB equation (2.30).

Proof. As before we need only show the concavity. To do this consider V'''. From (4.28), we get

$$\begin{aligned}
V'''(x) &= k_1(\alpha,\ \beta)\left(r_+^3(\alpha)e^{r_+(\alpha)x} - r_-^3(\alpha)e^{r_-(\alpha)x}\right) > 0, && 0 \le x < x_\alpha,\\
V'''(x) &= \tfrac{-\mu V''(x)}{\sigma^2 a(x)} + \tfrac{\mu a'(x)V'(x)}{\sigma^2(a(x))^2} > 0, && x_\alpha \le x < x_\beta,\\
V'''(x) &= k_1 r_+^3(\beta)e^{r_+(\beta)(x-x_1)} + k_2(\beta)r_-^3(\beta)e^{r_-(\beta)(x-x_1)} > 0, && x_\beta \le x < x_1.
\end{aligned}$$

Thus $V''(x) < V''(x_1) = 0$ for each $x < x_1$. On the other hand, $V''(x) = 0$ for each $x \ge x_1$. These lead to the concavity of V. □

4.2. Case of $\alpha \le \frac{2\delta}{\mu} < \beta$. Applying Propositions 7 and 8, we see that in this case $a(0) = \frac{2\delta}{\mu} \ge \alpha$ and $a(x) \ge \alpha$ for all $x \ge 0$. Then in the right neighborhood of 0, $\alpha \le a(x) < \beta$. It follows that for ϕ given by (4.3), equation (4.10) holds. Proceeding as in section 4.1, we see that $a(x)$ satisfies (4.12). Therefore

$$a(x) = G^{-1}\left(\frac{\mu^2 + 2\gamma\sigma^2}{\mu\sigma^2}x + G(2\delta/\mu)\right) \in [2\delta/\mu,\ \ \infty), \tag{4.29}$$

where G is given by (4.14). As a result $a(x)$ increases and $a(x_\beta) = \beta$, where

$$\begin{aligned}
x_\beta &= \frac{\mu\sigma^2}{\mu^2 + 2\gamma\sigma^2}[G(\beta) - G(2\delta/\mu)]\\
&= \frac{\mu\sigma^2}{\mu^2 + 2\gamma\sigma^2}(\beta - 2\delta/\mu) + \frac{2\delta\mu c}{\mu^2 + 2\gamma\sigma^2}\log\left(\frac{\beta - c}{2\delta/\mu - c}\right).
\end{aligned} \tag{4.30}$$

Integrating (4.4), we get

$$V(x) = V'(0)\frac{\mu a(x) - 2\delta}{2\gamma}\left(\frac{a(x) - c}{2\delta/\mu - c}\right)^{-\Gamma}, \qquad 0 \le x < x_\beta. \tag{4.31}$$

By virtue of Proposition 9 we have $a(x) \ge \beta$ for $x \in [x_\beta, x_1]$. Let x_1 be such that $V''(x_1) = 0$. Then for $x_\beta \le x < x_1$,

$$V(x) = k_1(\beta)e^{r_+(\beta)(x-x_1)} + k_2(\beta)e^{r_-(\beta)(x-x_1)}. \tag{4.32}$$

Using the principle of smooth fit for V in (4.32) at x_1, we see that $k_1(\beta)$ and $k_2(\beta)$ are given by (4.23). Put $\Delta = x_\beta - x_1$. Applying the principle of smooth fit at x_β for V' and V'', we deduce that Δ is given by (4.25). Therefore

$$x_1 = x_\beta + \Delta = x_\beta + \frac{1}{r_+(\beta) - r_-(\beta)}\log\left(\frac{\frac{1}{r_-(\beta)} + \frac{\sigma^2\beta}{\mu}}{\frac{1}{r_+(\beta)} + \frac{\sigma^2\beta}{\mu}}\right). \tag{4.33}$$

THEOREM 7. *Suppose $\alpha \le \frac{2\delta}{\mu} < \beta$. Let $a(x)$, c, Γ, x_β, x_1, $r_+(\beta)$, $r_-(\beta)$, $k_1(\beta)$, and $k_2(\beta)$ be given by* (4.29), (4.13), (4.19), (4.30), (4.33), (4.8), *and* (4.23), *respectively. Let $V'(0)$ be determined from* (4.26), *in which x_α and α are replaced by 0 and $2\delta/\mu$, respectively. Then*

$$V(x) = \begin{cases} V'(0)\frac{\mu a(x) - 2\delta}{2\gamma}\left(\frac{a(x) - c}{2\delta/\mu - c}\right)^{-\Gamma}, & 0 \le x < x_\beta,\\ k_1(\beta)e^{r_+(\beta)(x-x_1)} + k_2(\beta)e^{r_-(\beta)(x-x_1)}, & x_\beta \le x < x_1,\\ k_1(\beta) + k_2(\beta) + x - x_1, & x \ge x_1, \end{cases} \tag{4.34}$$

is a concave, twice continuously differentiable solution of the HJB equation (2.30).

Proof. The proof of this theorem is similar to that of Theorem 6. (One needs only to repeat the proof of Theorem 6, substituting x_α, x_β, and x_1 by 0, x_β, and x_1, respectively.) □

Remarks 1. Denote by $V_{\alpha,\,\beta}(x)$ the concave solution to the HJB equation (2.30) corresponding to the parameters (α, β). Then the results of this section show that

$$V_{\alpha,\,\beta}(x) = V_{\frac{2\delta}{\mu},\,\beta}(x), \qquad x \geq 0,$$

for each $\alpha \leq \frac{2\delta}{\mu}$. One can verify that as $\alpha \to 0+$ and $\beta = 1$, the expression (4.34) becomes the value function for the problem with $\alpha = 0$, $\beta = 1$, even though our methodology cannot be applied in the case of $\alpha = 0$.

It is interesting to notice that in the case of $\alpha = 0$, $\beta = 1$, and $\frac{2\delta}{\mu} < 1$ a different approach used in [18] yields

$$V_{TZ}(x) = \begin{cases} \int_0^x X^{-1}(y)dy, & 0 \leq x < x_\beta, \\ k_1(\beta)e^{r_+(\beta)(x-x_1)} + k_2(\beta)e^{r_-(\beta)(x-x_1)}, & x_\beta \leq x < x_1, \\ k_1(\beta) + k_2(\beta) + x - x_1, & x \geq x_1, \end{cases} \tag{4.35}$$

where $X(z) = Cz^{-1-\frac{2\gamma\sigma^2}{\mu^2}} + C_1 - \frac{\delta}{\mu^2/(2\sigma^2)+\gamma}\ln z$ for some constants C and C_1. The expression for the case $0 \leq x < x_\beta$ in (4.35) is very different from that in (4.34). However, the following result shows that they are in fact the same.

Proposition 11.

$$V(x) = V_{TZ}(x) \quad \textit{when } \alpha = 0,\ \beta = 1,\ \textit{and } \frac{2\delta}{\mu} < 1. \tag{4.36}$$

Proof. First note that $V(x)$ and $V_{TZ}(x)$ coincide for $x \geq x_\beta$. Due to the continuity of both V' and V'_{TZ} at x_β, we deduce that

$$V'(x_\beta) = V'_{TZ}(x_\beta) \equiv X^{-1}(x_\beta). \tag{4.37}$$

Now we prove that (4.36) holds if and only if

$$a(x) = \frac{\mu}{\sigma^2}\left(\frac{C}{\Gamma}\left(X^{-1}(x)\right)^{-1/\Gamma} + \frac{2\delta\sigma^2}{\mu^2 + 2\gamma\sigma^2}\right), \qquad 0 \leq x \leq x_\beta. \tag{4.38}$$

To this end, first suppose that (4.38) holds. Since $V(x)$ is derived from (4.4) after calculating $a(x)$, we insert in (4.4) the right-hand side of (4.38) and derive

$$\begin{aligned} \log\left(\frac{V'(x_\beta)}{V'(x)}\right) &= -\int_x^{x_\beta} \frac{dx}{\frac{C}{\Gamma}\left(X^{-1}(x)\right)^{-1/\Gamma} + \frac{2\delta\sigma^2}{\mu^2+2\gamma\sigma^2}} \\ &= \log\left(\frac{X^{-1}(x_\beta)}{X^{-1}(x)}\right) \qquad \forall 0 \leq x \leq x_\beta, \end{aligned}$$

where the second equality follows by considering the change of variable $y = X^{-1}(x)$. By (4.37), we conclude that $V'(x) = X^{-1}(x) \equiv V'_{TZ}(x)$ for all $0 \leq x \leq x_\beta$. This together with the fact that $V(0) = V_{TZ}(0) = 0$ leads to (4.36). Conversely, suppose $V(x) = V_{TZ}(x)$ for all $0 \leq x \leq x_\beta$. By differentiating V twice, we get

$$V'(x) = X^{-1}(x), \qquad V''(x) = -\frac{X^{-1}(x)}{\frac{C}{\Gamma}\left(X^{-1}(x)\right)^{-1/\Gamma} + \frac{2\delta\sigma^2}{\mu^2+2\gamma\sigma^2}}.$$

These two equations combined with (4.4) imply that (4.38) holds.

So now we need only to prove the validity of (4.38). Let

$$Z(x) = \frac{\mu}{\sigma^2}\left(\frac{C}{\Gamma}\left(X^{-1}(x)\right)^{-1/\Gamma} + \frac{2\delta\sigma^2}{\mu^2+2\gamma\sigma^2}\right), \qquad 0 \le x \le x_\beta. \tag{4.39}$$

Since $V(x) = V_{TZ}(x)$ for $x \ge x_\beta$ and V' and V'' are continuous at x_β, we obtain via a similar calculation as above that $Z(x_\beta) = a(x_\beta)$. On the other hand, by differentiating $Z(x)$, we derive

$$Z'(x) = \frac{\mu}{\sigma^2\Gamma}\frac{\frac{C}{\Gamma}\left(X^{-1}(x)\right)^{-1/\Gamma}}{\frac{C}{\Gamma}\left(X^{-1}(x)\right)^{-1/\Gamma} + \frac{2\delta\sigma^2}{\mu^2+2\gamma\sigma^2}} = \frac{\mu}{\sigma^2\Gamma} - \frac{2\delta\sigma^2}{\mu^2+2\gamma\sigma^2}\frac{1}{Z(x)\Gamma} \quad \forall 0 \le x \le x_\beta.$$

Therefore, $Z(x)$ satisfies (4.12) with the boundary condition $Z(x_\beta) = a(x_\beta)$. This leads to (4.38) and the proof is completed. □

4.3. Case of $\beta \le \frac{2\delta}{\mu}$. Since in this case $a(0) \ge \beta$, we can apply Proposition 9 to conclude $a(x) \ge \beta$ for all $x \ge 0$. Substituting $a = \beta$ in (4.2), we get

$$V(x) = k_1(\beta)e^{r_+(\beta)(x-x_1)} + k_2(\beta)e^{r_-(\beta)(x-x_1)}, \qquad 0 \le x < x_1. \tag{4.40}$$

Using the principle of smooth fit at x_1, we get that $k_1(\beta)$ and $k_2(\beta)$ are given by (4.23). Using the initial condition $V(0) = 0$, we obtain

$$\Delta = -x_1 = \frac{1}{r_+(\beta) - r_-(\beta)}\log\left(\frac{r_+^2(\beta)}{r_-^2(\beta)}\right). \tag{4.41}$$

Note that the expression on the right-hand side of (4.41) is negative if and only if $|r_+(\beta)| < |r_-(\beta)|$. The latter is true if and only if

$$\frac{\delta}{\mu} < \beta; \tag{4.42}$$

see (4.8).

THEOREM 8. *Suppose $\frac{\delta}{\mu} < \beta \le \frac{2\delta}{\mu}$. Let $k_1(\beta)$, $k_2(\beta)$, $r_+(\beta)$, $r_-(\beta)$, and x_1 be given by* (4.23), (4.8), *and* (4.41), *respectively. Then*

$$V(x) = \begin{cases} k_1(\beta)e^{r_+(\beta)(x-x_1)} + k_2(\beta)e^{r_-(\beta)(x-x_1)}, & 0 \le x < x_1, \\ k_1(\beta) + k_2(\beta) + x - x_1, & x \ge x_1, \end{cases} \tag{4.43}$$

is a concave, twice continuously differentiable solution of the HJB equation (2.30).

Proof. The proof of this theorem follows the lines of the proof of Theorem 7, in which one replaces x_β by 0. □

Remark 4. The case when (4.42) fails is a trivial case; see Theorem 5.

Remark 5. From the expressions for $V_{\alpha,\,\beta}$ obtained in this subsection and the previous subsection, one can see that

$$\lim_{\beta\to\infty} V_{\alpha,\,\beta}(x) = \infty, \qquad \lim_{\alpha\to 0,\ \beta\to 0} V_{\alpha,\,\beta}(x) = 0 \qquad \forall x > 0.$$

5. Optimal policies. In this section we construct the optimal control policies based on the solutions to the HJB equations obtained in the previous sections. Recall that x_1 is the smallest number such that V'' vanishes. For each $x \le x_1$ define

$$a^*(x) \equiv \arg \max_{\alpha \le a \le \beta} \left(\frac{1}{2}\sigma^2 a^2 V''(x) + (a\mu - \delta)V'(x) - \gamma V(x) \right). \tag{5.1}$$

As evident from below the function $a^*(x)$ represents the optimal feedback control function for the control component a_t^π, $t \ge 0$. More precisely, the value $a^*(x)$ is the optimal risk that one should take when the value of the current reserve is x. From the analysis in section 4, it follows that $a^*(x)$ can be represented as

$$a^*(x) = \begin{cases} \alpha, & 0 \le x \le x_\alpha, \\ a(x), & x_\alpha \le x \le x_\beta, \\ \beta, & x \ge x_\beta. \end{cases} \tag{5.2}$$

Note that the values of the critical points x_α, x_β as well as the function $a(x)$ depend on the three different cases studied in section 4. Specifically, in the case of $\frac{2\delta}{\mu} < \alpha$ the values of x_α and x_β are specified by Theorem 6 while $a(x)$ is given by (4.15); in the case of $\alpha \le \frac{2\delta}{\mu} < \beta$, $x_\alpha = 0$ and x_β is given by Theorem 7 while $a(x)$ is determined by (4.29); in the case of $\beta \le \frac{2\delta}{\mu}$, $x_\alpha = x_\beta = 0$.

To determine the other component of the optimal control, C_t^π, $t \ge 0$, which is the singular control in the terminology of control theory, we need to involve the reflection processes which solve the so-called Skorohod problem for the one-dimensional diffusion. Let (R_t^*, C_t^*) be a solution to the following Skorohod problem on $t \ge 0$:

$$\begin{aligned} &R_t^* = x + \textstyle\int_0^t (a^*(R_s^*)\mu - \delta)ds + \int_0^t a^*(R_s^*)\sigma dW_s - C_t^*, \\ &R_t^* \le x_1, \\ &\textstyle\int_0^\infty 1_{\{R_s^* < x_1\}} dC_s^* = 0. \end{aligned} \tag{5.3}$$

This solution yields two processes R_t^* and C_t^*. The first is a diffusion process on $(-\infty, x_1]$ reflected at the upper boundary, and the second is an increasing process. Subtracting C_t^* from R_t^* results in the reflection of R_t^* from x_1. The last condition is the requirement that this functional increases only when the controlled process is at the boundary x_1, thus not affecting the dynamics of R_t^* whenever R_t^* is below x_1. Existence of a solution to such a Skorohod problem follows from Theorem 3.1 in [12]. For a process R_t with a constant drift and diffusion term ($R_t = x + \mu t + \sigma W_t$), a solution to the Skorohod problem can be written in a closed form via the so-called running maximum

$$R_t^* = x + \mu t + \sigma W_t - C_t^*, \qquad C_t^* = \max_{0 \le s \le t} [(x + \mu s + \sigma W_t - x_1)^+].$$

In the case when drift and diffusion coefficients are not constants, by and large the solution to the equations (5.3) cannot be found in a closed form. An ε-approximation to the solution to the Skorohod problem by a jump diffusion R_t^ε can be the following. The process R_t^ε is a diffusion on $(-\infty, x_1]$, and whenever this process reaches x_1, it jumps down to $x_1 - \varepsilon$. The corresponding process C_t^ε in this case is a purely discontinuous functional which increases by ε when R_t^ε reaches x_1. The solution (R_t^*, C_t^*) to (5.3) can be viewed as a limiting case of $\varepsilon \to 0$.

THEOREM 9. *Let V be a concave, twice continuously differentiable solution of the HJB equation* (2.30) *and* $(R_t^*, C_t^*; t \geq 0)$ *be a solution to the Skorohod problem* (5.3). *Then for* $\pi^* = (a^*(R_t^*), C_t^*; t \geq 0)$, *we have*

$$J_x(\pi^*) = V(x) \qquad \forall x \geq 0. \tag{5.4}$$

Proof. For simplicity assume that the initial position $x \leq x_1$. In this case both processes R_t^* and C_t^* as a solution to the Skorohod problem are continuous. In view of (5.1),

$$L^{a^*(R_s^*)} V(R_s^*) = 0, \tag{5.5}$$

where the operator L^a is defined in (2.33). Repeating the argument in proving (2.62) and applying (5.5), we see that

$$E(e^{-\gamma(t\wedge\tau)} V(R_{t\wedge\tau}^*)) = V(x) - E\int_0^{t\wedge\tau} e^{-\gamma s} V'(R_s^*) dC_s^*. \tag{5.6}$$

Since $V'(x_1) = 1$ and in view of (5.3),

$$1_{\{R_s^* = x_1\}} dC_s^* = dC_s^*, \tag{5.7}$$

we can replace $V'(R_s^*)$ in the integrand on the right-hand side of (5.6) by $V'(R_s^*)1_{\{R_s^*=x_1\}} = V'(x_1)1_{\{R_s^*=x_1\}}$ to obtain

$$\begin{aligned} E(e^{-\gamma(t\wedge\tau)} V(R_{t\wedge\tau}^*)) &= V(x) - E\int_0^{t\wedge\tau} e^{-\gamma s} V'(x_1) 1_{\{R_s^*=x_1\}} dC_s^* \\ &= V(x) - E\int_0^{t\wedge\tau} e^{-\gamma s} V'(x_1) dC_s^* = V(x) - E\int_0^{t\wedge\tau} e^{-\gamma s} dC_s^*, \end{aligned} \tag{5.8}$$

where in the last two equalities we used once more (5.7) and the condition $V'(x_1) = 1$. Taking limit as $t \to \infty$, and applying (2.65), we obtain the desired result. □

Combining Theorems 3 and 9, we get the following result immediately.

COROLLARY 3. *The function V presented in the previous sections is the value function and π^* is the optimal policy.*

Remark 6. Theorem 9 and Corollary 3 also imply that the HJB equation (2.30) has a unique solution in the class of concave, twice continuously differentiable functions.

Next we summarize all the results we obtained in Table 1 for easy reference.

6. Economic interpretation and conclusions. The optimal policies obtained in the previous sections have clear economic meaning and are very easy to implement. Let us now elaborate.

Theorem 5 is a mathematical formulation of the intuition that if a company has a liability rate not smaller than the maximal expected profit rate, then it is optimal to declare bankruptcy immediately, distributing the whole reserve as the dividend. In this case the risk control policy is irrelevant.

When an immediate bankruptcy is not optimal, the optimal risk control policy is characterized by two critical reserve levels: x_α and x_β. The values of these two levels are further determined by three parameters: the minimum risk allowed (α), the maximum risk allowed (β), and the ratio between the debt rate and profit rate ($\frac{\delta}{\mu}$). If the company has very little debt compared to the potential profit (so that $\frac{2\delta}{\mu} \leq \alpha$),

TABLE 1
Summary of results.

Range for $\frac{\delta}{\mu}$	x_α	x_β	$a^*(x)$	Risk α ever attained	x_1
$\frac{2\delta}{\mu} < \alpha$	positive and finite; see (4.9)	positive and finite; see (4.16)	(i) α, for $x \in [0, x_\alpha]$; (ii) increases from α to β on $[x_\alpha,\ x_\beta]$; see (4.15); (iii) β, for $x \geq x_\beta$	yes	positive; see (4.25)
$\frac{2\delta}{\mu} = \alpha$ / $\alpha < \frac{2\delta}{\mu} < \beta$	0	positive and finite; see (4.30)	(i) increases from $2\delta/\mu$ to β on $[0,\ x_\beta]$; (ii) β, for $x \geq x_\beta$	yes / no	positive; see (4.33)
$\frac{\delta}{\mu} < \beta \leq \frac{2\delta}{\mu}$	0	0	β	no	positive; see (4.41)
$\frac{\delta}{\mu} \geq \beta$ (trivial case)	0	0	any	N/A	0

then both the critical reserve levels, x_α and x_β, are positive and finite. In this case, the company will minimize the business activity (i.e., take the minimum risk α) when the reserve is below the level x_α, then gradually increase the business activity when the reserve is between x_α and x_β, and then maximize the business activity (i.e., take the maximum risk β) when the reserve ever reaches or goes beyond the level x_β.

Next, if the company has a higher debt-profit ratio (so that $\alpha < \frac{2\delta}{\mu} < \beta$), then the company has to be a bit more aggressive in the sense that $x_\alpha = 0$ and x_β is positive and finite. In this case, no matter how small the reserve is the company will never take the minimum risk; rather it will start with the risk level $\frac{2\delta}{\mu}$ and gradually increase to the maximum risk level β when the reserve hits the level x_β and goes above this level. This can be explained by the fact that when the debt rate is high one needs to gamble on the higher potential profits in order to get out of the "bankruptcy zone" as fast as possible, even at the expense of assuming higher risk. The company becomes more aggressive when the debt-profit ratio is even higher (precisely when $\frac{\delta}{\mu} < \beta \leq \frac{2\delta}{\mu}$), in which case the maximum allowable risk β is taken throughout while the two critical levels x_α and x_β are both zero. Finally, when the debt-profit ratio is so high that the debt-profit ratio is greater than the maximum risk possible, then the company should declare bankruptcy and go out of business immediately. This is due to the fact that the expected net cash flow is negative in this case, no matter what the company's policy might be.

On the other hand, the optimal dividend policy is always of a threshold type with the threshold being equal to x_1. Namely, the reserve should be kept below the critical level x_1 while distributing any excess as dividends. A simple realistic approximation of this policy is distributing a small amount of dividends whenever the process reaches x_1. If the initial reserve x exceeds x_1, then the optimal policy requires to distribute instantaneously all the excess above x_1. From the structure of our solution we also see that the maximum business activity is always taken on *before* dividend distributions take place.

In conclusion, we would like to point out an intricate interplay between the liability and restrictions on the risk control of a financial company. The sheer number of qualitatively different optimal policies, which appears due to different possible relationships between exogenous parameters, shows the multiplicity of different economic environments which a financial company faces depending on the size of the debt and on the size of available business activity.

Acknowledgments. We thank the associate editor and the two reviewers for their careful reading of an earlier version of the paper and for their constructive comments that led to an improved version.

REFERENCES

[1] S. Asmussen, B. Højgaard, and M. Taksar (2000), *Optimal risk control and dividend distribution policies. Example of excess-of loss reinsurance*, Finance Stoch., 4, pp. 299–324.
[2] S. Asmussen and M. Taksar (1997), *Controlled diffusion models for optimal dividend pay-out*, Insurance Math. Econom., 20, pp. 1–15.
[3] P. Boyle, R. J. Elliott, and H. Yang (1998), *Controlled Diffusion Models of an Insurance Company*, preprint, Department of Statistics, The University of Hong Kong.
[4] T. Choulli, M. Taksar, and X. Y. Zhou (2001), *Excess-of-loss reinsurance for a company with debt liability and constraints on risk reduction*, Quantitative Finance, 1, pp. 573–596.
[5] C. Dellacherie and P. A. Meyer (1980), *Probabilité et potentiels: Théorie des martingales*, Hermann, Paris.
[6] W. H. Fleming and R. W. Rishel (1975), *Deterministic and Stochastic Optimal Control*, Springer–Verlag, Berlin–New York.
[7] W. H. Fleming and H. M. Soner (1993), *Controlled Markov Processes and Viscosity Solutions*, Springer–Verlag, New York.
[8] B. Højgaard and M. Taksar (1998a), *Optimal proportional reinsurance policies for diffusion models with transaction costs*, Insurance Math. Econom., 22, pp. 41–51.
[9] B. Højgaard and M. Taksar (1998b), *Optimal proportional reinsurance policies for diffusion models*, Scand. Actuar. J., 2, pp. 166–168.
[10] B. Højgaard and M. Taksar (1999), *Controlling risk exposure and dividends pay-out schemes: Insurance company example*, Math. Finance, 2, pp. 153–182.
[11] M. Jeanblanc-Picque and A. N. Shiryaev (1995), *Optimization of the flow of dividends*, Russian Math. Surveys, 50, pp. 257–277.
[12] P.-L. Lions and A.-S. Sznitman (1984), *Stochastic differential equations with reflecting boundary conditions*, Comm. Pure Appl. Math., 37, pp. 511–537.
[13] J. Paulsen and H. K. Gjessing (1997), *Optimal choice of dividend barriers for a risk process with stochastic return on investments*, Insurance Math. Econom., 20, pp. 215–223.
[14] D. Revuz and M. Yor (1999), *Continuous Martingales and Brownian Motion*, 3rd ed., Springer–Verlag, Berlin.
[15] H. L. Royden (1988), *Real Analysis*, 3rd ed., Prentice–Hall, Upper Saddle River, NJ.
[16] R. Radner and L. Shepp (1996), *Risk vs. profit potential: A model for corporate strategy*, J. Econ. Dynam. Control, 20, pp. 1373–1393.
[17] M. Taksar (2000), *Optimal risk and dividend distribution control models for an insurance company*, Math. Methods Oper. Res., 51, pp. 1–42.
[18] M. Taksar and X. Y. Zhou (1998), *Optimal risk and dividend control for a company with a debt liability*, Insurance Math. Econom., 22, pp. 105–122.
[19] J. Yong and X. Y. Zhou (1999), *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer–Verlag, New York.